A Functional Workbench for *Anopheles gambiae* Micro Array Analysis

[1]Marion Adebiyi, [1]Josiah Oghuan, [1]Segun Fatumo,
[1]Ezekiel Adebiyi
[1]Department of Computer and Information Sciences
Covenant University
Ota, Nigeria
ezekiel.adebiyi@covenantuniversity.edu.ng

[2]Jason Rasgon
[2]Department of Entomology,
The Center for Infectious Disease Dynamics, and the
Huck, Institutes of the Life Sciences,
The Pennsylvania State University, University Park, PA,
1680

Abstract— **Insecticide resistance, a character inherited that encompasses alteration in one or more of insect's genes is now a major public health challenge combating world efforts on malaria control strategies. *Anopheles* has developed heavy resistance to pyrethroids, the only World Health Organization (WHO) recommended class for Indoor Residual Spray (IRS) and Long-Lasting Insecticide Treated Nets (LLITNs) through P450 pathways. We used the biochemical network of *Anopheles gambiae* (henceforth *Ag*) to deduce its resistance mechanism(s) using two expression data (when *Ag* is treated with pyrethroid and when controlled). The employed computational techniques are accessible by a robust, multi-faceted and friendly automated graphic user interface (GUI) tagged 'workbench' with JavaFX Scenebuilder. In this work, we introduced a computational platform to determine and also elucidate for the first time resistance mechanism to a commonly used class of insecticide, Pyrethroid. Significantly, our work is the first computational work to identify genes associated or involved in the efflux system in *Ag* and as a resistance mechanism in the A*nopheles.***

Keywords*: *Anopheles gambiae, biochemical network, Microarray data, resistance mechanism, and Features extraction.

I. INTRODUCTION

Anopheles gambiae (henceforth, Ag) is a complex of at least seven morphologically indistinguishable species of mosquitoes in the genus *Anopheles* [4]. The insect was first identified in the 1960s as the principal vector (carrier) of the most dangerous species of malaria parasite *Plasmodium falciparum* (henceforth, *Pf*) in Africa.

The intensive exposure to insecticides has inevitably resulted in the evolution of insecticide resistance in the *Ag*. It has formed resistance to all the recommended classes of insecticides, especially pyrethroid, the only class recommended for Indoor Residual Spray (IRS), the proven Insecticide Treated bed Nets (ITNs), the accepted and the only mainstay of the malaria control strategies [1]. The level and mechanisms of resistance deployed by this malaria vector varies, depending on the class of insecticides used and the nature of alteration in the system of the insect. Since the release of the genome sequence of the malaria mosquito, *Ag* [5, 6], several works had attempted to use genomics to explain the mode of its resistance [7 - 10], a very important problem in malaria research. Functionally, meaningful

patterns can be derived if genes are reasonably combined using knowledge about their functional interactions [11-13]. Besides, biochemical research has elucidated an increasingly better image of the metabolic architectures of *Anopheles,* (see www.biocyc.org, www.anobase.org, and www.sciencemag.org).

Thus, in this work, using the metabolic network data of *Ag AgamP3* from AnoCyc database, http://www.biocyc.org, biochemical reactions are functionally linked together by the re-construction of the metabolic network, where each gene was mapped onto its corresponding enzyme and reactions. Such interaction knowledge has been used to support the regulatory grouping of genes with gene expression profiles of yeast [14] based on the reconstructed network, cocktails of computational techniques were combined to extract some tightly linked genes that are predicted to have aided *Ag* to flush out pyrethroid speedily from its system. In other words, these genes participate in the efflux system of the *Ag*, a serious mechanism deployed by vectors to resist insecticidal actions.

.

