COMPUTATIONAL ANALYSIS OF *ANOPHELES GAMBIAE* METABOLISM TO FACILITATE INSECTICIDAL TARGET AND COMPLEX RESISTANCE MECHANISM DISCOVERY

BY

ADEBIYI, MARION OLUBUNMI

(CUGP060188)

DOCTOR OF PHILOSOPHY DEGREE IN COMPUTER SCIENCE

JUNE, 2014

CERTIFICATION

I hereby certify that this is an original research work carried out by Marion Olubunmi ADEBIYI in the Department of Computer and Information Sciences, School of Natural and Applied Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria, under my supervision.

1. Name: Professor Ezekiel Adebiyi
   (Supervisor)

   Signature: _____ Date: _____

2. Name: Dr. Jason Rasgon
   (Co-Supervisor)

   Signature: _____ Date: _____

3. Name: Dr. Nicholas Omoregbe
   (Head of Department)

   Signature: _____ Date: _____

4. Name: Professor Emmanuel Onibere
   (External Examiner)

   Signature: _____ Date: _____

DECLARATION

It is hereby declared that this research was undertaken by Marion Olubunmi Adebiyi. The thesis is based on her original study in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, under the supervision of Prof. Ezekiel Adebiyi and Dr. Jason Rasgon. Ideas and views of this research work are products of the original research undertaken by Marion Olubunmi Adebiyi and the views of other researchers have been duly expressed and acknowledged.

Prof. Ezekiel Adebiyi

(Supervisor)

Signature: _____ Date: _____

Dr. Jason Rasgon

(Co- Supervisor)

Signature: _____ Date: _____.

DEDICATION

I dedicate this work to **GOD** Almighty, the giver of life, wisdom and inspirations. He is the one that supplied inspiration and strength to achieve this great task.

ACKNOWLEDGEMENT

I am forever grateful to the Alpha and the Omega of my life, the one who knows the end from the beginning and the Faithful one who has successfully seen me through my PhD programme. I do not take your deeds for granted Oh Lord, be glorified forever!

I want specially appreciate the Chancellor of Covenant University, Bishop David Oyedepo, who through his obedience to a divine call created the platform and the wings on which I am riding on to destiny today. Sir, you are indeed a shepherd and a father, you lead, guide and ensure safety for your children. Thank you Daddy.

My immense and unreserved thanks goes to the Vice Chancellor, Professor C. K Ayo, who always makes me hopeful and had continually been a place of succor for me. God will continually lead and guide you as you lead us to greatness.

I would sincerely like to thank my Supervisor, Prof. E. F. Adebiyi, he has been one of the most patience and helpful guides I could ever ask for, I could not have completed this task without him. Working on this thesis has been a great learning experience for me and I will forever be grateful to him for giving me this opportunity to learn and extract from his wealth of experience. Thank you so much sir, for your excellent supervision. I must thank Dr. Jason Rasgon for his great professional touch, attention and influence on this work, I sincerely appreciate you sir.

My unquantifiable gratitude and bunch of appreciation goes to Dr. Segun Fatumo whose valuable inputs, and numerous ideas that considerably expanded and improved this thesis cannot, but be mentioned. I am ever grateful to Mrs. Itunu Isewon and Mr. Joshua Okeniyi for many useful discussions and materials we exchanged during the course of this work.

My profound gratitude also goes to my Head of Department, Dr. Nicolas A. Omoregbe for his understanding and co-operation throughout this research work. I appreciate the support of my family friend and senior colleague, Mr. and Dr. (Mrs.) Oladipupo in the course of the project, their tremendous words of encouragement and support saw me through the stagnant moments. My sincere thanks to Mr. Oluranti Jonathan of CIS Department, you were there for me when I needed men to run around physically, you will not lack God's mercy and support in all your endeavors. My appreciation also goes to Dr. Okuboyejo, Dr. Oni, Dr. Azeta, Dr. Osamor, Dr. Oyelade, Mrs. Agoha, Dr. Ajayi of Biological Sciences and everyone that has been praying for the success and speedy finishing of this PhD. Thanks to everybody in CIS Department for their encouragement and support.

