PREDICTION OF GENETIC VARIANTS ASSOCIATED WITH ANTIMALARIAL DRUG RESISTANCE USING SET COVERING MACHINE

APATA OLUWABUKOLA RACHEAL
19PBF02173

MAY, 2022

PREDICTION OF GENETIC VARIANTS ASSOCIATED WITH ANTIMALARIAL DRUG RESISTANCE USING SET COVERING MACHINE

BY

APATA OLUWABUKOLA RACHEAL
(19PBF02173)
B.Sc (Hons) Microbiology, Obafemi Awolowo University

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE (M.Sc) DEGREE IN BIOINFORMATICS IN THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY COVENANT UNIVERSITY, OTA, OGUN STATE, NIGERIA

MAY, 2022

ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfilment of the requirements for the award of the Master's degree in **Bioinformatics** in the department of **Computer and Information Sciences**, College of Science and Technology, Covenant University, Ota.

Mr. Taiwo B. Erewumi
(Secretary, School of Postgraduate Studies) **Signature & Date**

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies) **Signature & Date**

DECLARATION

I, **APATA OLUWABUKOLA RACHEAL (19PBF02173)**, declare that this research was carried out by me under the supervision of Dr. Isewon I.M of Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. I attest that this thesis has not been presented either wholly or partly for the award of any degree elsewhere. All sources of data and scholarly information used in this thesis are duly acknowledged.

APATA OLUWABUKOLA RACHEAL

Signature & Date

CERTIFICATION

We certify that the thesis titled "**PREDICTION OF GENETIC VARIANTS ASSOCIATED WITH ANTIMALARIAL DRUG RESISTANCE USING SET COVERING MACHINE**" is an original work carried out by **APATA OLUWABUKOLA RACHEAL, (19PBF02173)**, in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria, under the supervision of Dr. Isewon I.M. We have examined and found that the work acceptable for the award of Master's degree in Bioinformatics.

Dr. Itunuoluwa M. Isewon
(Supervisor) Signature & Date

Prof. Olufunke O. Oladipupo
(Head of Department) Signature & Date

Dr. Victor T. Odumuyiwa
(External Examiner) Signature & Date

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies) Signature & Date

DEDICATION

I dedicate this thesis to God almighty and the lovelies that hold a special place in my heart - my mother, and siblings.

ACKNOWLEDGEMENTS

I give all praise and glory to God the Father, the Son, and the Holy spirit for showers of blessings throughout my research work and grace for successful completion. I want to express my heartfelt gratitude to Dr. Itunuola Isewon, my research supervisor, for her selfless commitment in giving invaluable guidance during this research. Her dynamism, motivation, sincerity, and vision profoundly inspire me. I thank the management of Covenant University, CUBRe, and CApIC-ACE for the funding and support on this work. A profound thank you to the HoD, Associate Professor Oladipupo Grace, the PG coordinator, Dr. Oni and other faculty of Computer and Information Science depart for academic supports and contributions.

I am incredibly grateful to my lovely mother, my gold - Mrs. Apata Mobolanle J., for her love, prayers, care, and sacrifices to educate and prepare me for adulthood. Thank you, mummy. I cannot love you less. I appreciate my best half - Fabelurin Omololu O. for his love, understanding, prayers, and continuing support in the course of this research work. I express my appreciation to my brother – Olusegun Apata, sisters – Mobolaji, Afolakemi, Temitope and Oluwafunmilayo Apata, brothers-in-law, sisters-in-law and other relatives for their unwavering support and prayers. I extend my gratitude to my friends, co-research assistants and colleagues – Owolabi Jesujoba, Oluwamuyiwa Fesobi, Bisi-Adeniyi Titilayo, Molo Mbasa, Vingi Patrick, Enoma David, Oladejo David, Suraj Sadeeq for their constant support and companionship. I thank Prof. Jelili Oyelade for his support and words of encouragement. My Sincere appreciation also goes to the Admin staffs of CUBRe and CApIC-ACE – Mr. Babajide, Miss Helen, Mr. Obaoye. Finally, I want to express my gratitude to everyone who has helped me complete the research work, whether directly or indirectly. God bless you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVATIONS

ABBREVIATION	MEANING
ACT	Artemisinin-based Combination Therapy
ADR	Antimicrobial Drug Resistance
AMR	Antimalarial Resistance
BP	BasePair
CNV	Copy-Number Variation
CQ	Chloroquine
DHA	Dihydroartemisinin
DHFR	Dihydrofolate reductase-thymidylate synthase
DHPS	Dihydropteroate synthase
DNA	Deoxyribonucleic Acid
ENA	European Nucleotides Archive
GLM	Generalized Linear Model
GWAS	Genome-Wide Association Study
IC50	50% inhibitory concentration
ICEs	Integrative and Conjugative Elements
INDELs	Insertions and Deletions
ISs	Insertion Sequences
Kb	Kilobases
LUM	Lumefantrine
MDR	Multidrug Resistance
MGEs	Mobile Genetic Elements
MQ	Mefloquine

MTB	*Mycobacterium tuberculosis*
NCBI	National Center for Biotechnology Information
PFCRT	Plasmodium falciparum chloroquine resistance transporter gene
PFMDR1	Plasmodium Falciparum Multidrug Resistance
Pgh1	P-glycoprotein homologue 1
PQ	Primaquine
PVL	Panton-Valentine Leucocidin
PYR	Pyremethamine
SCM	Set Covering Machine
SIs	Sequence Inversions
SNP	Single Nucleotide Polymorphism
TESs	Therapeutic Efficacy Studies
WHO	World Health Organization

ABSTRACT

Antimalarial resistance (AMR) has become a major issue in malaria-endemic countries, and novel methods for identifying strains resistant or susceptible to specific medications are critical in the fight against antimalarial-resistant Plasmodium parasites. The growing availability of genetic information has enabled the application of computational methods in surveying resistance patterns. K-mer-based machine learning approaches have shown considerable potential as a diagnostic and research tool. In this work, Set Covering Machine (SCM) algorithm was applied to predict antimalarial drug response outcomes and their genetic determinants. The model predicted six antimalarial drugs *(*Chloroquine, Dihydroartemisinin, Lumafantrine, Primaquine, Pyrimethamine, and Mefloquine) response phenotype in *Plasmodium falciparum.* The model used the most compact set of k-mers generated from the genomes of the parasite isolates to learn and predict binary drug response outcomes. To avoid model overfitting, ten-fold cross-validation was conducted on the training set to choose the optimal hyperparameter values. Regardless of the resistance mechanism, whether acquired resistance or point mutations in the chromosome, the training accuracy (mean cross-validation score) and testing accuracy of SCM prediction of the six antimalarial drug resistance was above 85%. The model significantly classified the resistant isolates from the sensitive isolates of the parasite and could be used as potential tools in antimalarial resistance surveillance and clinical studies. A number of sequence k-mers associated with antimalarial drug resistance were identified. We identified several already known genes and loci associated with the six drugs, including those containing *pfcrt* and *pfdhfr*. Novel genes and loci were also discovered. Of particular interest are the variant regions on the var genes on chromosomes 6, 8, 10, and 13 containing the *Plasmodium falciparum* erythrocyte membrane *protein 1 (PfEMP1*). The *PfEMP1* variant k-mers were found to be associated with chloroquine, dihydroartemisinin, and pyrimethamine resistance. The var genes encode PfEMP1. The genes have extreme variability and are a principal virulence factor of malaria parasite with extreme antigenic variability. The variations in these var genes were found to play a role in antimalarial drug resistance in *P. falciparum.*

Keywords**: Machine learning, Malaria, Plasmodium falciparum, Genome-Wide Association Study, Phenotype prediction