European Journal of Biotechnology and Bioscience

Available online at www.biotechjournal.info

E J B B

European Journal of Biotechnology and Bioscience

ISSN: 2321-9122
www.biosciencejournals.com
EJBB 2014; 2 (1): 24-31
Received: 08-7-2014
Accepted: 20-7-2014

Balarabe R. Mohammed
School of Science Engineering and Technology, Abertay University Dundee, DD1 1HG, UK

Craig S. Wilding
Vector Group, Liverpool School of Tropical Medicine, L3 5QA, Liverpool, UK

Phillip J. Collier
School of Science Engineering and Technology, Abertay University Dundee, DD1 1HG, UK

Yusuf Y. Deeni
School of Science Engineering and Technology, Abertay University Dundee, DD1 1HG, UK

Correspondence
Balarabe R. Mohammed
School of Science Engineering and Technology, Abertay University Dundee, DD1 1HG, UK

Bioinformatic analysis of regulatory elements within the promoter region of the cytochrome P450 Gene, *CYP6M2* in *Anopheles gambiae*

Balarabe R. Mohammed, Craig S. Wilding, Phillip J. Collier, Yusuf Y. Deeni

Abstract
Cytochrome P450s including *CYP6M2* have been demonstrated to be involved in the metabolism of insecticides, typically through up regulation in resistant individuals. The difference in gene expression levels seen in insecticide resistant mosquitoes may result from sequence differences in the 5' upstream region of *CYP6M2* including those in the promoter elements. Understanding the complex mechanisms regulating P450 expression, including that of *CYP6M2* in insecticide resistant *Anopheles gambiae* remains a great challenge. In this study, extensive bioinformatics resources were used to predict regulatory elements and cross-species comparison in the *cis-acting* elements within *CYP6M2* (896 bp) and *CYP6G1* (896 bp) known to be up regulated by the orthologs of Nuclear factor erythroid 2- related factor 2(*Nrf2*)/Kelch -like ECH-associated protein 1(*Keap* 1) and Aryl hydrocarbon receptor (*AhR*)/ Aryl hydrocarbon nuclear translocator (*ARNT*) in *Drosophila melanogaster*. Searches were also made for the *cis- acting* elements within a 879 bp region up stream of *CYP6M2* hypothesised to contain the promoter for this gene in both the Tiassalé multiple insecticide-resistant and Kisumu susceptible strains of *An. gambiae*. Results revealed the presence of *Nrf2*/*Keap* 1 and *AhR*/*ARNT* as putative transcription factor binding sites (TFBS) within the *CYP6M2* promoters. Further, we identified the orthologs of those transcription factors which bind to these elements (Cap 'n' collar isoform C (*CnCC*) / (*Keap* 1) & Spineless (*ss*)/ Tango in *Drosophila melanogaster*) as AGAP010295/ AGAP000748 & AGAP005300/ AGAP003645 in *Anopheles gambiae*. These data suggest the presence of putative AGAP010295/AGAP009748 and AGAP005300/ AGAP003645 binding sites in the promoter of *Anopheles gambiae CYP6M2*, which may potentially be associated with the up regulation of *CYP6M2* involved in insecticide resistance. These if established have implications in the control of malaria.

Keywords: *Anopheles gambiae*, *Drosophila melanogaster*, *CYP6M2*, *CYP6G1*, Cytochrome P450

1. Introduction

Cytochrome P450 (CYPs) are known to play a central role in insecticide resistance, allowing resistant insects to detoxify insecticides at a higher rate [1]. Among them, members of *CYP6M2* gene have been repeatedly implicated in resistance to insecticides [2].The development of insecticide resistance to all the commonly used insecticides in *Anopheles gambiae* necessitates the evolvement of these novel strategies to combat malaria in Africa [3,4]. Often through microarray analyses of resistant strains. Whilst specific genes such as *CYP6M2* and *CYP6P3* in *An. gambiae* and *CYP6P9* in *Anopheles funestus* are repeatedly identified as up regulated in resistant strains and have been demonstrated to be capable of insecticide metabolism *in vitro*, the specific sequence elements responsible for driving the up regulation and hence resistance remain to be identified. Both *cis*- elements (on the same chromosome, typically immediately upstream of the transcription start site) and *trans*- elements (distant from the gene – including on separate chromosomes) may have evolved to increase gene expression. However, identifying these is a challenging process necessitating both experimental and bioinformatics techniques. As available sequence resources increase (e.g. the sequencing of 16 *Anopheles* genomes [5] and over 1000 *An. gambiae* genomes [6] so bioinformatic methods will aid in understanding the complex mechanisms underlying gene regulation and expression in insecticide resistance [7]. Here we use the term promoter as the DNA sequences, usually between specific binding sites for proteins involved in the initiation and regulation of gene transcription [8, 9].This transcription is controlled primarily by transcription factors (TFs) which recognize and bind to specific short DNA sequence motifs [10]. The most familiar of these motifs is the TATA box (found in 30-40% of core.