An immeasurable heartfelt love goes to Ayodele, my husband and best friend, who put up with the clutter, gave every support, sacrificed with maturity and patiently gave his encouragement and kept me focused on the work. My deep love and hip of gratitude to my little Angel, IreOluwa Joy and lovely charming Prince, IranwoOluwa Joshua for their love, understanding, prayers, concern and acceptance throughout the program.

Words cannot be enough to say thank you to my parents Chief (Mrs.) Alice E. Owolabi and parent in law, Pa. Amos Abiodun Adebiyi for being so supportive. Thank you for your prayers, advice and encouragements.

Finally, my siblings, who encouraged my curiosity and the desire to sit up and never give up; Dr. & Mrs. Olu Owolabi, Mr.& Mrs. Kunle Owolabi, Mr & Mrs. J. Dada, Engr. & Mrs. Femi Owolabi, Pst. & Dr. Grace Olasehinde, Kehinde Adebiyi, Esther Bamikole, Uncle Mayo, Uncle Yomi & Aunty Faith, God bless you.

Table of Contents

CERTIFICATION I
DECLARATION II
DEDICATION III
ACKNOWLEDGEMENT IV
TABLE OF CONTENTS VI
LIST OF TABLES X
LIST OF FIGURES XI
ABRBREVIATIONS XIII
DEFINITIONS OF TERMS XV
ABSTRACTS XVI

CHAPTER ONE: INTRODUCTION 1
1.1 BACKGROUND TO THE STUDY 1
1.2 STATEMENT OF THE PROBLEM 4
1.3 AIM AND OBJECTIVES OF THE STUDY 5
1.4 SIGNIFICANCE OF THE STUDY 6
1.5 CONTRIBUTION TO KNOWLEDGE 7
1.6 LIMITATIONS OF THE STUDY 7
1.7 OVERVIEW OF THE THESIS 8

CHAPTER TWO: LITERATURE REVIEW 10
2.1 INTRODUCTION 10
2.2 THE MALARIA VECTOR 10
2.2.1 Life Stages (Cycle) of *Anopheles Gambiae* 13
2.3 INSECTICIDES 18
2.3.1 Insecticide Resistance 20
2.3.2 Insecticide Resistance Mechanisms 21
2.3.3 Insecticide Resistance Management 23
2.4 MALARIA VECTOR CONTROL PROGRAMMES 26
2.4.1 Indoor Residual Spraying (IRS) 27

2.4.2 Pyrethroid Treated Bednets (LLITNs) 30
2.5 OMICS 33
2.5.1 Genomics 33
2.5.2 Proteomics 36
2.5.3 Metabolome: 37
2.5.4 Transcriptome 37
2.5.5 Microarrays 39
2.5.5.1 Categories of microarrays 42
2.5.5.2 Application of microarrays 43
2.5.5.3 Understanding the Concept of Microarray Technology 44
2.5.5.4 Microarray Data 48
2.6 PATHWAY/GENOME DATABASE (PGDB) 49
2.6.1 Pathway Tools 52
2.6.2 Pathologic 56
2.6.3 Hole Filler 57
2.6.4 Pathway Tools Navigator and Editor 57
2.6.5 Other Pathway/Genome Databases 57
2.7 WEB BASED DATABASES AND RESOURCES 60
2.7.1 GenBank 60
2.7.2 AnoBase 61
2.7.3 Kyoto Encyclopedia of Genes and Genomes (KEGG) 62
2.7.4 PubMed 62
2.7.5 BioCyc 63
2.7.6 ParCrys 63
2.7.7 UniProt 64
2.7.8 Protein Data Bank (PDB) 64
2.8 ANNOTATION TOOLS 65
2.8.1 GOPET 66
2.8.2 Domain Sweep 66
2.9 RECONSTRUCTING METABOLIC NETWORK 67