II. REVIEW OF LITERATURE

Current work in this area is still on-going and both experimentalists and computational scientists in the field have proposed different methods and approaches on how this can be achieved [13, 15, 16]. For instance;

Potts spin clustering technique was deployed to investigate the network of a less complex organism, *E.coli* and discover its fragmentations when the organism is treated, e.g. with drugs or starved or if it is mutated [17]. They represented a metabolic network as a bipartite graph, consisting of alternating enzymes and metabolites. The network was investigated by a nearest neighbor based clustering algorithm. Their algorithm was based on the potts spin model, a method adopted from statistical physics, which was purposed to model magnetic behaviour of glassy systems [18, 19].

In another work, the use of an image-like representation of the cellular metabolism was accomplished by interpreting the expression levels of the microarray data as pixel, intensities with grey values and the network topology as relationships between pixels [20]. In their method, *Haar wavelet* transform was used with image processing method

for analyzing the micro array data. This image processing methods extracted crucial features from the image.

Recently, this technology has been improved by the introduction of PathWave is a software that enables easy analysis of gene expression data with the topology of metabolic pathways of organisms. This extensively simplifies interpretation of results [21]. PathWave improved wavelet transforms and compliments normal enrichment test to detect functionally related regulation patterns.

It is important to point out that the most current method described in c above is not applicable to the available pamethrin-induced gene expression data of *Ag*. This expression data is a single time-point exposure of 30 minutes for the treated samples [2, 23].

III. MATERIALS AND METHODS

The metabolic network data of *Ag* was extracted from the flatfiles of the AnoCyc from http://www.biocyc.org. AnoCyc is a pathway/ genome database (PGDB) of the malaria vector *Ag* that contains information of both the pathway and that of the whole genome. Such information includes; Genes, Pathways, Enzymes, Reaction, Metabolites/Compounds, Proteins and so on.

The profiling data used in this work is the gene expression data of *Ag* when treated with Permethrin [2, 23].

A. Normalizing Microarray Data

The microarray data of *Ag* was normalized using the VSN normalization package for R 2.3.0, available from the CRAN repository [http://www.r-project.org]. This is part of the Bioconductor project [http://www.bioconductor.org]. The top 100 differentially expressed genes was extracted using R script. Other vital contents obtained alongside include the corresponding p-values, which has been adjusted to control the false discovery rate. The logFC column gives the value of the contrast; which represents a log2-fold change between the two experimental conditions (treated and control). t is the empirical Bayes moderated t-statistics and B is the empirical Bayes log odds of differential expression. Every blank and empty probe sets were removed since they will not be useful in the analysis. Lastly, we averaged the values for probe sets that represented the same gene or appeared more than once.

B. Construction of the metabolic network from AnoCyc obtained from BioCyc [www.biocyc.org]

AnoCyc is a flat file description of the various reactions that make up the static image of *Ag*. It described logically how various metabolite compounds of *Ag* interact via various reactions catalyzed by a variety of enzymes that are produced by different genes.

From AnoCyc we have 206 pathways, 1402 reactions and 1608 metabolites. All metabolites and reactions are taken as nodes. If two metabolites are connected by an enzymatic reaction, then we have two edges. The incoming edge is an educt and the outgoing a product [20]. We assigned every metabolite an ID and also every reaction. So we have a list of metabolites as well as of reactions and the edges describing the input and output compounds (metabolites) of every reaction. We discarded some metabolites such as ATP for example, as these metabolites are needed in a lot of reactions (hubs). The reconstructed network has a total of 1328 metabolites and 1215 reactions.

C. Model for Analyzing Gene Expression Data on Metabolic Network

Computational steps from previous work [17-20] were employed to analyse the gene expression data mapped to the metabolic networks of *Ag*. The model used for this analysis explored a feature extraction technique based on a combinatorial approach.

To analyze gene expression data on the metabolic network of *Ag,* we designed and developed a user interface and embedded the adopted computational pipeline.