Transcription factor binding leads to the activation or repression of transcription in response to changes in the environment, as well as during development [7]. Each gene contains a set of unique combination of TFBSs in the promoter that determines its temporal and spatial expression [12, 13].The TFBS only carry the potential to bind to the gene promoter. However, the TFBS binding can occur everywhere within the genome and are therefore not restricted to regulatory regions [14] .The control of gene transcription initiation is a significant mechanism for the determination of gene expression and how much mRNA and consequently protein is being produced [14]. Therefore, one major task of understanding transcriptional regulation networks is to identify all TFBSs bound by all TFs encoded in a genome, which eventually will provide the information necessary to construct models for describing transcriptional regulatory networks [15,16,8]. Although recent studies in mosquitoes revealed that, the vast majority of described promoters are from *Anopheles* and *Aedes* species [17]. No study appears to have been made in understanding the complex mechanisms regulating *CYP6M2* gene expression in insecticide resistance in *Anopheles gambiae*. In this paper therefore, as a prerequisite for understanding the molecular mechanism involved in the regulation of *CYP6M2*, characterization of its promoter sequence and searches for potential regulatory elements were made. Similar searches for the regulatory elements of 5' upstream region of *CYP6G1*, a single P450 allele responsible for insecticide resistance in *Anopheles gambiae*. This *CYP6G1* gene is also known to be up regulated by the orthologs of Nuclear factor erythroid 2- related factor 2(*Nrf2*)/Kelch -like ECH-associated protein 1(*Keap* 1) and Aryl hydrocarbon receptor (*AhR*)/ Aryl hydrocarbon nuclear translocator (*ARNT*) [18,19]**.**The recognition of these transcription factor binding sites (TFBSs) has the potential to improve knowledge on how wild populations of *Anopheles gambiae* become resistant to insecticide and are activated by different endogenous and exogenous challenges. This promoter analysis only infers the binding potential and not the functionality of a site. The functionality can only be proven through wet- laboratory experiments with predetermined parameters, particularly since a potential binding site in a promoter can be functional in certain cells and non- functional under different conditions [14].

2. Materials and Methods

2.1 Phylogenetic Tree of the *Anopheles gambiae CYP6M2* gene sequence

In order to analyse the evolutionary similarities, differences and relationships between various insect species and *Anopheles gambiae*, phylogenetic analysis was carried out comparative phylogenetic and bioinformatics enable sequence homology within several species to be identified. A phylogenetic analysis of *CYP6M2* orthologs in *Anopheles gambiae* was carried out using Ensembl Metazoa Genome Browser (http://metazoa ensembl.org/ Anopheles gambiae/Info/Index).

2.2 Genomic DNA and mRNA sequence retrieval

We used the VectorBase database to retrieve sequences from the *Anopheles gambiae* genome (http://www.vectorbase.org). The National Centre for Biotechnology information (http:// www. ncbi. nlm. nih.gov/) and the Ensembl Metazoa Genome Browser (http://www.ensembl.org/ index.html) (Ensemble Metazoa Genome Browser; National Centre for Biotechnology Information). Furthermore, 5' upstream region of (*CYP6G1* (896 bp) in *Drosophila melanogaster* was also retrieved and analysed along with the *CYP6M2* promoter sequences using the same tools to determine the degree of conservation in both promoter structures and perhaps function. In order to successfully clone the 5' upstream *CYP6M2* gene, we cloned into pJET1.2 cloning vector pJET-*CYP6M2*_Tiassalé (Tias) (930 bp) and pJET-*CYP6M2*_Kisumu (Kis) (896 bp) [20].

2.3 Analysis of *CYP6M2* gene sequence and *CYP6M2* gene promoter

In order to understand the regulatory mechanism of the *CYP6M2* gene promoter, the presence and location of transcription factor binding sites in the 5' regulatory regions of the retrieved gene sequence and promoter sequence analysis was done using several Bioinformatics tools. These sequenced constructs were further analysed for the presence of putative GATA transcription factors binding factors considered as positive cis-acting regulatory elements sequences [21, 27].

2.3.1 Recognition of promoter/ Enhancer elements within the cloned and sequenced *CYP6M2* gene promoters

In order to identify putative regulatory motifs within the cloned and sequenced *CYP6M2*_ (Kis) and *CYP6M2*_ (Tias) gene promoter sequences, several online tools were used. CPGFinder tool was used to predict the existence of cPG dinucleotides within the gene promoter sequence (cPGisland islands. USC.edu). Clusters of cPGs/GC rich regions called cPG islands, (regions containing >200 cPG sequences) are often found close to genes, with cPG sequences appearing in the promoters and first exons but also in downstream regions. These so-called cPG islands (cPG dinucleotides in invertebrates), first defined by Bird in 1986, are on average 400–500 bp of length, have a C+G content of 0.5 or higher and an observed to expected cPG ratio of 0.6 or higher within a range of 200 bp or greater [22]. Analysis of the sequenced *CYP6M2*_Kis -1 & *CYP6M2*_Kis-2 and *CYP6M2*_Tias-1 and *CYP6M2*_Tias-2 promoter using cPG island searcher (http://cPGislands.usc.edu/) [23]**.**The TSSW (Recognition of human PolII promoter region and start of transcription) tool predicts the existence of Enhancer and general transcription factor binding sites (e.g. TATA Box). Further analysis using McPromoter defined as a probabilistic promoter prediction system that identifies likely TSSs in large genomic sequences tool [11]. (McPromoter MM: II http://genes. mit.edu/ McPromoter.html).