2.9.1 Social and economic networks 68
2.9.2 Information networks 68
2.9.3 Technological networks 68
2.9.4 Infrastructural networks 68
2.9.5 Temporary networks 68
2.9.6 Biological/ Biochemical Networks 68
2.9.6.1 Signal Transduction Networks (Pathways) 70
2.9.6.2 Regulatory Networks 71
2.9.6.3 Metabolic networks (Metabolic pathway) 71
2.9.6.4 Metabolic Network Reconstruction 72
2.10 NETWORK TOPOLOGICAL ANALYSIS 74
2.10.1 Choke Point 74
2.10.2 Load Point 75
2.10.3 Damage 76
2.10.4 Flux Balance Analysis (FBA) 76
2.10.5 Reaction without Deviation (RWD) 76
2.11 CLUSTERING TECHNIQUES 78
2.11.1 Data based Clustering Algorithms 78
2.11.2 Graph Based Network Clustering Modeling 81
2.12 ANALYZING TRANSCRIPTION DATA ON METABOLIC NETWORK 86

CHAPTER THREE: RESEARCH METHODOLOGY 88
3.1 INTRODUCTION 88
3.2 CONSTRUCTION OF A NEW PATHWAY/GENOME DATABASE (PGDB) FOR *A. GAMBIAE* 90
3.3 COMPUTATIONAL ANALYSIS OF *A. GAMBIAE* METABOLIC NETWORK 92
3.4 RE-CONSTRUCTING THE *A. GAMBIAE* METABOLIC NETWORK 94
3.5 COMPUTATIONAL ANALYSIS OF THE RECONSTRUCTED METABOLIC NETWORK 96
3.5.1 Reaction without Deviation (RWD) Algorithm (Steps 3-4) 98
3.5.2 Searching PDB for Every Gene that has its Structure or Homologue (>= 30) (from 102

3.5.3 Ranking Proteins by their Success in Crystallization Using OB and ParCrys Prediction (from step 6) 102
3.6 DISCOVERY OF *A.GAMBIAE* RESISTANCE MECHANISMS AGAINST PYRETHROIDS 105
3.6.1 Description of the Data Used 106
3.6.2 Construction of Metabolic Network from AnoCyc 111
3.6.3 Model for Analyzing Gene Expression Data on Metabolic Networks 112
3.6.4 Network Clustering Using Simulated Annealing Algorithm 117
3.6.5 Mapping Gene Expression Data onto Reactions 119
3.6.6 Feature Extraction Using Combinatorial Approach 120
3.6.7 Manual Analysis of Stimulated or Repressed Pathways 120

CHAPTER FOUR: RESULTS 122
4.1 INTRODUCTION 122
4. 2 CONSTRUCTION OF A NEW PGDB FOR *A. GAMBIAE* 122
4.3 COMPUTATIONAL ANALYSIS OF *A. GAMBIAE* METABOLIC NETWORK 127
4.4 DISCOVERY OF *A. GAMBIAE* RESISTANCE MECHANISMS AGAINST PYRETHROIDS 138

CHAPTER FIVE: DISCUSSION 145
5.1 SUMMARY 145
5.2 CONCLUSION 146
5.3 RECOMMENDATION 147
5.4 FUTURE WORK 148

PUBLICATIONS FROM THE WORK 149

REFERENCES 150

WEB REFERENCES 166

APPENDIX 168

LIST OF TABLES

CHAPTER 2

Table 2.1: The Omes Table (http://bioinfo.mbb.yale.edu) ……..38

Table 2.2: Tier 1 Databases and their Scope……..……..53

Table 2.3: Other Pathway Genome Database (PGDB), Outside BioCyc…….…….…….58

CHAPTER 3

Table 3.1: Important Statistics from CP and RWD Analysis …….……..101

Table 3.2 OB & ParCrys Ranking Table for our Target List..……...103

Table 3.3: The Gold Standard List……... ……....104

CHAPTER 4

Table 4.1 Distribution of Genes within Replicons on Existing AnoCyc……...….……123

Table 4.2 Classification Table for 206 Pathways in First AnoCyc Version…..…….....125