D. The GUI Platform

Three steps make up the adopted computational pipeline.
- Network clustering using a simulated annealing clustering procedure [25]
- Mapping gene expression data onto reactions
- Feature extraction using a combinatorial approach

E. Design of the GUI Workbench

The GUI was designed with JavaFX Scenebuilder Version 2.1 with NetBeans IDE 7.2.1 built on Java Development Kit (JKD-6) for Windows. The GUI multi functional workbench is a system embedded with technique for executing 3 major tasks of our computational pipeline when in use. The first pipeline it executes at the click of 'Begin Analysis' is to clutch the reconstructed network into clusters of genes ranked according to the transpose vector of highly connected sub-graph. Secondly, it maps the gene expression data onto the corresponding enzymatic reaction node on the clustered network at the click of 'MAP' on the analysis screen. Finally, it executes the feature extraction step using combinatorial approach, ranking results clusters according to the highest p-values.

The welcome screen for our workbench is depicted in figure 1 below.

Figure 1.   The Splash Screen

Figure 2 is the main data analysis interface and it encapsulates the multi-faceted phases of our workbench that has capabilities for loading transcription profiling data of organism as a table on the screen, import its metabolic network and enzyme data file from system or storage device by clicking on the file tab on the screen, and mapping the 3 data contents together.
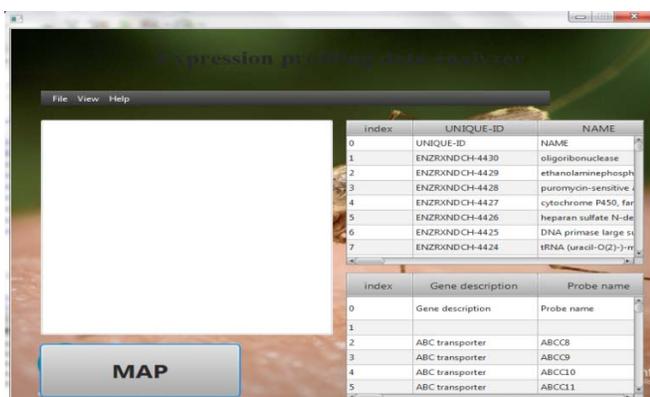
Figure 2.   Workbench in data loading stage, just before analysis begins

Features extraction is a module embedded into this interface. Features extraction is the discovery of clusters whose genes are differentially expressed within different time points of the microarray under the control condition (no insecticide) and the insecticide influenced [17]. It is also the identification of regions with a varying pattern between two different conditions. We used combinatorial approach for our feature extraction (see details in the section immediately before). The procedure has been set to consider genes with p-values <=0.001 and so such criteria were used to rank cluster output. It generates output as clusters of genes which were adequately ranked according to the highest p value. It has been established that genes of top ranking differential expressions have correlation coefficients greater than -0.5 but less than 0.6. [27]. Result pane is also localized on this page at a corner of the workbench as represented in figure 3. Thus, the end result of the whole analysis was generated and presented on this interface.
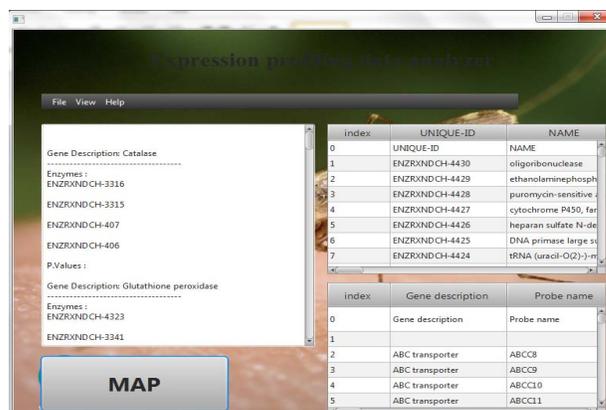
Figure 3. The workbench Interface with result in display

For all of the three steps in the computational pipeline, see [13, 22, 24, and 25]. For full details, we implemented all three steps in java in this workbench. For completion, we include some details in features extraction in the next section.