2.3.2 Prediction of putative regulatory elements within the cloned and sequenced *CYP6M2* gene promoters

In order to predict the putative transcription factor binding sites (TFBS) within the cloned and sequenced *CYP6M2*_Kis and *CYP6M2*_Tias gene promoters, an online analysis using ConSite (http://asp.ii.Uib. no :8090/cgi-bin/CONSITE/consite/) was performed. The putative transcription binding sites for *AhR*/*ARNT* and *Nrf2*/ *ARE* were searched using Consite web site (http://consite.genereg.net/cgi-bin/consite?rm=t_input) [24, 3]. ConSite is a user friendly web-based interface that is used conventionally to identify cis-regulatory elements within genomic sequences.

2.3.3 Prediction of conserved *Anopheles gambiae* (*CYP6M2*) and *Drosophila melanogaster* (*CYP6G1*) putative TFs in the promoter region (Consite)

In order to predict the conservation between the regulatory regions of *CYP6M2* gene (896 bp) and *CYP6G1* gene (896 bp) putative TFs in the promoter regions, the Consite tool was used (http:// consite. genereg.net/cgi-bin/consite?rm=tinput) [24, 5].

2.3.4 Multiple *CYP6M2* gene promoter peptide sequence alignments

Multiple sequence alignments demonstrate conserved amino acid sequences. The online gene sequence alignment tool Vector base (http://www.vectorbase.org) was used to align previously isolated *Drosophila melanogaster*, *Aedes aegypti and Culex quinquefasciatus* 5' upstream *CYP6M2* sequence to the reference *Anopheles gambiae CYP6M2* sequence. DNA sequence homology was further examined in terms of percentage, using the ClustalW2 online tool from EMBL-EBI (http:// www.ebi.ac.uk/) (EMBL-EBI). This programme aligned the *CYP6M2* promoter sequences earlier retrieved to produce an alignment profile and a summary table displaying sequence homology within all the selected species (EMBL-EBI).

2.3.5 Identification of Aryl hydrocarbon receptor (*AhR*) and Nuclear *factor*-erythroid *2-related factor-2* (*Nrf2*) orthologous and their receptors

2.3.5.1 Identification of *AhR/ARNT* orthologs

In order to search for the orthologs of *AhR/ARNT* in *Anopheles gambiae*, we used the Ensembl genome browser (http://www.ensembl.org/Multi/Search/Results?q=AhR;y=3;site = ensembl all;x=3) to search for *AhR* gene in Human. While Fly base (http://flybase.org/reports/ FBgn0003513.html) for *Drosophila melanogaster* (Dm) and VectorBase (https:// www.vectorbase. org/ Anopheles gambiae/Gene) for *Anopheles gambiae* (Ag) data bases were used to search for AhR orthologs.

2.3.5.2 Identification of *Nrf2/ARE* orthologs

In order to search for the orthologs of *Nrf2/ARE* in *Anopheles gambiae*, we used the ensembl genome browser (http:// www. ensembl. org/ Human/ Search/ Results?q=%20 Nrf2;%20y=% 203;site=ensemblall x=3;;page=1;facet_species=Human) was used to search for the *Nrf2* gene in Human.While Fly base (http://flybase.org/reports/FBgn0262975.html) for *Drosophila melanogaster* and (https://www.Vectorbase.org/Anopheles gambiae/ Gene) for *Anopheles gambiae* data bases were used to search for the *Nrf2* orthologs.

3. Results

3.1 Phylogenetic tree of *CYP6M2* gene sequence

To determine how conserved the *Anopheles gambiae CYP6M2* gene sequence is between different species of insects, a phylogenetic tree of the *Anopheles gambiae CYP6M2* gene sequence was obtained from Ensembl (Ensembl Genome Browser).

Fig 3.1: Phylogenetic tree of *CYP6M2* gene sequence (Ensembl Genome Browser). The following tree was generated by Ensembl and takes gene orthologs and paralogs into account. The phylogenic tree highlights "*CYP6M2*, *Anopheles gambiae*" in red font. Paralogues have been shown in blue.

Figure 3.1 shows the inferred evolutionary relationship between various species based on *CYP6M2* gene sequence similarities. The *Anopheles gambiae CYP6M2* gene sequence is the reference sequence in this regard. This result indicates that the *Anopheles gambiae CYP6M2* gene sequence shares particular homology with members of the Diptera order**.**

3.2 *CYP6M2* promoter genomic sequence retrieval

An 896 bp sequence upstream of the *CYP6M2* TSS was retrieved from VectorBase. Upstream of this region is a large transposable element (TE) upstream of *CYP6M2* and we hypothesise that promoter elements must be between this TE and the TSS.

Fig 3.2: Shows image of the retrieved *CYP6M2* promoter gene image indicating the *CYP6M2* protein coding (Red boxes) entire forward and reverse strands of 896 bp (Black arrows).

3.3 Promoter searching and analysis of *Anopheles gambiae, Aedes aegypti, Drosophila melanogaster, Culex quinquefasciatus and Homo sapiens Cytochrome P450s:*

The sequences comparison of 896 bp promoter region of *Anopheles gambiae, Aedes aegypti, Drosophila melanogaster, Culex quinquefasciatus and Homo sapiens* showed similarity using percent identity matrix (**Table 3.1**).

