

Apr 01, 2022

# Screening whole proteome of *Aedes aegypti* and identification of potential targets for in-silico molecular and structural interaction studies against natural bioactives

Anagha S Setlur<sup>1</sup>, Chandrashekar K<sup>1</sup>, Manas Sarkar<sup>2</sup>, Vidya Niranjana<sup>1</sup>

<sup>1</sup>Department of Biotechnology, RV College of Engineering, Bangalore- 560059;

<sup>2</sup>Research and Development, Reckitt Benckiser India Pvt. Ltd., Gurgaon, Haryana- 122001

1



[dx.doi.org/10.17504/protocols.io.e6nvwbj2vmk/v1](https://dx.doi.org/10.17504/protocols.io.e6nvwbj2vmk/v1)

Centre of Excellence in Computational Genomics (Vidya Lab)

Vidya Niranjana

As a dangerous etiological agent for dengue, chikungunya, Zika and yellow fever, it is essential to combat the incidences of *Aedes aegypti*, by using repellents. However, chronic overuse of synthetic repellents has led to possibilities of adverse side effects in humans. As a consequence, scientists and researchers are now shifting the focus of research on developing natural alternatives to these repellents. In such a case, the present study aimed to devise a standard protocol that can screen the whole proteome of *A. aegypti* and identify the major proteins that can be targeted by natural bioactives to produce repellents. To study the binding of the natural actives and the targets, a whole proteome analysis was carried out by finding the reference proteome of the organism, performing a literature survey to identify the potential targets, understanding the circadian rhythm of *A. aegypti* to identify the proteins expressed in the dark and light cycles, and shortlisting the targets by analyzing the common conserved domains of query sequences. Twenty protein target categories were identified, out of which 309 protein sequences were modelled using standalone tool- RaptorX. These structures were validated using Ramachandran plots from SAVES v6.0. Molecular docking studies using POAP, between the selected representative of the twenty protein targets and the natural bioactives revealed negative binding energies. Those that had the least negative energies were taken forward for 100ns molecular dynamic simulation studies, from which the docked complex stabilities were noted and the conformational changes induced during simulations were revealed. This protocol allows whole proteome analysis that will enable identification of major protein targets that the naturals can act upon, and further reveals the effectiveness of the use of naturals against these proteins, thereby, implying the use of this methodology for whole proteome analysis of other organisms as well.

**Keywords:** Whole proteome, *Aedes aegypti*, conserved domains, homology modelling, docking, molecular dynamic simulations

DOI

[dx.doi.org/10.17504/protocols.io.e6nvwbj2vmk/v1](https://dx.doi.org/10.17504/protocols.io.e6nvwbj2vmk/v1)

Anagha S Setlur, Chandrashekar K, Manas Sarkar, Vidya Niranjana 2022. Screening whole proteome of *Aedes aegypti* and identification of potential targets for in-silico molecular and structural interaction studies against natural bioactives. **protocols.io**  
<https://dx.doi.org/10.17504/protocols.io.e6nvwbj2vmk/v1>

Reckitt Benckiser India Pvt. Ltd.  
Grant ID: CRW96972

Whole proteome, *Aedes aegypti*, conserved domains, homology modelling, docking, molecular dynamic simulations

\_\_\_\_\_ protocol ,

Mar 29, 2022

Apr 01, 2022

60038

## IDENTIFICATION OF PROTEIN TARGETS

### 1 Reference proteome identification

*Aedes aegypti* is a yellow fever mosquito, that is also an etiological agent for dengue, chikungunya and Zika. Hence, identification of the significant proteins in the organism's reference proteome by a whole proteome screening is essential to use these as major targets against given set of natural bioactive compounds. Thus, the reference proteome for *Aedes aegypti* was identified from UniProt database. The taxonomic ID with [7159](#), is the reference proteome for *A. aegypti*, with UniProt ID UP000008820. A total of 21,496 proteins are present in *A. aegypti*, expressed by 14,555 genes. The selected strain was LVP\_AGWG and its genome assembly and annotation is [GCA\\_002204515.1 from EnsemblMetazoa](#).

