Predicting the structure of *Anopheles gambiae* Cytochrome P450 protein using computational methods

Trust Odia, Marion Adebiyi
Department of Computer and Information Sciences,
Covenant University,
Ota, Nigeria.
trust.odia@covenantuniversity.edu.ng, marion.adebiyi@covenantuniversity.edu.ng

Abstract*—the CYP12F4 protein is a member of the Cytochrome P450 super-family of monooxygenanses, a large and diverse group of enzymes that catalyzes the oxidation of organic substances and metabolic reactions. It is found in the female African mosquito *Anopheles gambiae (A. gambiae)* that carries and transmits the most deadly malaria parasite, *Plasmodium falciparum (Pf)*. Presently experimental structure is not available for this protein; it has thus remained uncharacterized with unknown function. This work employs *in-silico* methods to predict the structure of this metabolic catalyzer and further deduced a specific function for the same protein. Using 6 template proteins, 29 residues were modeled with homology. Several web servers were deployed to predict a computational model for CYP12F4. GOPET web tool was finally used to deduce the unique function of this protein. The folds were identified and analyzed and the protein was specifically found to be active in binding of molecules with 86% confidence value with various catalysis activities. 21 helices, 6 strands, 51 beta turn and 348 hydrogen bonds were elucidated and analyzed on the structure. Several literatures have confirmed these findings. CYP12F4 is a heme and iron ion binding protein that carries heme and catalyses the incorporation of one atom from molecular oxygen into a compound and reduces the other atom of oxygen to water. A deep understanding of these properties of heme and its binding with respect to CYP12F4 protein is vital in malaria control.

***Index Terms*—protein structure prediction, homology, protein structure, protein function, template, cytochrome P450**

I. Introduction

Anopheles gambiae is the female mosquito of the *gambiae* specie that carries human malaria parasite *P.falciparum*. The following characteristics makes *Anopheles gambiae* the primary vector of malaria, they include: rapidly colonizing small pools of rain water, harbouring the parasite *P.falciparum* in its body under a wide range of environmental conditions, its acute anthropophilic nature, feeding on humans and resting indoors i.e about 95% indoor resting catch in Kenya [9]. The parasite *P.falciparum* is transmitted to humans through a bite from the vector *A. gambiae.* This mosquito is highly dominant in tropical African regions due to the environmental conditions which favours its reproduction and survival.

Malaria is the disease caused by the parasite *P.falciparum*, carried by *A. gambiae*. According to (W.H.O, 2014) there were 198 million malaria cases and an estimated 584,000 malaria deaths, affecting mostly children under the age of 10. Pregnant women are also at risk of this disease when they are immune-suppressed and this causes miscarriage, still birth and maternal death in pregnant women. Most of the feeding by the mosquito is done at night and dawn, when the host is at rest. Anti-malaria strategies like insecticides have been in place for years till date to control the transmission of this disease but haven't been efficient. The mosquito has built resistance over time to these insecticides.

A.gambiae is made up of several proteins which include cytochrome P450 sub-family of monoxygenases, which are oxidoreductase enzymes that catalyze the oxidation of organic substances and metabolic reactions of endo and exogeneous compounds. Cytochrome P450s are hemoproteins, they partake in the synthesis and degradation of the mosquito's hormones and pheromones. *A. gambiae* P450 are microsomal enzymes, thus require flavoprotein NADPH cytochrome P450 reductase as electron donor. *A. gambiae* cytochrome P450s have been implicated in insecticide resistance and malaria infection [4].

The CYP12F4 belongs to a sub-family (F4) in the large family of cytochrome P450s and is found in the *Anopheles gambiae* mosquito. Presently experimental structure for this protein has not been resolved, which hence little knowledge about its function and activities in *A. gambiae* is exists. According to [4], CYP12F4 a mitochondrial cytochrome P450 has been implicated in response to malaria infection, but its function and activities remain unclear. CYP12F4 is an enzyme and it has been found in metabolizing deltamethrin which is used in the production of insecticide-treated bednets and also implicated in xenobiotic detoxification [2]. The molecular function and biological process of a protein can be derived from its structure. Protein structure prediction deals with resolving the structure of the target protein. The structure of a protein can be generated experimentally (NMR, X-ray crystallography and Cryo-Em) or computationally (ab initio, fold recognition and homology modeling). Computationally predicted structures exits for the target protein CYP12F4, but of low structural quality and functional relationship.