Table 3.1: Percent Identity Matrix - created by Clustal2.1

Gene ID	AAEL009496	AGAP008212	Fbgn0000473	CPIJ016809	ENSG00000059377
AAEL009496	100.00	44.99	37.26	35.27	35.18
AGAP008212	44.99	100.00	36.18	37.77	36.79
Fbgn0000473	37.26	36.18	100.00	42.40	40.75

Nucleotide homologies between *CYP6M2* 896 bp *Anopheles gambiae* (AGAP08212), *Aedes* (AAE009496), *Culex* (CPI016809) *Drosophila* (FBgn0000473) and *Homo sapiens* (*CYP5A1*) species 896 bp promoters. Matrix indicates the percentage identities of aligned nucleotide identities between the *Anopheles*, *Aedes*, *Culex*, *Drosophila* and *Homo sapiens* promoters.

Table 3.2: Shows the prediction of TSSs with upstream region in the cloned *CYP6M2*_Kis and *CYP6M2*_Tias

Promoter	Length (bp)	TATA Box position	(TSS)	(cPG count)	Ratio of O/E	(%GC)
*CYP6M2*_Kis-1	915	824	-97	42	0.75	50.50 - 53.00
*CYP6M2*_Kis-2	896	818	-108	31	0.99	50.50 - 52.50
*CYP6M2*_Tias-1	929	838	-50	56	0.78	50.50 - 53.00
*CYP6M2*_Tias-2	929	838	-50	56	0.78	50.50 – 53.00

Key : Length in base pairs (bp), TATA Box position , Transcription start site (TSS)-position relative to the first coding base in the sequence, cPG count- Total cPG count within the sequence, Ratio of observed over expected cPG dinucleotide and % G+C(GC content)=50.50- 53.00)

The cPG count is the number of CG dinucleotides. The Percentage cPG is the ratio of cPG nucleotide bases (twice the cPG count) to the length. The ratio of observed to expected cPG is calculated according to the formula cited in [25].
Tables 3.3 and 3.4 show the prediction of putative CpG dinucleotides was performed using the CpG finder at the European Bioinformatics Institute (EMBL, http: //www.ebi.ac. uk/) and the following parameters: CpG length >200 bp, G+C >50% and a “CpG value” of at least 0.75. Further analysis of promoter region and start prediction using McPromoter invertebrate analysis tool) predicted the existence of a promoter at position 360 and a TATA box at positions 824 bp, 818 bp, 838 bp and 838 bp for *CYP6M2*_Kis-1, *CYP6M2*_Kis-2, *CYP6M2*_Tias-1 and *CYP6M2*_Tias-2 respectively.

cPG dinucleotide analysis by carrying out the cPG dinucleotide search, we generated data on the cPG dinucleotide content in the Tiassalé resistant and Kisumu susceptible strains of *Anopheles gambiae* (**Table 3.3**). The region of cPG dinucleotide content is more or less similar with each other, although the length of the cPG is different from each other. The GC contents of these strains range from 50.50- 53.00 %. The transcription start sequence (TSS) of the susceptible and resistant strains is different from each other with *CYP6M2*_Tias-1 and *CYP6M2*_Tias-2 (871 bp) longer than *CYP6M2*_Kis-1 (854 bp) and *CYP6M2*_Tias-2 (788 bp) respective.

Table 3.3: Shows McPromoter invertebrate analysis of the *CYP6M2* cloned sequences

Promoter	*Length (bp)*	*TSS Positions*	*Threshold*
CYP6M2_Kis-1	915	-861 -862	+0.06703
CYP6M2_Kis-2	896	-821 - 822	0.04599
CYP6M2_Tias-1	929	-871 - 872	+0.05115
CYP6M2_Tias-2	929	-871 - 872	+0.05115

3.4 Prediction of putative promoter elements

McPromoter was used to search for promoter elements within the Kisumu and Tiassalé clones. McPromoter is a probabilistic promoter prediction system that identifies likely location of eukaryotic RNA polymerase II transcription start sites (TSSs) in large genomic sequences [11]. The McPromoter analysis incorporates data on DNA structure and binding site information thereby generating a graph that displays the likelihood of a promoter occurring at each position. In each of the cloned promoter regions within the CpG island (enclosed by the highest threshold **Figure 3.4**) was identified as the most likely location of a promoter element. The threshold used as a promoter element was identified in all the four clones of *CYP6M2*, predicted to lie within the CpG Island with a TSS at nucleotides (861-862 bp), (-821-82 bp), (-871-871 bp) and (-871-872bp) for *CYP6M2*_Kis-1, *CYP6M2*_Kis-2, *CYP6M2*_Tias-1 and *CYP6M2*_Tias-2 clones respectively (Table 3. 3). The TSS has a predictive score of 0.00999, on a range of -0.5 to 1 in which higher values indicate a greater likelihood of promoter activity.