Overview	
Status	Reference proteome
Proteins	21,496
Gene count	14,555 - Download one protein sequence per gene (FASTA)
Proteome ID	UP000008820
Taxonomy	7159 - Aedes aegypti
Strain	LVP_AGWG
Last modified	May 27, 2021
Genome assembly and annotation	GCA_002204515.1 from EnsemblMetazoa full
Busco	C:99.4%;S:61.4%;D:38%;F:0.2%;M:0.4%;n:3285 diptera_odb10
Completeness	Standard

Aedes aegypti is the principal vector of the etiological agents of yellow fever, dengue fever and chikungunya fever. In 1901, Walter Reed and colleagues showed that yellow fever was transmitted by Aedes aegypti. Their work was based on Carlos Finlay's hypothesis that mosquito bites could transmit the organism causing the disease.

Due to its easy adaptation to laboratory culture it is the most studied species within the Culicidae, and has extensively contributed to our understanding of mosquito biology, physiology, genetics, and vector competence.

The Aedes aegypti complete genome sequence was first published in 2007. About 47% of the genome consists of transposable elements, probably contributing to the expansion in size and organization of the Aedes aegypti genome over time.

## The reference proteome selected for whole proteome screening of *Aedes aegypti*

### 1.1 Identification of major protein targets by whole proteome screening

Literature survey revealed several categories of proteins that could potentially act as targets against natural repellents. From a whole proteome screen, twenty different categories of proteins were selected after a thorough scrutiny. Previous studies have stated that from among the total expressed genes in *Aedes aegypti*, only about 7.9% of the genes were known to be involved in rhythmicity, pointing towards those expressed during the 24-hour light and dark cycles in the organism (Leming *et al.*, 2014). Since these genes play critical roles the genome of the organism, twenty different protein targets were identified after careful inspection of the genes that fall under the circadian rhythm of the mosquito. The shortlisted proteins for analysis were also cross-checked with previously published microarray data, to comprehend the level of expression of each of these proteins during the circadian cycle. Moreover, each of these twenty categories of proteins had several proteins available in UniProt, some fragmented and some whole sequences of the proteins. The categorization of the proteins was carried out depending on the functions of the targets identified. The sequences for all these twenty categories were retrieved in FASTA format, for the reference proteome identified previously. The sequence retrieval was carried out in batch mode for each protein category. The downloaded sequences were filtered to distinguish the complete and fragmented ones. Only the complete sequences were chosen for structure prediction and modelling, except for those proteins where the complete sequences were not available, wherein, the fragmented sequences were selected for modelling.

UniProtKB 2021\_04 results

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**  
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**  
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by: **Reviewed (4)** **Unreviewed (291)**

Popular organisms: **AEDAE (295)**

Proteomes: **UP000008820 (205)**

Subcellular location: **Multi-pass membrane protein (194)**

Do you mean odorant binding protein: "Aedes aegypti (Yellowfever mosquito) (Culex aegypti) [7159]"  
Quote terms: "odorant binding"

Entry	Entry name	Protein names	Gene names	Organism	Length	Mass
Q16EI9	OR4_AEDAE	Odorant receptor 4	GPROR4 Or4, AAEL015147	Aedes aegypti (Yellowfever mosquito) (Culex aegypti)	406	46,099
Q178U6	ORCO_AEDAE	Odorant receptor coreceptor	SGPRor7 Orco, AAEL005776	Aedes aegypti (Yellowfever mosquito) (Culex aegypti)	478	54,277
Q16S34	OBP45_AEDAE	General odorant-binding protein 45	Obp45 AAEL010714	Aedes aegypti (Yellowfever mosquito) (Culex aegypti)	260	29,321
Q16HV2	Q16HV2_AEDAE	Odorant receptor	GPROR125 5578890, GPRor125, AAEL013893	Aedes aegypti (Yellowfever mosquito) (Culex aegypti)	382	43,987
Q17NP2	Q17NP2_AEDAE	Odorant receptor	GPROR45 5564843, GPRor45,	Aedes aegypti (Yellowfever mosquito)	384	43,891