The objective of this paper is to predict the 3D structure of the target protein CYP12F4 using computational methods and show the structural properties and features of this enzyme in confirmation to its implicated activities, role, molecular function and biological process in *A. gambiae*.

II. RELATED WORK

Protein structure and function prediction is a very crucial study in cell biology, proteomic and drug therapeutics as well. Research has been carried out in this field from literature and has produced good results. A very recent work from [27] showed the use of phylogeny (evolutionary relationship to analyze all CYPs in Tribolium Castaneum with genes in other insect species to deduce genetic evolution and function of T. Castaneum CYP gene family. The integrated use of annotations, molecular modeling, docking, phylogenetic analysis, gene expression revealed 143 CYPS in *T. casteneum* which may contribute to insecticide resistance in beetle. Their work also provided insights into the evolution of T. Castaneum *CYP* gene superfamily and developed a valuable resource for the functional genomics research necessary for understanding the strategies employed by insects in coping with their environment and to harness potential insecticides targets for pest control. Homology modeling was also used by [16] to model three CYPs (CYP6AA3, CYP6P7 and CYP6P8) implicated in insecticides. Modeling of these structures showed better understanding about different substrate preferences among the enzymes, variations among predicted substrates channels and geometry of active site was the reason behind their pyrethroids binding differences (i.e. differences in their active sites structure may impact substrate binding and selectivity). Their result showed that the differences in metabolic activities in insect P450 can attribute to structural differences responsible for selectivity in their activities against insecticides. They concluded by saying that the predicted models may be used to explore target P450 inhibitors and in the analysis of the binding and metabolism of insecticide compounds that have potential for use in the control of *A. gambiae*. Similar study [22] involved the use of homology modeling (comparative modeling) to infer the structure of a cytochrome P450 enzyme (AnCYPOR) with rat as template (CYPOR). Detailed analysis revealed major differences in FMN-and FAD/ NAD(P) H binding domains that might lead to (differences in enzymatic properties and catalysis of mosquito CYPOR from mammalian CYPOR (rat), also mutagenesis study showed that C427 supports FAD binding in AnCYPOR and that NAD(P) H binding and catalysis differs from mosquito and rat. Computational approaches have proved to be faster, cost-effective and efficient than experimental methods from these related works.

Other approaches not peculiar to structure protein structure prediction have been used for in-depth analysis of the target cytochrome P450 protein at metabolic and catalytic level. The combination of molecular modeling and quantitative structure-activity relationship (QSAR) study by [15] was used to understand the factors that determine substrate selectivity and binding to the human drug metabolizing P450s. Detailed review by [4] expatiated on the general structure of P450 Cytochromes, role of Cytochrome P450s in *A. gambiae* as detoxifying enzymes, highlighting the link between *A.gambiae* P450 cytochrome and insecticide resistance, response to malaria infection in *P. berghi* and *P. falciparum* invasion [4]. Further details showed that effect of chloroquine (in the abundance of transcripts responsible for encoding proteins involved in a variety of processes) including P450 cytochromes (CYP9L1, CYP304B1 and CYP305A1) expressed differently in a blood meal containing *P. berghi.* In relation to insecticides resistance Pyrethroid resistance has been detected in the *A. gambiae* [16] due to a combination of target site insensitivity and increased oxidative metabolism, catalyzed by P450s cytochrome.

III. MATERIALS AND METHODS

The workflow for this paper is shown below in Fig. 1, the protein sequence of the target was downloaded from Uniprot web-server in FASTA format.

