

**INFERRING GENE REGULATORY NETWORK FOR THE
SPOROGENIC STAGE OF *PLASMODIUM FALCIPARUM* LIFE
CYCLE USING GRAPH NEURAL NETWORKS AND scRNA-SEQ
DATASET**

**ONYEMAECHI, PROMISE
(22PBF02398)**

AUGUST, 2024

**INFERRING GENE REGULATORY NETWORK FOR
SPOROGENIC STAGE OF *PLASMODIUM FALCIPARUM* LIFE
CYCLE USING GRAPH NEURAL NETWORKS AND scRNA-SEQ
DATASET**

BY

**ONYEMAECHE, PROMISE
(22PBF02398)**

**B. Sc. (Hons) Plant Biology and Biotechnology, University of Benin,
Benin City.**

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

AUGUST, 2024

ACCEPTANCE

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

Miss Adefunke F. Oyinloye
(Secretary, School of Postgraduate Studies)

Signature and Date

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

Signature and Date

DECLARATION

I, ONYEMAECHI, PROMISE (22PBF02398) declare that this research was carried out by me under the supervision of Prof. Marion Adebisi of the Department of Computer, Landmark University, Omu-Aran, Nigeria. I attest that this dissertation has not been submitted either wholly or partially for the award of any degree elsewhere. All sources of data and scholarly information used in this dissertation are duly acknowledged.

ONYEMAECHI, PROMISE

Signature and Date

CERTIFICATION

This is to certify that this dissertation titled "**INFERRING GENE REGULATORY NETWORK FOR THE SPOROGENIC STAGE OF *PLASMODIUM FALCIPARUM* LIFE CYCLE USING GRAPH NEURAL NETWORKS AND scRNA-SEQ DATASET**" is original research earned out by **ONYEMAECHI, PROMISE (22PBF02398)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria under the supervision of Prof. Marion Adebisi. We have examined and found this work acceptable as part of the requirements for the award of Master of Science (M.Sc.) in Bioinformatics.

Prof Marion O. Adebisi
(Supervisor)



Signature and Date

Prof. Olufunke. O Oladipupo
(Head of Department)

Signature and Date

Prof. Olusegun Folorunso
(External Examiner)

Signature and Date

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

Signature and Date

DEDICATION

I dedicate this project to God Almighty for His unending supply of grace, and strength given to me during my master's degree programme. Furthermore, to biomedical and computational researchers Who paved the way, providing platforms for insightful biomedical investigations.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to Almighty God for His unwavering guidance and blessings to this point. Special thanks to my supervisor, Prof. Marion Adebisi, for her invaluable guidance, support, and mentorship in this research journey. I am also deeply appreciative of the Head of Department, Prof. O. Oladipupo, and the Postgraduate Coordinator, Dr. Itunuoluwa Isewon, for their invaluable input and organization. I extend my sincere appreciation to CApIC-ACE for supporting my studentship as well as the members of the faculty, staff, and students of the Department, for their contributions and support in making this proposal possible.

TABLE OF CONTENTS

CONTENTS	PAGES
TITLE PAGE	i
COVER PAGE	ii
ACCEPTANCES	iii
DECLARATION	iv
CERTIFICATION	v
DEDICATION	vi
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS	vii
LIST OF FIGURES	xi
LIST OF TABLES	xii
LIST OF ABBREVIATIONS	xiii
ABSTRACT	xiv
CHAPTER ONE: INTRODUCTION	1
1.1 Background of the Study	1
1.2 Statement of the Problem	5
1.3 Research Questions	6
1.4 Aim and Objectives	6
1.5 Research Methodology	6
1.6 Significance of the Study	8
1.7 Scope of the Study	9
1.8 Limitations of the Study	9
1.9 Organization of Dissertation	9
CHAPTER TWO: LITERATURE REVIEW	10
2.1 Preamble	10
2.2 Malaria	10
2.3 Malaria Parasites	11
2.4 Life cycle of <i>Plasmodium falciparum</i>	13
2.5 The <i>Plasmodium</i> genomes	14
2.6 Graph Theory	16
2.6.1 Directionality: directed graphs and undirected graphs	17
2.6.2 Degrees and neighbourhoods	17
2.6.3 Incident	18
2.6.4 Path	18
2.6.5 Loop or cycle	18