### Example output for odorant binding proteins in *A. aegypti* reference proteome

```

uniprot-odorant+binding+protein+AND+organism__Aedes+aegypti+[7159]_fasta - Notepad
File Edit Format View Help
>sp|Q16EI9|OR4_AEDAE Odorant receptor 4 OS=Aedes aegypti OX=7159 GN=GPROR4 PE=2 SV=1
MTQSLFDQTFGFIKVLQIMIGYPSCLAPYPTTFASRLKSSAGFVFCFLMLTYCVFGQII
NIGLLMMGHRQTDQVVEEVAIQVSSTGFCIIGLAKMYSLSYNRAILSWLIADFRVKWNAG
ELTDKDRSIRDGTLRPTVAITTTVAALGNIIMVSANFQPVVEMIVGRVVTGEWVKLFPYV
IWFPPFNSTHGAIIYLVYLFVYSGVIVAVGNVGFNCFCLLTSLSMLQLKLLCSWIEDMV
EVEDEKGVQSKKKLYRIVRYHQDLIRGRNALQSMFSTTLFLNFSASSVLMCMQLYLITTA
GITLMVKFTLFMLCILMEIFILCYGEEILANSSSIAAGAFNSNWWYQSKVSQNPFRGKN
LIPIIQQGQRPMLTAWKFWPITIRTFSAILQTSWSYFTLLKTVMH
>sp|Q178U6|ORCO_AEDAE Odorant receptor coreceptor OS=Aedes aegypti OX=7159 GN=SGPRor7 PE=1 SV=1
MNVQPTKYHGLVLDLMPNIRLMQGFHGLFRYVNGPVLIRKLYSWNLIMILLQYFAIMG
NLVNMNTGDVNELTANTITTLFFTHSVTKFIYVAVNSEHFYRTLGIWNQPNSHSLFAESDA
RYHSIALAKMRKLLVMVMTTVL SVVAWITITIFFGDSVKNVFDEKNETYTYVEIPRLPIK
AWYPWDMASGVVPYFFSFIYQAYFLFMSQCANLADVMFCSWLLFTCEQLQHLKGIMRPLM
ELSATLDTRYPNASALFRVASAGSKSELILNEEKDPDKDFDLNGIYNKADWGAQFRAP
STLQTFGDNINGNPNGLTKKQELMVRSAIKYWVERHKHVRLVSAIGETYGAALLHML
TSTIKLTLLAYQATKIDALNVYGLTVIGYLVYALAQVFLFCIFGNRIIESSSVMEAAYS
CHWYDGSSEAKTFVQIVCQCQKAMTISGAKFFTVSLDLFASVLGAVVTYFHVVLVQLK
>sp|Q16S34|OBP45_AEDAE General odorant-binding protein 45 OS=Aedes aegypti OX=7159 GN=Obp45 PE=1 SV=1
MELKTCICILPLLCTLSYAEALDHAAILKSPNELQLECSKYLPISDVSNVDCYDRSIG
LVGRFVNDISGRPAQTIARYYQPDGSDHITRTDQCLCEKVLTPVRNAYCQRASSGLQC
YLDNYGQLLTGTPQFVPVTEIRAAQIFWDCQMLQISDRDLTQIFKDGYNKTSSEGRCLIR
CFLVRAGLYSDRQGNIGRFVQCEGYTAEYEQAVVSCYDRLLKESLDSCLATRTMDEC
IQGNQFSSSDIDGLEKLEVO
>tr|Q16HV2|Q16HV2_AEDAE Odorant receptor OS=Aedes aegypti OX=7159 GN=GPROR125 PE=3 SV=1
MLNVLLLCAGIRTLQAVDWTWERPVRYLVHLISAYHSMVMILQAAHVVAEQNDVMDTAF
LIKIVGMGSAYIKILLTYHANSVDEVEHFRSKPMSSGEDQYDSTVRGKFLRSTLV MIR
WVLGVLIVDEILFAFPNSQRNKLFLKPPAMSSLSGSGITGWLANFLVNWMLIWL SKYLC
CTTKLGLVLLMGLRVEFKILTHKLEQITRQAKSIQSVEDHCKFLKKELEVLKYQAE LRRN
IELIRPLLGMAFMIYYALFFIGTLMVYTHHGGFEFYSLTFASSVVVTLLECYWMCQLV
DSIQDDAESMGNELYDICAGIPYARQSHRTYVEMRTSLMIIINARHSLAIDCVGIFSIS
TAIFVQMLNTSYSVLMFLINMG
>tr|Q17NP2|Q17NP2_AEDAE Odorant receptor OS=Aedes aegypti OX=7159 GN=GPROR45 PE=3 SV=2
MVYNTSL EAYQDIIRLARRMASFCGVLDLADNYKPNWRTALTFAGCISYLTSCYSAWY
YDPVFKMLQALAPNGIAWQGLKLFIAVQQRKFQGRKAYLEEFHRIHVNRPEDNAVLG
LMVKMHLACRVLLSAYITAVLGFGLYPVYFFVMYGERTFAINVLVPGIDPDSQWGYFLT
SYQIFLLSIAMAGISAFDITFLIFVCNLAGLVDFMNNKLELDRLLEKEKADLAEIRDHV
REILIDHYGIVSYESDLNRYIAINLMQVGSVACLISLFLCYMTNYPGYAFILGAIF
QLLTYCCLGTGISVKNDDAILAINGTKWYLLEKSQLKMVGFMHLRQCNPSNLTVGGFAPL
NIETFVEIMKTIYQFFAMINFN
>tr|Q17EY1|Q17EY1_AEDAE Odorant receptor OS=Aedes aegypti OX=7159 GN=GPROR25 PE=3 SV=2
MEDAKFGLIFRFRIRRALSVAGCDIFEENWRPSAWTFIVLLFASIFPYFAVFLLNHDDM
SYERLAESVAIFITSLDGIYGLLEFIVNRNKNWNEVMKNIHSRRFQYKSKTISELFDLYY