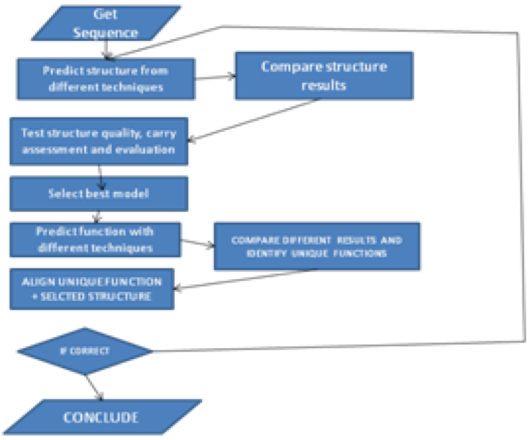
Get Sequence
Predict structure from different techniques
Compare structure results
Test structure quality, carry assessment and evaluation
Select best model
Predict function with different techniques
COMPARE DIFFERENT RESULTS AND IDENTIFY UNIQUE FUNCTIONS
ALIGN UNIQUE FUNCTION + SELCTED STRUCTURE
IF CORRECT
CONCLUDE

Fig. 1. research workflow

A BLAST was done on NCBI web server with the target protein against protein databank server to identify a closely related i.e. homologous protein (template). The template that was identified was a 3K9V (crystal structure of rat mitochondrial P450 24A1 S57D in complex with CHAPS). SWISS-MODEL web server was used with the template to predict the structure on target-template mode [18]. SWISS-MODEL uses homology modeling, thus Phyre2 was also used with the same identified template to predict a structure, which uses ab initio technique [11]. I-TASSER a hierarchical method that involves protein threading was used to predict a structure also [17]. A selected template from the BLAST was used to predict the final model based on the accuracy and reliability of the BLAST algorithms at NCBI. The target sequence was aligned against the structure (secondary and tertiary) of the template to produce a model. An automatically selected template by these tools was also used to compare the selected template (3k9v) as proof. Basically the tools used in predicting the structure of the target protein were selected from the top 5 ranked web servers on Critical Assessment of Techniques for protein Structure Prediction (CASP 10) website for high performance, these includes Phyre2, I-TASSER and SWISS-MODEL.

The structures generated from these different tools were compared with PDBSUM a webserver at EBI website. PDBSUM counts the amount of alpha helices, beta sheets, turns; motifs present in the secondary structure of model and it gives a summary of the key information of macromolecular structures. [14]. JMOL is a protein structure visualization software; it was used to visualize the generated structures. The structures were accessed using PROCHECK a tool under PDBSUM, which checks the stereochemical quality of the structures by generating a ramachandran plot of the torsion phi and psi angles of the amino acid residues in the structure [18]. Ramachandran plot is a way of visualizing the backbone structure of polypeptides, using datapoints in a graph as shown in figure 2c. Ramachandran plot calculates the percentage of amino acid residues in the favoured and allowed region which is at the top right and disallowed region where outliers are located, which is bottom right. Residues that are functionally relevant to the structure must lie in the allowed regions. Ramachandran plot was used as criteria for evaluating the generated structures and the structure with the highest percentage of amino acids concentration at the allowed region was selected as the best model [9].

Protein function was predicted with GOPET, it provides biological process and molecular functional annotation based on Gene Ontology, performs homology searches on GO-mapped protein databases and predictions with confidence values through SVM [21]. The predicted functions were checked on QuickGO at EBI for correctness. I-TASSER was used along with GOPET to predict functions for the target protein. I-TASSER predicts functions by threading 3D models through BioLiP database and other state of the art algorithms [17]. From the various predicted functions, the unique molecular function of CYP12F4 was selected and it was heme binding at 86% confidence level on GOPET. The features and characteristics of CYP12F4 which is substrate binding site, cytochrome reductase/cyp b5 binding site, heme binding site and NADPH-binding site were identified in the predicted structure. Structural alignment was carried out on the predicted model; 3D-BLAST [20] at BioXGEM was used to search for related structures and the molecular function of the most related structure was compared with that of the CYP12F4 (predicted function). This was used as a validation that the predicted function of CYP12F4 was correct, the related structure was the crystal structure of rat mitochondrial P450 24A1 S57D in complex with CYMAL-5. The structural alignment of the target protein model and that of the crystal structure of rat mitochondrial P450 24A1 S57D chain B was at 0.64A RMSD, 7.9 Z-score, aligned length was 422 and sequence identity was 33.2%. The related structure and the target protein model were later aligned together for more structural analysis with Dalilite at EBI web server.