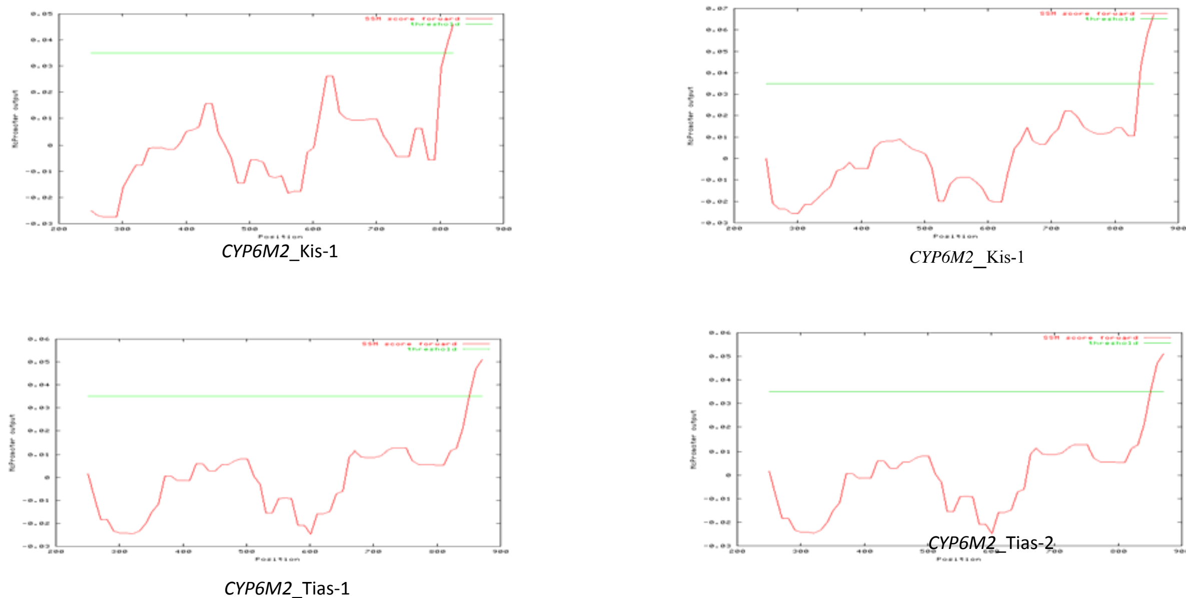

Figure 3.4: McPromoter analysis of the 896 bp region upstream *CYP6M2*_Kis-1 promoter. Grey lines indicate the boundaries of CpG dinucleotides identified using the UCSC browser (http://tools.genome.duke.edu/generegulation/McPromoter/). The height of the graph at any point indicates the likelihood of a promoter occurring at that site, on a scale of -0.5 to 1 with 1 being optimal. The threshold used is for an intermediate sensitivity of 50%, for which the threshold range for all the clones is between +0.04599 to +0.06703. However, the CpG dinucleotide content is identified as the most likely to contain a promoter element. Lowering the threshold will increase the sensitivity and the number of promoters predicted, but will also increase the chance of false-positives.

3.5 Prediction of conserved *Anopheles gambiae* (*CYP6M2*) and *Drosophila melanogaster* (*CYP6G1*) putative transcriptional binding sites (TBS) in the promoter regions (Consite)

Previous studies have shown that sequence conservation between species can be used to predict functional TFBS. Pairwise alignment and subsequent analysis have shown the following transcription factor binding sites as being conserved between the two species.