- *Features Extraction Using Combinatorial Approach*

Features extraction is the discovery of clusters whose genes are differentially expressed, under the control condition (no insecticide) and the insecticide treated condition. In the combinatorial approach we applied, all possible combinations of sums and differences of expression values in each cluster were calculated. For instance, if one of our clusters has three (3) reactions and we already mapped the expression values to the corresponding genes of the reactions. Let these expression values be 1, 2, and 3. Then we will have four (4) possible combinations, namely +1+2+3, +1+2-3, +1-2+3 and +1-2-3. Next, we compared (subtracted) all combinations, the largest difference was taken as p-value and the clusters are ranked according to their p-values. The largest p-value indicates highest probability of the cluster not to be differentially expressed [11, 13].

For each cluster, all combinations were calculated as described above. Normalization was performed on the microarray data to extract a wilcoxon test p-value. This distinguished differences between the two different states (control versus insecticide treated) for each cluster. The clusters were then ranked according to the lowest p-value that was achieved for the respective clusters.

TABLE I.   CLUSTER 1 GENES, P-VALUES AND PATHWAYS

Gene ID	Coef	p- value	Pathway
COEAE1D	0.380391144	3.71E-08	Esterase
ABCC11	0.253134829	1.15E-06	ABC Transport
ABCC9	0.379594455	2.01E-06	ABC Transport
ABCC12	-0.189323085	1.56E-05	ABC Transport

COE15O	-0.218963002	2.00E-05	Esterase
JV1	0.219007436	3.89E-05	Ecdysteroid hormones

A.   Analysis of Stimulated or Repressed Clusters

Our features extraction technique in compliance with the yardstick of [27], using the Pearson positive correlation coefficient calculated to check the patterns of the values of p-values and positions of the clusters under a single time point of the microarray under the normal and treated conditions, identified clusters labeled 1, 3 and 5 to contain statistically significant differentially expressed genes. All enzymes found in clusters 3 and 5 are Cytochrome P450 monooxygenase genes. We did not give any interpretations or analysis to genes belonging to any other classes of insecticides apart from enzymes from class pyrethroid (our insecticide of interest in this work). Furthermore, we did not find any significant differential expression in clusters 2 and 4.

That makes clusters 3 and 5 our centre of attraction because it captured 5 enzymes whose encoding genes are only insect P450 genes, that has been characterized and classified as pyrethroid metabolizers. From cluster 3, we have 2 P450 genes. One of the genes has neither been characterized nor implicated in pyrethroid metabolism by any known literature till date, but known to be involved in xenobiotic detoxification [33, 34]. This gene should be subjected to further experimental analysis for tendencies of showing up as pyrethroid detoxifier and /or metabolizer.

Cluster 1 has a total of 6 genes belonging to 3 important pathways as captured by our method, see Table I. In all, there are 2 Carboxyl esterase genes belonging to Organophosphate, 3 ABC transporter genes belonging to ATP binding cassette super family, a eukaryotic cell specific gene belonging to globular protein family and 1 Superoxide dismutase gene belonging to Organophosphate.

The 3 genes captured in cluster 5 are all cytochrome P450s genes, a pathway that has been known to produce cholesterol that synthesizes sodium cholate in the liver of A*nopheles*, and facilitates digestion of triacylglycerols [29, 36], a major component of pyrethroid class of insecticide. The over expression of the genes in cluster 5 demonstrates that they are all enzymes that may metabolize pyrethroid. A close result and conclusion was reached by experimental analysis, which revealed that one of the most consistently over expressed probe sets in DDT-resistant samples encoded for CYP6M2 (one of the enzymes implicated in cluster 5 of our result), same enzyme that has previously been implicated in pyrethroid resistance in *Ag* [2, 3, 22 and 29). Functional validation also confirmed that this enzyme can metabolize both DDT and pyrethroid insecticides [31]. P450 enzymes from these clades have been linked to insecticide resistance and metabolism of natural products and xenobiotics [31].