IV. RESULTS

The final selected model is shown in figure 2a below gotten from SWISS-MODEL. It shows 2 sheets, 2 beta hairpins, 1 psi loop, 1 beta bulge, 6 strands, 21 helices, 32 helix-helix interacs, 51 beta turns, 2 gamma turns.

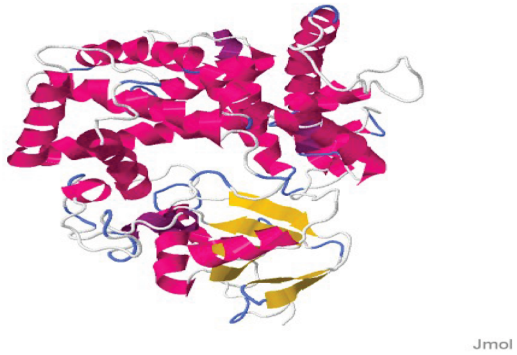

Fig. 2. final model

Alpha helices are in pink colour, sheets are in yellow. The ligand binding site was predicted with 3D-ligand binding site, it shows the. The ramachandran plot by PROCHECK is shown in figure 2c below. 88.2% of the 343 amino acids residues in the final model lie in the most favoured regions of the plot, labeled A,B,L at positions 90 to 180, 0 to -90 and 0 to 90 (psi degree) and -45 to -180, -45 to -180 and 90 to 45(Phi degree) respectively. 9.8% of 38 amino acids residues lie in the additional allowed regions, labeled a, b, l and p. 1.3% of 5 amino acid residues lie in the generously allowed regions and 0.8% of 3 amino acid residues lie in the disallowed regions labeled XX. 33 Glycine residues, 29 Proline residues and 389 non-glycine and non-proline residues making a total of 453 residues in the final structure.

The ligand binding site of the final model is shown in grey colour in Fig. 3. below.

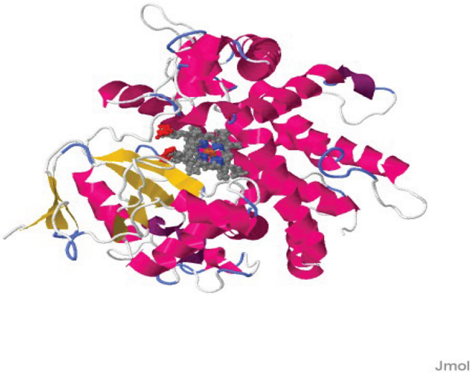

Fig. 3. Ligand binding site

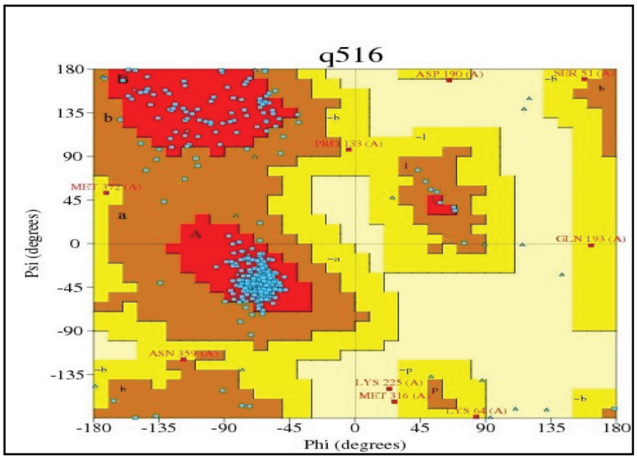

Fig. 4. final model ramachandran plot

GOPET predicted heme binding 86% confidence level, catalytic activity 83% confidence level (molecular function) and QuickGO revealed biochemical process of oxidation-reduction, molecular function of monooxygenase activity, iron ion binding, oxidoreductase activity acting on pair donor, with incorporation of reduction of molecular oxygen and electron carrier activity.