Table 4.3: List of Generated Data with File Name and File Type.....…..…...…….....126

Table 4.4: 61 Predicted Insecticidal Targets and Information for Each ……129

Table 4.5: The 61 Predicted Insecticidal Targets with their Homolog
E-values and Corresponding Transcripts in Human,
Bird (Chicken) and Fish (Tilipia)……134

Table 4.6: Cluster 5 genes……...……139

Table 4.7: Cluster 3 genes.....143

Table 4.8: Cluster 1 genes.....143

APPENDIX

Table 1: Predicted insecticidal targets and homology test with the Human…….….168

Table 2: Predicted insecticidal targets and homology test with Bird (Chicken)……....171

Table 3: Predicted insecticidal targets and homology test with Fish (Tilapia)……..…….174

LIST OF FIGURES

CHAPTER 2

Figure 2.1: An Adult *Anopheles gambiae* (WHO, 2012)…….10

Figure 2.2: *Anopheles gambiae* Life Cycle (www.mosqitoes.com).....13

Figure 2.3: *Anopheles* Larva…….14

Figure 2.4: Malaria Life Cycle (http://www.cartercenter.org).....17

Figure 2.5: A Spray-man in IRS, (MAPS, 2013)…….30

Figure 2.6: Woman in a Tanzanian Village Re-dipping Bed-net in a Pyrethroid Emulsion after 6 Months Usage (MAPs, 2013)…….32

Figure 2.7: cDNA Microarray Experiment…….46

Figure 2.8: Oligonucleotide Chip Experiment……47

Figure 2.9: Networks Analysis Flowchart…….68

Figure 2.10: The Knocked out Reaction is a Choke-point but may not be Essential for the Organism if the dashed lines exist in the Metabolic Network. ...…….77

Fig 2.11: Evolving Clustering Algorithms and K-means variants……...80

Figure 2.12: Compound Graph…….83

Figure 2.13: Reaction Graph…….84

Figure 2.14: Bipartite graph showing Reaction 1:
glutamate + ATP ⟶ gamma-glutamylphosphate + ADP
(catalysed by gamma glutamyl kinase; EC 2.7.2.11)…….86

CHAPTER 3

Figure 3.1: Frontpage of AnoCyc on BioCyc……90

Figure 3.2: Schematic Diagram of Ano2Cyc ver. 1.1 Developmental Stages……..91

Figure 3.3: Bipartite Graph Representation of One of the Reactions on our Network..…...96

Figure 3.4: *In-silico* Knock-out of Essential Reactions.....…...100

Figure 3.5 Graphical Representation of set of Reactions Given two Different Formats….112

Figure 3.6: The Splash Screen……..114

Figure 3.7: Workbench in Data Loading Stage, Just Before Analysis Begins…….....115

Figure 3.8: The Workbench Interface with Results in Display……..……116

CHAPTER 4

Figure 4.1: Overview of the Metabolic Map for *A. gambiae*……..…...124

Figure 4.2: Venn Diagram Showing Details of Enzymes from List of CP, RWD and *gambiae* Target (GT) in Comparison with the Gold standard List of 20 Enzymes……128

ABBREVIATIONS

Abbreviations	Meaning
DDT	Dichloro-diphenyl-trichloroethane
ITN	Insecticide treated nets
IRS	Indoor residual spraying
WHO	World Health Organizations
BHC	Benzenehexachloride
GST	Glutathione-S-transferases
OP	Organophosphate
OC	Organochloride
GABA	Gama-(γ)-aminobutyric Acid
Bti	*Bacillus Thuringiensis* Insecticides
A. gambiae	*Anopheles gambiae*
Kdr	Knock down resistance
P450s	Cytochrome-P450-monooxygenases
HCH	Hexacholine
GM	Genetically modified
Na^+	Sodium ion
KEGG	Kyoto Encyclopedia of Genes and Genomes
Pf	*Plasmodium falciparum*
PDB	Protein Data Bank
GUI	Graphic user interface