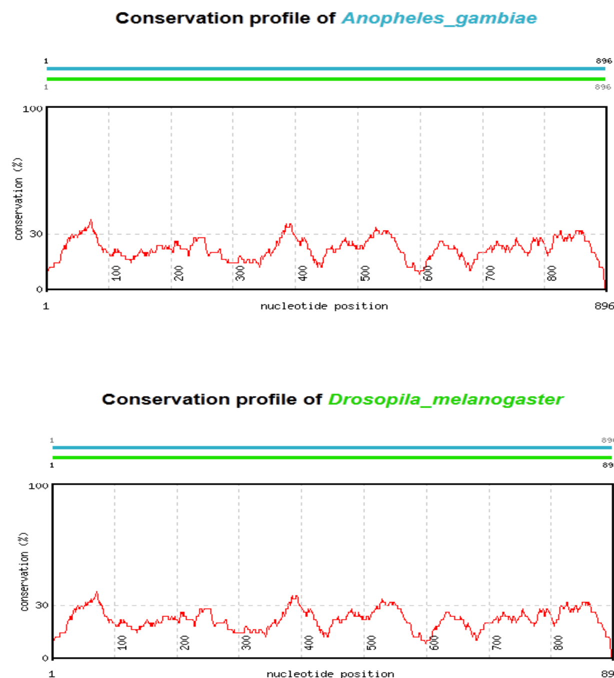

Figure 3.5: **Sequence alignment and TFBS analysis of *Anopheles gambiae* (*CYP6M2*) and *Drosophila melanogaster* (*CYP6G1*) gene promoter using conservation cut off Score of 30% and TF score threshold of 80% (Phylofoot.org).** The *Anopheles gambiae* (*CYP6M2* 896 bp) and (*CYP6G1* 896 bp) promoter sequences were reverse complimented (The Sequence Manipulation Suite). The “Analyze orthologous pairs of genomic sequences” tool (http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite?rm=t_input) from ConSite was used to visualise TFBS conservation (Phylofoot.org). The sequences were transcribed into the ConSite tool and the *AhR*/*ARNT* and *Nrf2*/*ARE* were selected to produce the above grap.

```
FBgn0262975
FBtr0306748          DIINLPMDEFNERLSKYDLSENQLSLIRDIRRRGKNKVAAQNCRKRKLDQ 1219
AGAP005300
AGAP005300-RB        DIINLPMDEFNERLSKYDLSETQLSLIRDIRRRGKNKVAAQNCRKRKLDQ 892
ENSG00000116044
ENST00000039706      KIINLPVVDFNEMMSKEQFNEAQLALIRDIRRRGKNKVAAQNCRKRKLEN 505

FBgn0262975
FBtr0306748          ILTLEDEVNAVVKRKTQLNQDRDHLESERKRISNKFAMLHRHVFQYLRDP 1269
AGAP005300
AGAP005300-RB        IVTLADEVKDMKMRKERLMRDREMVLSEHKKIRDKFSMLYRHVFQNLRDA 942
ENSG00000116044
ENST00000039706      IVELEQDLDHLKDEKEKLLKEKGENDKSLHLLKKQLSTLYLEVFSMLRDE 555
```

A

```
ENST00000242057              KLDFTPIGCDAKGRIVLGYTEAELCTRGSGYQFIHAADMLYCAESHIRMIKTGESGMIVF
FBgn0003513:FBtr0306026      KLDFSLVSMDQRGKHILGYADAELVNMG-GYDLVHYDDLAYVASAHQELLKTGASGMIAY
AGAP010259-RA                KLDFSLVSMDHKGKNTLGYSDSELANMG-GYDLVHYDDLAYVASAHQELLKTGASGMIAY

ENST00000242057              RLLTKNNRWTWVQSNARLLYKNGRPDYIIVTQRPLTDEEGTEHLRKRNTKLPFMFTT-GE
FBgn0003513:FBtr0306026      RYQKKDGEWQWLQTSSRLVYKNSKPDFVICTHRQLMDEEGHDLLGKRTMDFKVSYLDTGL
AGAP010259-RA                RYQKKDGAWQWLQTSSRLVYKNSKPDFVICTHRQLMEEEGRDLLGKRTMDFKVSYLDAGI
```

B

Fig 3.6: A - Amino acid alignment of FBgn0262975 (*CnCC*) sequences *Drosophila melanogaster (Dm)*, with AGAP005300 *Anopheles gambiae* (*Ag*) and ENSG00000116044.11 (*Nrf2*) *Homo sapiens* (*Hs*) species. Where * indicate identical residues; conserved substitutions; Semi-conserved substitutions. The alignment was performed using ClustalX (http://www.ebi. ac.uk/). Dashes are used to denote gaps introduced for a maximum alignment. **B -** Amino acid alignment of FBgn003513 (*ss*) sequences *Drosophila melanogaster (Dm)*, with AGAP010259 *Anopheles gambiae* (*Ag*) and ENST000242057 (*AhR*) in *Homo sapiens* (*Hs*) species. Where * indicate identical residues; conserved substitutions; Semi-conserved substitutions. The alignment was performed using ClustalX (http://www.ebi.ac.uk/). Dashes are used to denote gaps introduced for a maximum alignment.

3.7 Putative transcription binding sites found along *Anopheles gambiae* at 70% cut off score

-885-890 bp
-349-359 bp
-33-42 bp
-6-11 bp

Fig 3.7: TFBS analysis on 897 bp *CYP6M2* promoter. Reverse compliment (http://www.Bioinformatics.org /sms/ index.html) was performed on the sequenced *CYP6M2*.The reverse compliment sequence was analysed using the "analyze single sequence" tool available from ConSite (Phylofoot.org,). Sequence analysis at 70% cut off score or stringency revealed 49 *AhR*/*ARNT* and *Nrf2*/*ARE* specifically selected TFBS at specified positions. A total of 49 *AhR*/*ARNT* and 10 *Nrf2*/*ARE* putative transcription binding sites were found within the 896 bp *CYP6M2* gene.

Table 3.4: Shows the results of the orthologous for *AhR*/*ARNT* and *Nrf2*/*ARE* of *Homo sapiens* in *Drosophila melanogaster* and *Anopheles gambiae*.

Mammalian	*Drosophila* melanogaster	*Anopheles gambiae*
AhR	*ss*	AGAP010259
ARNT	*Tgo*	AGAP009748
Nrf2	*Cncc*	*Nf2e1*
Keap 1	*Keap 1*	AGAP003645

Key: *AhR* (Aryl hydrocarbon receptor)**,** *ARNT* (Aryl hydrocarbon nuclear translocator), *Ss* (*Spineless*), *Tgo* (*Tango*), *Nrf2* (Nuclear factor erythroid 2 –related factor 2), *CnCC* (Cap 'n' collar isoform C), *Nf2e1* (Nuclear factor erythroid 2 *invertebrate*)

Table 2.4 Shows the orthologs to *AhR* in higher mammalian as spineless (ss) and AGAP010259 in *Drosophila melanogaster* and *Anopheles gambiae* respectively. Whilst *Tango* and AGAP009748 are the orthologs to *ARNT* in higher mammalian. Similarly, orthologs to *Nrf2* in higher mammalian are *CnCC* and *Nf2e1* in *Drosophila melanogaster* and *Anopheles gambiae* respectively. Whilst *Keap* 1and AGAP003645 are orthologs to *Keap* 1 in higher mammalian**.**