The heme binding site of the final structure showing the iron ion was predicted by 3DLigandSite, it shows the iron ion binding site in Fig. 5. below and the iron in orange colour.

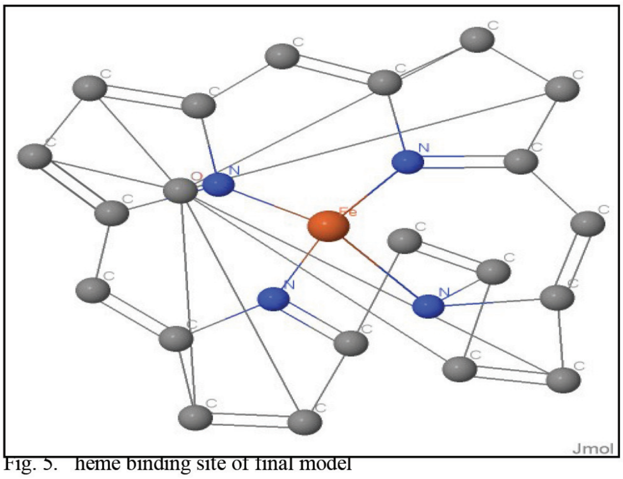

Fig. 5. heme binding site of final model

The structural alignment of molA of final model and molB of related structure as displayed by Dalilite is shown below in figure 2 (e) with Z-score of 51.6, 422 number of equivalent residues and 0.7 RMSD.

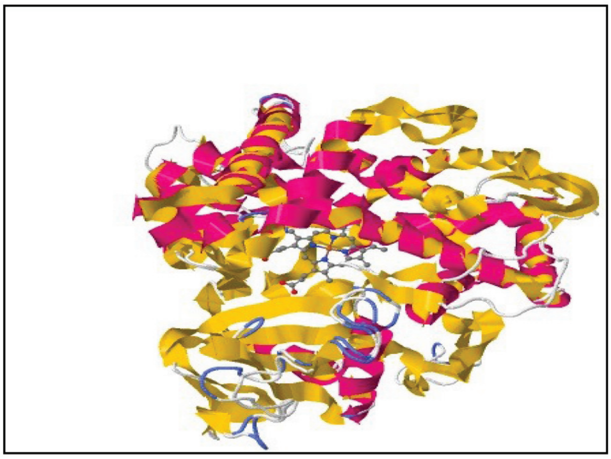

Fig. 6. structural alignment

The pink colour represents the final model and the yellow colour represents the related structure (rat mitochondrial P450 24A1 S57D in complex with CYMAL-5). The heme binding site is visible in Fig. 6. above gotten from 3D-BLAST at BioXGEM.