```

### Sample FASTA sequences downloaded for a specific protein category

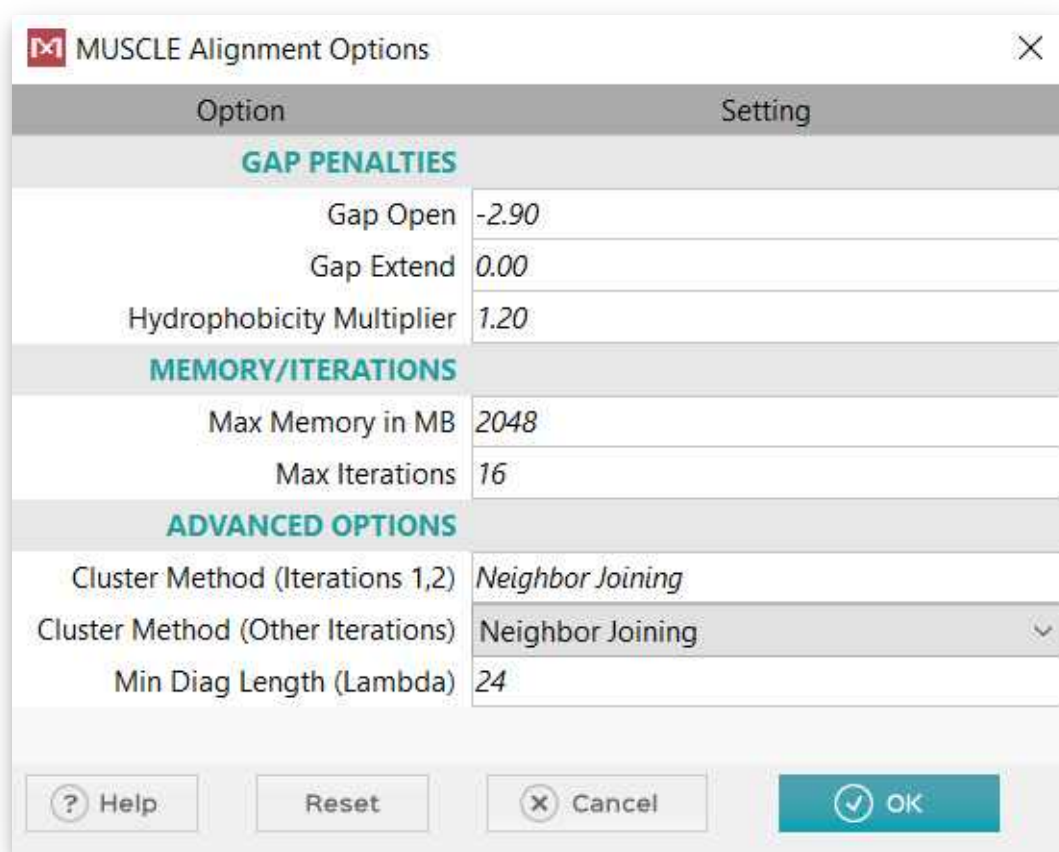
## ALIGNMENT AND CONSERVED DOMAIN ANALYSIS