4. Discussion

The focus of this study was to understand the molecular mechanism involved in the regulation of *CYP6M2*, characterization of its promoter sequence and searches for potential regulatory elements were made. The recognition of these transcription factor binding sites (TFBSs) has the potential to improve knowledge on how wild populations of *Anopheles gambiae* become resistant to insecticide and are activated by different endogenous and exogenous challenges. *Anopheles gambiae* genomic resources were used to assess the 5' upstream region hypothesised to contain the promoter region of *CYP6M2*, a gene associated with the detoxification of insecticides including permethrin [26, 2]. Bioinformatic analysis of the *CYP6M2* represents the characterization of this gene that performs important functions in the metabolism of insecticides resistance. Conserved putative TFBS within the cloned and sequenced *Anopheles gambiae* promoter were identified. Transcription factors in their active form have the ability to bind to the putative TFBS. To establish the differences between the pJET-*CYP6M2*_Tias and pJET-*CYP6M2*_Kis promoters, two separate analysis were carried out to firstly predict the position of the promoter, TATA Box and Enhancer sequences, and secondly to predict the position of CpG dinucleotides in the cloned and sequenced DNA. Consistent with the previous reports, several putative GATA binding sites, recognized as a positive regulatory element in several invertebrate genes were observed in all the promoter regions of the four constructs [27,19].The *CYP6M2* sequence contains a number of CpG dinucleotide , TATA BOX and a promoter sequence. cPG dinucleotides could stimulate gene expression under gene expression in several ways due to their binding affinity to powerful transcription factors such as Sp1 [21]. Organisms such as the invertebrates *Drosophila melanogaster* and *Caenorhabditis elegans* and the fungus *Saccharomyces cereviceae* have little or no DNA methylation and, as a result, CpG dinucleotides occur at the expected frequency throughout the genome [28]**.** DNA methylation in invertebrates is associated with transcription level, alternative splicing, and genome evolution [29]**.** Our findings are in accord with the conclusions that 'The promoters of vertebrate and invertebrate organisms differ in that invertebrate genomes do not contain cPG islands, a feature for more than half of vertebrate genes' [25]**.** Comparison of *Homo sapiens AhR*/*Nrf2* and their orthologs in *Drosophila melanogaster* and *An*. *gambiae* show significant degree of conservation in a number of regions (Figures 3.5 A and B). According to the Bioinformatics analysis of *An. gambiae* Cytochrome P450 related genes including *CYP6M2*, significant progress in understanding the complex process of regulation of this gene by *AhR* (AGAP010295) and *Nrf2* (AGAP005300) have been achieved. It appears that this work only focuses on *CYP6M2*, but the significance of the work is far more important than the *CYP6M2* itself. The approach in this work can be carried out with other Cytochrome P450 genes to gather more information on the pathways involved in the regulation of genes involved in the metabolism of insecticides. This research therefore is a fundamental work in the fields of the bio informational analysis, and also put forward a new way for the bioinformatics analysis of other detoxification genes.

5. Conclusions

The present findings established insilico presence of putative *AhR*/*ARNT* (AGAP010295/AGAP009748) and/or *Nrf2*/*Keap* 1 (AGAP005300/ AGAP003645) binding sites in *Anopheles gambiae* .This potentially suggests that *CYP6M2* is under the control of *Nrf2* through *ARE* and or *AhR* through Dioxin response element (*DRE*) in the 5' –flanking region. Thus we propose that *AhR*/*ARNT*/ or *Nrf2*/*Keap* 1 pathways are involved in the cellular network that maintains redox homeostasis in order to protect cells from oxidative stress induced by their respective ligands. Verification could be made by subsequent wet lab experiments such as Dual luciferase assay and/ or Quantitative polymerase chain reaction (qPCR).

6. Acknowledgments

This work was supported by funding from the Tertiary Education Trust Fund (TETF) and University of Abuja, Nigeria.

7. References

1. David JP, Ismail HM, Chandor-Proust A, Paine MJI. Role of cytochrome P450s in insecticide resistance: impact on the control of mosquito-borne diseases and use of insecticides on Earth. Role of cytochrome P450s in insecticide resistance: impact on the control of mosquito-borne diseases and use of insecticides on Earth. *Phil Trans R Soc B*, 2013; 368:20120429.
2. Edi CV, Djogbenou L, Jenkins AM, Regna K, Muskavitch MAT. CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito *Anopheles gambiae*. *PloS Genet,* 2014; 10(3) e1004236. doi: 10. 1371 /journa .pgen.1004236,
3. Miglani M. Gakhar SK. Cloning and Bioinformatic Analysis of Vitellogenin Gene of the India Malaria Vector *Anopheles culifefacies* (Diptera: Culicidae). *Acta*

Entamoligica Sinica 2013; 56(9):1063-1074.