V. DISCUSSION

The structural alignment between the final model and the related structure confirms that the chain A of the predicted structure of the target protein is related to the experimental chain B structure of rat mitochondrial P450 24A1 S57D in complex with CYMAL-5 structurally and functionally. Both structures (chain A and B) have the same molecular function which is monooxygenase activity, iron ion binding, heme binding, oxidoreductase activity to name the top 4. The heme binding function relates to the heme binding site of CYP12F4 predicted structure in Fig. 5. and the heme binding characteristics of cytochrome P450s. From literature unique features of the CYP12F4 protein which includes its substrate binding sites (at the F, R, V, I, 400 position), cytochrome reductase/cyp b5 binding site, heme binding site (Ligand Heme name: HEME B, PROTOPORPHYRIN IX CONTAINING FE with chemical formular $C_{34}H_{32}FeN_4O_4$) and NADPH-binding site which is the source of electron transfer [8]. These are peculiarities to the family of cytochrome P450s. CYPs have shown resistance to organic compounds like insecticides and sub-family members of CYP12F4 have been implicated in this [4]. The most highly conserved regions among P450s lie between the I and L helices, and are involved in heme binding. The heme of P450s is covalently bound to an invariant cysteine, which is enveloped by a β–bulge region called Cys-pocket (located at position 457 to 470) and meander loop located from position 428 to 436 from Phyre2 analysis. Three residues besides the cysteine are very strictly conserved, two Glycine (is in a position that allows the formation of the β–hairpin turn, it serves two roles: allowing for a sharp turn from the Cys-pocket into the L helix and for proximity to the heme) and one Phenylalanine (close to the sulfur-iron bond which is based from Human_CYP11A1 cholesterol side-chain cleavage enzyme, mitochondrial). Multiple sequence alignment between CYP12F4 and sub-family members revealed the following conserved regions 466(P) PRO 87%, 455 (P) PRO 62%, 437 (R) ARG 87%, 436 (E) GLU 62%, 470 (G) GLY 87%, 327 (T) THR 75%, 438(W) TRP 62%, 435 (P) PRO 75%, 433(F) PHE 84%, 430 (P) PRO 72%, 427(F) PHE 62%. The substrates (insecticides) bind to the substrate binding site of the P450 enzyme due to their high affinity [4].

VI. CONCLUSION

From literature sub-family members of CYP12F4 have been highly expressed in A. gambiae (1) to insecticides resistance by detoxifying the organic compounds (2) 1 day after blood meal containing P.Berghi up-regulated in fat body(3) in A. gambiae midgut due to malaria infection from P. Berghi and P. falciparum. The sub-family of cytochrome P450 (CYP12) have shown response to malaria infection by hemocytes [4]. This means that CYP12F4 seems to be expressed in response to insecticide resistance in A. gambiae. From this study, the predicted structure is hypothesized to be a possible structure for CYP12F4, but future work seeks experimental confirmation for the predicted structure of the CYP12F4 protein and its function as well. If these findings are true then strategies should be developed to suppress the express of these detoxifying proteins in order to control malaria from A. gambiae.

REFERENCES

[1] Anorld, K., Bordoli, L. Kopp, J. andSchwede, T. (2005). The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling. Oxford Journals. Science & Mathematics. Bioinformatics. Vol. 22. Issue 2. Pp 195-201

[2] Adebiyi, M. (2014). Computational analysis of *Anopheles gambiae* metabolism to facilitate insecticidal target and complex resistance mechanism discovery. Thesis.covenantuniversity.edu.ng/123456789/953.

[3] Bonneau, R. and Baker, D. (2001) Ab initio protein structure prediction: progress and prospects. Annual Reviews 30: 173-89 pg 1.

[4] Eisenhaber, F., Persson, B. and Argos, P. (1995) Protein structure prediction: recognition of primary, secondary and tertiary structural features from amino acid sequence. Pubmed, PMID 758 7278; 30(1):1

[5] Félix, RC and Silveira, H (2012). The Role of Anopheles gambiae P450 Cytochrome in Insecticide Resistance and Infection, Insecticides - Pest Engineering, Dr. Farzana Perveen (Ed.), ISBN: 978-953-307-895-3, InTech, Available from: http://www.intechopen.com/books/insecticides-pest-engineering/the-role-ofanopheles-gambiae-p450-cytochrome-in-insecticide-resistance-and-infection

[6] Froimowitz, M. and Fasman, G.D (1974) Prediction of the secondary structure of proteins using the helix-coil transition theory. Macromolecules 7(5), 539-9. PMID 4371089.

[7] Garnier J., Osgurthorpe D.J. and Robson, B. (1998). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. JMOL. Biol. 120 (1) 97-120.

[8] Hasemann, C.A.,Kurumbail, R.G., Boddupali, S.S., Peterson, J.A. and Deisenhofer, J. (1995). Structure and function of cytochrome P450: a comparative analysis of three crystal structures. Structure. Vol.3. (1). Pg 41-62.