2.6.6	Cyclic and acyclic graph	18
2.6.7	Representing graphs	18
2.6.8	Adjacency List	19
2.6.9	Adjacency Matrix	19
2.6.10	Adjacency Arrays	20
2.7	Deep learning	21
2.7.1	Graph Neural Networks (GNNs)	21
2.7.2	Graph Convolutional Networks	25
2.7.3	Graph Autoencoder Networks	26
2.8	Gene Regulatory Network	28
2.8.1	Methodological foundations of GRN inference	30
2.9	Single-cell RNA-sequencing (scRNA-seq) data	34
2.10	Adoption of Single-cell RNA-sequencing (scRNA-seq) data in GRN inference	35
2.11	Review of Related Works	37
CHAPTER THREE: RESEARCH METHODOLOGY		41
3.1	Preamble	41
3.2	Dataset	41
3.3	Preprocessing of scRNA-sequencing dataset	41
3.4	Gene Clustering	42
3.5	Construction of correlation-based co-expression network	42
3.6	Reconstruction of GRN using Graph convolutional network	42
3.7	Model Training	43
3.8	Performance Evaluation	43
CHAPTER 4: RESULT AND DISCUSSION		46
4.1	Preamble	46
4.2	Results for Objective One	46
4.2.1	Single-cell RNA seq dataset: A Valuable Tool for GRN Inference in <i>Plasmodium falciparum</i>	46
4.2.2	Preprocessing	47
4.2.3	Clustering of genes	47
4.3	Result from Objective 2	50
4.3.1	Construction of adjacency matrix and weighted gene co-expression graph network	50

4.3.2	GRN reconstruction	51
4.3.3	Performance Evaluation	52
4.4	Result from Objective 3	54
4.4.1	Computing centrality score and identification of master regulators	54
4.4.2	GRN Visualization	57
CHAPTER FIVE: CONCLUSION AND RECOMMENDATION		60
5.1	Summary	60
5.2	Conclusion	60
5.3	Contribution to Knowledge	61
5.4	Recommendation	61
REFERENCES		63
APPENDICES		73

LIST OF FIGURES

FIGURES	TITLE OF FIGURES	PAGES
Figure 1.1:	Overview of the workflow GRN using scRNA-seq data based on graph convolutional neural network model.	4
Figure 2.1:	The life cycle of Plasmodium falciparum parasites.	14
Figure 2.2:	A visual representation of graph G.	16
Figure 2.3:	Example of directed(a) and undirected(b) graph on 4 vertices.	17
Figure 2.4:	A graph and its adjacency lists.	19
Figure 2.5:	A graph and its adjacency matrix.	20
Figure 2.6:	A graph and its adjacency array.	21
Figure 2.7:	Representation of how information is gathered from the nearby neighbourhood of a node.	24
Figure 2.8:	An example of spatial GCNs. Every node in the feed-forward computation compiles data from both its own and its neighbours' representations.	26
Figure 2.9:	(A) GRN showing the directed interactions of genes and their regulatory targets (Dautle et al., 2023), (B) and master regulators and target genes in a biological network.	29
Figure 3.1:	Flow diagram showing the pipeline of the study.	45
Figure 4.1:	Gene cluster represented in the scRNA-seq dataset, embedded in two dimensions using UMAP and clustered using the Leiden algorithm.	49
Figure 4.2:	Number of genes in each Leiden cluster.	49
Figure 4.3:	Distribution of the weight of gene interactions in the four inferred GRNs labelled as GRN 1, GRN 2, GRN 3, GRNN 4.	51
Figure 4.4:	Heatmap of performance metrics for GRN predicted for the four networks.	53
Figure 4.5:	Distribution of Degree Centrality, Betweenness Centrality, and PageRank of genes across the network for each cluster.	55
Figure 4.6:	Visualization of the inferred GRN for gene cluster 1.	57
Figure 4.7:	Visualization of the inferred GRN for gene cluster 2.	58
Figure 4.8:	Visualization of the inferred GRN for gene cluster 3.	58
Figure 4.9:	Visualization of the inferred GRN for gene cluster 4.	59