Enzymes encoding P450 genes captured in clusters 3 and 5 are all CYP6 genes. CYP6 genes are the first genes of this clade to be cloned and characterized from insects. Asides,

CYP6 genes belong to the CYP3 family of insect P450 clade [31]. Genes in this category outnumber any other type of genes among insect P450 genes. Validated evidences proved that members of this clade of genes are greatly involved in xenobiotic metabolism and insecticide resistance [33]. They are also inducible by pesticides and natural products [28].

IV.   RESULTS AND DISCUSSION

Our metabolic network construction resulted into 1328 metabolites, 1215 reactions with each reaction considered to be reversible.

TABLE II.   CLUSTER3 GENES, P-VALUES AND PATHWAYS

Gene ID	Coef	P. value	Pathway
CYP12F4	-0.245259084	3.67E-06	Cytochrome P450 monooxygenase
CYP6Z3	0.231102388	7.56E-05	Cytochrome P450 monooxygenase

Overall, we have an output of five (5) clusters. There were no repressed clusters amongst them all. Clusters 2 and 4 products are insignificant. Clusters 1, 3 and 5 are all stimulated clusters. 7 genes (ABCC9, ABCC11, ABCC12, COEAE1D, COE15O, and JV1) were implicated in cluster 1 as shown in Table I above, 2 genes (CYP12F4, CYP6Z3) in cluster 3 (Table II) and 3 genes (CYP6P3, CYP6Z2, CYP6M2) in cluster 5 (Table III). Every gene captured in cluster 1 belongs to pathways that control range of biological functions including molting, behavior, and response to stimuli, the development of the nervous system, reproduction and production of sex pheromones [27, 34].

TABLE III.   CLUSTER 5 GENES, P-VALUES AND PATHWAYS

Gene ID	Coef	P. value	Pathway	Pathways of actions
CYP6P3	0.682940705	6.69E-19	CytochromeP450 P450 monooxygenase	Cytochrome P450 monooxygenase
CYP6Z2	0.371302073	6.28E-07	CytochromeP450 P450 monooxygenase	Cytochrome P450 monooxygenase
CYP6M2	-0.273416045	1.5617E-04	CytochromeP450 P450 monooxygenase	Cytochrome P450 monooxygenase

Genes captured in clusters 3 and 5 specifically are Cytochrome P450 monooxygenase genes which all play a role in xenobiotic detoxification, they are also an important metabolic pathway and validated mechanism by which several insect species become resistant to insecticides [34].

Also, over expression of these validated permethrin metabolizers are participants in resistance phenotype of *Ag*. population. This may be due to their reductase effect as confirmed also in [31, 34]. One of the outstanding significantly over expressed gene from our result has been experimentally validated to metabolize pyrethroid [23]. The other genes captured have also been proved to enhance metabolic detoxification in *A. gambiae* [2, 3 and 34].

The p-values indication from our normalization in this work may suggest some capability of genes in cluster 3 and 5 to catalyze pyrethroid. Thus, we hypothesize that over expression of these validated permethrin metabolizers are participants in resistance phenotype of *Ag* population. This was also found previously by experimental means (Mitchell *et al*., 2012). Our findings also show that more than one gene that are systematically linked (network of multiple genes) contributed to Pyrethroid resistance in *Ag*. Obviously, there are multiple metabolic resistance mechanisms deployed by *Ag* population to resist insecticides. Finally, we suspect that the P450s genes captured in clusters 3 and 5 aids *Ag* to flush out pyrethroid speedily from its system. In other words, these genes participate in the efflux system of the *Anopheles,* a serious mechanism deployed by vectors to resist insecticidal actions.