4. Nkya TE, Akhouayri, Poupardin R, Batengana B, Mosha F, Magesa S *et al*. Insecticide resistance mechanisms associated with different environments in the malaria vector *Anopheles gambiae*: a case study in Tanzania. Malaria Journal 2014; 13:28.
5. Neafsey DE, Christophides GK, Collins FH, Emrich, SJ, Fontaine MC, Gelbart W *et al.* Waterhouse1, R.M., Williams, L.J. and Besansky, N.J. The Evolution of the *Anopheles* 16 Genomes Project. G3: *Genes/ Genomes/Genetics* as doi: 10. 1534/ g3.113.006247, 2014.
6. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR *et al*. The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*. *SCIENCE,* 2002; (298):4.
7. Pavesi P, Mauri G, Pesole G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* 2004; 32:199-203.
8. Qiu, P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochemical and Biophysical Research Communications*,* 2003; 309:495-501.
9. Amit R, Hernan HG, Phillips R, Fraser SE. Building Enhancers from the Ground up. A Synthetic Biology Approach. *Cell* 2011; 146: 105-118.
10. Talebzadeh M, Zare-Mirakabad F. Transcription Factor Binding Sites Prediction Based on Modified Nucleosomes. *PLOS ONE*, 2014; 9(2): e89226, doi: 10. 1371/ Journal. pone.0089226.
11. Ohler W, Liao G, Niemann H, Rubin GM. Computational analysis of core promoters in the *Drosophila* genome. Genome Biology, 2002; 3(12):1-12.
12. Hu Z, and Gallo SM. Identification of interacting transcription factors regulating tissue gene expression in human. BMC Genomics 2010; 11(49):1-15.
13. Pettinato L, Calistri E, Di Patti F, Livi R, Luccioli. Genome-Wide Analysis of Promoters. Clustering by Alignment and Analysis of Regular Patterns. *PLoS ONE*, 2014; 9(1): e85260, doi: 10.1371/journal.pone.0085260.
14. Cartharius K, Frech K, Grote, B, Klocke M, Haltmeier A, Klingenhoff M *et al* MatInspector and beyond. promoter analysis based on transcription factor binding sites 2005; 21(13):2933-2942.
15. Nain V, Sahi S, Kumar PA. In Silico Identification of Regulatory Elements in Promoters. Computational Biology and Applied Bioinformatics. Prof. Heitor Lopes (Ed.) 2011.
16. Lee TI, Rinaldi NJ, Robert DT, Odom Z. Bar-Joseph, GK Gerber *et al*. Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science* 2002; 298:799-804.
17. Chen S, Rasgon JL. *Culex tarsalis* Vitellogenin Gene Promoters Investigated In Silico and In Vivo Using Transgenic *Drosophila melanogaster*. *PLoS ONE,* 2014;9(2):e88994.
18. Liu N. Pyrethroid Resistance in Insects Genes, Mechanisms, and Regulation, Insecticides -Advances in Integrated Pest Management, Dr. Farzana Perveen (Ed.), ISBN: 978- 953-307-780-2, 2012.
19. Good, R. T., Gramzow, L., Battlay, P., Sztal, T., Batterham, P. and Charles, R. The Molecular Evolution of Cytochrome P450 genes within and between *Drosophila* specie Genome Biology and Evolution Advance Access published April 20, 2014 doi:10.1093/gbe/evu083, 2014.
20. Mohammed R.B., Wilding, C.S., Collier, P.J. and Deeni, Y. Y. Cloning of *Anopheles gambiae CYP6M2* Gene Promoter and Construction of its Luciferase Reporter System .*International Journal of Scientific and Technology Research*, 3(7), 2014, 259-264.
21. Kadalayil, L., Petersen, U-M. and Engström, Y. Adjacent GATA and κB-like motifs regulate the expression of a Drosophila immune gene. *Nucleic Acids Research*, 1997 (25): 6 1233– 1239.
22. Krinner, S., Heitzer, A.P., Diermeier, S.D., Obermeier, I., La¨ngst, G. and Wagner, R.: CPG domains downstream of TSSs promote high levels of gene expression. *Nucleic Acids Research*, 2014; 1–14.
23. Sharif, J., Endo, T.A., Toyoda, T. and Koseki, H.: Divergence of CPG island promoters: A consequence or cause of evolution. *Develop*. *Growth Differ*. 2010; 52:545–554.
24. Sandelin A, Bailey P, Bruce S. Engström PG, Klos Mj, Wasserman WW *et al.* Arrays of ultra-conserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*.2004; 5-9.
25. Gardiner-Garden and Frommer M. CpG islands in vertebrate genomes. J Mol Biol 1987; 196(2):261-82.
26. Mitchell SN, Stevenson BJ, Müller P, Wilding CS, Egyir-Yawson A, Field SG *et al.* Identification and validation of a gene causing cross-resistance between insecticide classes in *Anopheles gambiae* from Ghana. Proceedings of the National Academy of Sciences, USA, 109, 6147–6152, 2012.
27. Felix RC, Silveira H. Promoter analysis of three P450 cytochromes in *Anopheles gambiae*. Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal, Unpublished Thesis 2011; 71-98.
28. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Dev. 2011; 25: 1010-1022, 29.
29. Rivière G. Epigenetic features in the oyster Crassostrea gigas suggestive of functionally relevant promoter DNA methylation in invertebrates. Frontiers in Physiology 2014; 5 (129).