MAP	Malaria Atlas Project
KOBAS	KEGG Orthology Based Annotation System
GO	Gene Ontology
PPCpred	Predictor of Protein Production, Purification and Crystallization
ParCrys	Parzen Window to estimate a protein's propensity to produce diffraction-quality crystals
LLITNs	Long-Lasting Insecticide Treated Nets
PGDB	Pathway/Genome Database
DKFZ	German Cancer Research Center
GOPET	Gene ontology prediction and evaluation tool
OB - Score	Overton and Barton Score
GRF	Genome Resource Facility
EC	Enzyme Classification
NGS	Next Generation Sequencing

DEFINITION OF TERMS

Choke Point (CP): These are biochemically essential points in the network of an organism. They are reactions that either uniquely produce a specific product or consume a certain specific substrate in a metabolic network.

Reaction without Deviation (RWD): Are also known as **Essential Reactions** if the producing product downstream of it was indicated to hamper.

Load Point (Load Point): Is the identification of nodes with a high ratio of k-shortest paths to the number of nearest neighbor edges or connectivity providing many alternative metabolic pathways.

Damage: Is the procedure that allows access to the enzymes that may serve as drug targets when their inhibition influences a broader amount of downstream reactions and products.

Gene: Is the basic unit of heredity in a living organism. All living things depend on genes. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring.

Enzymes: Substance produced by a living organism which acts as a catalyst to bring about or speed up a specific biochemical reaction.

Protein: Proteins are the main building blocks and functional molecules of the cell

Deoxyribonucleic Acid: (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses.

Ribonucleic Acid: (RNA) is a biologically important type of molecule that consists of a long chain of nucleotide units. Each nucleotide consists of a nitrogenous base, a ribose sugar, and a phosphate.

Insecticidal Target: An enzymatic reactions is considered a potential insecticidal target if such an enzymatic reaction is a choke point (CP) as well as reaction without deviation (RWD)

Resistance Genes: These are the genes that do the flushing, excretion and or refusal of insecticides; such genes are even altered or change their mode of response at some time in the system of the vector in order to neutralize the effects of these insecticides used against them.

ABSTRACT

Insecticide resistance is a genetic characteristic involving changes in one or more insect genes. It is also a major public health challenge combating world efforts on malaria control and strategies. The Malaria vector, *Anopheles gambiae* (*A. gambiae*) has formed resistance to the existing classes of insecticides, especially pyrethroid, the only class approved for Indoor Residual Spray (IRS) and Long-Lasting Insecticide Treated Net (LLITNs). Identification of novel insecticidal targets for the development of more effective insecticides is therefore urgent. However, deciding which gene products are ideal insecticidal targets remains a difficult task in the search. To this end, it has been shown that the dissection and comprehensive studies of biochemical metabolic networks has great potential to effectively and specifically identify and extract essential enzymes as potential insecticidal targets. Using the PathoLogic programme, AnoCyc, a pathway/genome database (PGDB) for *A. gambiae* AgamP3 was constructed, using its annotated genomic sequence and other annotated information from ANOBASE, VECTORBASE, UNIPROT and KEGG databases. Furthermore, additional annotations to proteins annotated as "hypothetical" was gathered using specifically two annotation tools from the DKFZ HUSAR open servers, namely GOPET and DomainSweep and present a more comprehensive annotated PGDB for *A. gambiae* AgamP3. The resulting PGDB for *A. gambiae* AgamP3 has been deployed under the www.bioCyc.org databases. Next, a graph based model that analyzed the topology of the metabolic network of *Anopheles gambiae* was developed to determine the essential enzymatic reactions in the networks. A refined list of 61 new potential insecticidal candidate targets was obtained, which include one clinically validated insecticidal target and host of others with biological evidence in the literature. Finally, the biochemical network of *A. gambiae* was overlaid with two gene expression data obtained from the treatment of *A. gambiae* with pyrethroid (permethrin) to elucidate some tightly linked resistance genes and deduce computationally, for the first time, its resistance mechanism(s) toward this insecticide.