V. CONCLUSION

In this work, we have used cuisine of computational tools and platforms to contribute to furthering research in combating malaria management and control. For instance, our work introduced a more annotated biochemical network for the most deadly malaria vector, *Anopheles gambiae.* We also introduced a computational platform to determine and also elucidate for the first time resistance mechanism to a commonly used class of insecticides, Pyrethroid. Significantly, our work is the first computational work to identify genes associated or involved in the efflux system in *Ag* and as a resistance mechanism in the *Anopheles.* Lastly, we also introduced a multi-functional GUI array analyzer that is capable of analyzing microarray data for other vectors and organisms. Our results also suggest which genes can be inhibited during insecticidal target discovery exploits *in silico*.

REFERENCES

[1] World Health Organization, "*World Malaria Report,"* Geneva, Switzerland: WHO Press; 2010 [http://whqlibdoc.who.int/publications/2010/9789241564106_eng.pdf]

[2] P. Müller, M. Donnelly, H. Ranson, "Transcription profiling of a recently colonized pyrethroid resistant Anopheles gambiae strain from Ghana," BMC Genomics, Vol. 8:36, 2007.

[3] R. Djouaka et al., "Expression of the cytochrome P450s, CYP6P3 and CYP6M2 are significantly elevated in multiple pyrethroid resistant populations of Anopheles gambiae s.s from Southern Benin and Nigeria," BMC Genomics, 9:538, 2008.

[4] G. M. Giles, "A handbook of the gnats or mosquitoes giving the anatomy and life history of the Culicidae together with descriptions of all species noticed up to the present date," John Bale, Sons & Danielsson, Limited. London, United Kingdom. 530pp 1902.

[5] R. Holt et al., "The Genome Sequence of the Malaria Mosquito *Anopheles gambiae,"* Journal of Science, Vol. 298, pp.129 – 149, 2002.

[6] R. Lyne et al., "An integrated database for *Drosophila* and *Anopheles* genomics," Genome Biology, Vol.8, No.7, pp. 1- 16, 2007.

[7] H. Ranson et al., "Evolution of super-gene families associated with insecticide resistance," Journal of Science, Vol. 298, pp.179-181, 2002.

[8] J. Vontas et al., "Gene expression in insecticide resistant and susceptible *Anopheles gambiae* strains constitutively or after insecticide exposure," Insect Molecular Biology, Vol.14, No.5, pp. 509–521, 2005.

.[9] P. Daborn et al., "A single P450 allele associated with insecticide resistance in *Drosophila,"* Journal of Science, Vol. 297, pp.2253-2256, 2002.

[10] J. Oakeshott, I. Horne, T. Sutherland and R. Russell, "The genomics of insecticide resistance," Genome Biology, Vol. 4, pp.1-4, 2003.

[11] M. Adebiyi, S. Fatumo, and E. Adebiyi, "Computational Discovery of Drugs Resistance Mechanism(s) of the Malaria Parasite to Tetracyclines and Chloroquines," Journal of Computer Science and its Application, Vol. 17, No 2, 2009.

[12] E. Adebiyi, "Code Malaria: Eradication development for the decade," Covenant University Press, Ota. ISSN 2006-0327, 2011.

[13] S. Fatumo, M. Adebiyi and E. Adebiyi, "In-Silico Models for Drug Resistance. "In-Silico Model for Drug Discovery," Edited by Sandhya Kortagere. "Methods in Mole cular Biology", Vol. 993, pp. 39-66, 2013. http://www.springer.com/biomed/pharmaceutical+science/book/978-1-62703-341-1.

[14] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer , "Co-clustering of biological networks and gene expression data,". Bioinformatics, vol. l, No. 8, pp. 145-154, 2002.

[15] Y. Deville, D. Gilbert, J. Helden and S. Wodak, "An overview of data models for the analysis of biochemical pathways," Briefings in Bioinformatics, Vol. 4, No. 3, pp. 246–259, 2003.

[16] S. Shirinivas, S.Vetrivel and N. Elango, "Applications of Graph Theory In Computer Science, An Overview," International Journal of Engineering Science and Technology, Vol. 2 No. 9, pp. 4610-4621, 2010.