LIST OF TABLES

TABLES	TITLE OF TABLES	PAGES
Table 1.1:	Summary of Objectives, methods, and outcome.	8
Table 2.1:	An illustration of data representation in scRNA-seq dataset (count matrix)	35
Table 4.1:	AUC, AUROC ratio, and EPR of the 4 predicted networks	52
Table 4.2:	Top 5 Genes with the highest composite score for Degree Centrality, Betweenness Centrality, and PageRank across four inferred Gene Regulatory Networks (GRNs).	56
Table 4.3:	Master regulators across the four inferred GRNs based on genes with the highest composite centrality score.	56

LIST OF ABBREVIATIONS

CNN: Convolutional Neural Network

CGN: Graph Convolutional Neural Network

GNN: Graph Neural Network

GRN: Gene Regulatory Network

GWAS: Genome-Wide Association Studies

ITNs: Insecticide-Treated Bed Nets

KO: Knock Out

NGS: Next-Generation Sequencing

PCA: Principal Component Analysis

RBC: Red Blood Cell

RNA-seq: RNA Sequencing

RNN: Recurrent Neural Network (deep learning model)

scRNA-seq: Single-Cell RNA Sequencing

UMAP: Uniform Manifold Approximation and Projection

WHO: World Health Organization

ABSTRACT

Malaria, primarily caused by the parasite *Plasmodium falciparum*, remains one of the most severe infectious diseases globally, particularly in sub-Saharan Africa, where it leads to significant morbidity and mortality. A deep understanding of the molecular mechanisms that govern *P. falciparum* infection, especially during the sporogonic stage of the parasite's life cycle, is crucial for controlling malaria and identifying new therapeutic targets. In this context, gene regulatory networks (GRNs) provide a valuable framework for exploring the gene interactions that drive the parasite's life cycle and its response to environmental changes. This research aims to the GRNs for the sporogonic stage of *P. Falciparum* life cycle using Graph Neural Networks (GNNs), a deep learning technique applied to single-cell RNA sequencing (scRNA-seq) data. With the advent of scRNA-seq, gene expression can now be studied at the individual cell level. This is particularly important in the life cycle of *P. falciparum*, whose sporogonic stage is heterogeneous and involves different cell types throughout its life cycle. The study began with the preprocessing of scRNA-seq data, followed by clustering the cells to identify distinct stages of the parasite. A correlation-based co-expression network was then constructed to capture the relationships between genes within these clusters. Graph convolutional neural networks were employed to reconstruct the GRNs, leveraging their ability to learn the interactions between genes in the network. The performance of the inferred GRNs was evaluated using the AUC, AUPRC ratio, and EPR metrics, with average values of 0.6756, 1.51, and 2.48, respectively. This research not only enhances the understanding of the sporogonic stage of the life cycle of *P. falciparum*'s gene regulatory mechanisms but also identifies potential master regulators and genes that play central roles in the parasite's survival and pathogenicity. Four master regulators were identified for four selected clusters of genes associated with the life cycle of the sporogonic stage, based on the centrality scores of genes in the network. Among these master regulators, PF3D7_0304500, PF3D7_1357200, PF3D7_0822300, and PF3D7_0212300 stood out as the most critical genes within their respective clusters.

Keywords: *Plasmodium falciparum*, malaria, Graph Neural Network, scRNA seq, Gene regulatory network, master regulator.