[9] Hollingsworth, S.A. and Karplus, P.A. (2010). A fresh look at the ramachandran plot and the occurrence of standard structures in proteins. Biomol. Concepts. Vol. 1(3-4). Pg 271-283.

[10] Jannat, N. K. (2010) Effect of larval environment on some life history parameters in anopheles gambiae s.s (Diptera: Culicidae). Simon Fraser University, Library. Canada pg 3.

[11] Kelly, L. A. and Slernberg, M.J.E. (2009) Protein structure prediction on the web: a case study using the phyre server. Nature Proc 4, 363-371.

[12] Kihara, D., Zhang, Y., Lu, H., Kolinski, A. and Skolnick, J. (2002) Ab initio protein structure prediction on a genomic scale: Application to the mycoplasma genitalium genome. PNAS Vol.9 (9), pg 5993-5998.

[13] Lakizadeh, A. and Marashi, S.A (2009). Addition of contact number information can improve protein secondary structure prediction by nueral networks. Excil J. 8 pg 66-73.

[14] Laskowski, R.A. (2008). PDBsum new things. Nucleic Acids Res. D. pg 355-359.

[15] Lewis, D.F. and Ito, Y. (2009). Informa Health Care. Xenobiotica. Issuetec. Vol.39. (8). Pg 625-635.

[16] Lertkiatmongkol, P. Jewitheesuk, E. and Rongnoparut, P. (2011). Homology modeling of mosquito cytochrome p450 enzymes involved in pyrethroid metabolism: insights into differences in substrate activity. BMC Research Notes. 4(321).

[17] Mount, D.M. (2004). Bioinformatics: sequence and genome analysis2. Cold Spring Harbor Laboratory Press. ISBN 0-87969-712.

[18] Nikou, D., Ranson, H. and Hemingway, J. (2003). An adult-specific CYP6 P450 gene is overexpressed in a pyrethroid resistant strain of the malaria vector, *Anopheles gambiae.* Elsevier. Gene. Vol. 318. Pg 91-102.

[19] Roy, A., Kucukural, A. and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. NIHPA Manuscripts. Vol. 5. 4. Pg 725-738.

[20] Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology. Vol.7. Pg 95-99.

[21] Schwede, T., Kopp, J., Guex, N. and Peitsch, MC. (2003).SWISS-MODEL: an automated protein homology-modelling server. Oxford Journals. Science & Mathematics. Nucelic Acids Research. Vol. 31. Issue 13. Pp 3381-3385.

[22] Sarapusit, S. Lerkiatmongkol, P., Duangkaew, P. and Rongnoparut, P. (2013). Modelling of *Anopheles minimus* mosquito NADPH-Cytochrome P450 oxidoreductase (CYPOR) and mutagenesis analysis. Int. J. Mol. Sci. Vol. 14 (1). Pg 1788-1801.

[23] Tung, C.H., Huang, J.W. and Yang, J.M. (2007). Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search. Genome Biology. Vol.8.pg R31.1-R31.16.

[24] Vinayagam, A., Val, C., Schubert, F., Eils, R., Glatting, K.H., Suhai, S and Konig, R. (2006). GOPET: A tool for automated predictions of gene ontology terms. BMC Bioinformatics.Vol. 7. 161

[25] Wass, N.M, Kelley, L.A and Sternberg (2010). 3DLigandSite: Predicting ligand-binding sites using similar structures. Oxford Journals. Science & Mathematics. Nucleic Acids Research. Vol. 38. Issue suppl 2. pp W469-W473.

[26] Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. Vol. 9. 40

[27] Zhang, Y. (2008). Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol., 18 (3), 34 2-8 PMC 2680823 PMID 18436442.

[28] Zhou, D.O. (2006). Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins. 66 (4), 8 38-45. PMID 17177203.

[29] Zhu, F., Moural, T.W., Shah, K. and Palli, R. S. (2013). Integrated analysis of cytochrome p450 gene superfamily in the red flour beetle Tribolium casteneum. BMC genomics. 14 (173).