[17] R. Koenig and R. Eils, "Gene expression analysis on biochemical networks using the Potts spin model," Bioinformatics, Vol. 20, pp.1500-1505, 2004.

[18] M. Blatt, S. Wiseman and E. Domany, "Data clustering using a model granular magnet," Neural Comput*ing*, vol. **9**, 1805–1850, 1997.

[19] E. Domany, "Super-paramagnetic clustering of data—the definitive solution of an ill-posed problem,". *Physica A*, Vol. 263, 158–169, 1999.

[20] R. Koenig et al., "Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms," BMC Bioinformatics, Vol.7, No.119 pp.1-14, 2006.

[21] *G.* Schramm *et al.,* " PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways,**"**. Oxford University Press, *Vol. 26, No. 9. pp. 1225–1231 doi:10.1093/bioinformatics/btq113,* 2010.

[22] E. Adebiyi and O. Ogunlana, "AnoCyc: Introducing an Automated Biochemical Metabolic Network for the malaria vector, *Anopheles gambiae,"* 2012, www.biocyc.org. –

[23] P. Müller et al.," Field-caught permethrin-resistant Anopheles gambiae overexpress CYP6P3, a P450 that metabolises pyrethroids,". PLoS Genet, Vol. 4, 2008.

[24] S. Fatumo et al., "In silico knock-out screening of the metabolic network of Plasmodium falciparum to yield potential drug targets," Proceedings of the DKFZ structural and functional genomics retreat, Heilbronn, Germany, 2006.

[25] D. E. Brown and C. L. Huntley, "A practical application of simulated annealing to clustering," Technical Report, IPC-TR-91-

003, Institute for Parallel Computing, University of Virginia, Charlottesville, VA, 1991.

[26] L. Nardini et al., "Detoxification enzymes associated with insecticide resistance in laboratory strains of Anopheles arabiensis of different geographic origin," *Parasites & Vector, vol.* **5, No.** 113, 2012. doi:10.1186/1756-3305-5-113.

[27] J. Cohen, "Statistical power analysis for the behavioral sciences," Hillsdale, NJ: Lawrence Erlbaum Associates, (2nd ed.) ISBN 0-8058-0283-5, 1988.

[28] R. Feyereisen, "Comprehensive Molecular Insect Science," (Gilbert, L.I., Iatrou, K. and Gill, S.S., eds.), Elsevier, Oxford. Vol. 4. pp. 1–77, 2005.

[29] P. Champe, R. Harvey, D. Ferrier, "Cholesterol and Steroid Metabolism," Biochemistry (Lippincott's Illustrated Reviews), eds Harvey RA, Champe PC (Lippincott Williams & Wilkins, Baltimore), 3rd Ed, pp 222–223, 2005.

[30] B. Stevenson et al., "Cytochrome P450 6M2 from the malaria vector Anopheles gambiae metabolizes pyrethroids: Sequential metabolism of deltamethrin revealed," Insect Biochem & Mol. Biol., Vol. 41, pp. 492–502, 2011.

[31] S. Mitchell et al., "Identification and validation of a gene causing cross-resistance between insecticide classes in *Anopheles gambiae* from Ghana," Genetics, 2012. www.pnas.org/cgi/doi/10.1073/pnas.1203452109.

[32] R. Feyereisen, "Evolution of Insects P450," Vol. 34, Biochemical Society Transactions, pp.1252-1255, 2006.

[33] C. McLaughlin et al., "Characterization of *Anopheles gambiae* CYP6Z2: role in xenobiotic metabolism and insecticide resistance," Insect Mol. Biol., Vol. 17, pp. 125–135, 2008.

[34] J. G. Scott, "Cytochromes P450 and insecticide resistance," *Insect Biochem* Mol Biol, 29(9):757-777, 1999,