## 2 Prediction of conserved domains for identified protein targets

Since there were several proteins falling under the same category, all the sequences had to be assessed for similarity. This was performed using multiple sequence alignment (MSA) via MUSCLE algorithm in MEGA-X software. Performing an MSA provided clarity on the sequence similarity and the number of sequences that was to be selected for structure predictions and modelling. When this confirmation was obtained from MSA studies, a conserved domain analysis to streamline the protein targets was carried out using NCBI Conserved Domain search tool, accessible via <https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>.

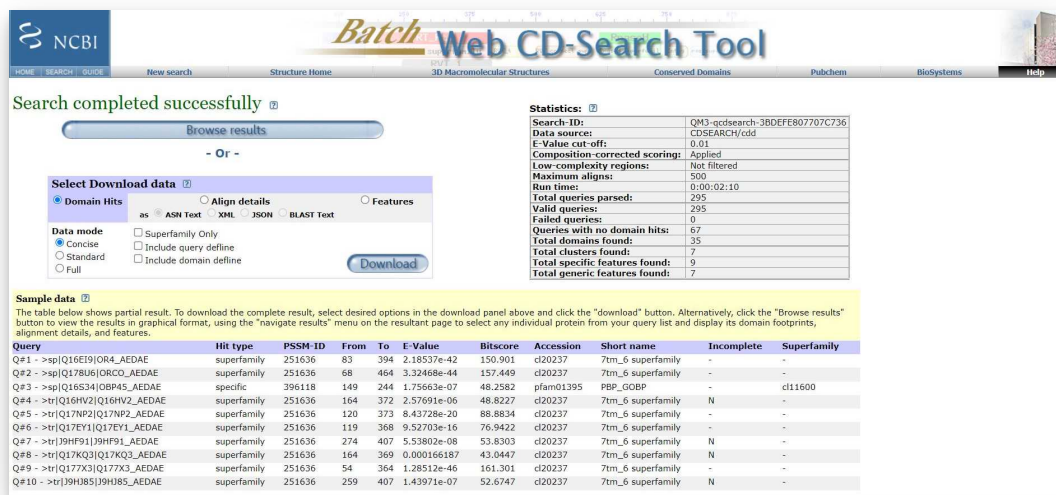
The query sequences for each protein categories were submitted individually to the online webserver and the predictions were carried out in batch mode to examine multiple sequences concurrently. All default parameters were maintained and search was performed against CDD-58235 position specific scoring matrices database. The search was set to composition corrected scoring and the expected threshold value was fixed to 0.01, with maximum 500 hits to be delivered as the search outcome. Suitable job titles were provided prior to running the operation. The results obtained were thoroughly analyzed for presence of common domains in the complete query sequences for each protein category, so that a representative sequence, having the consensus domain responsible for providing the protein its function, can be utilized for structure modelling and eventual docking studies.



Option	Setting
<b>GAP PENALTIES</b>	
Gap Open	-2.90
Gap Extend	0.00
Hydrophobicity Multiplier	1.20
<b>MEMORY/ITERATIONS</b>	
Max Memory in MB	2048
Max Iterations	16
<b>ADVANCED OPTIONS</b>	
Cluster Method (Iterations 1,2)	Neighbor Joining
Cluster Method (Other Iterations)	Neighbor Joining
Min Diag Length (Lambda)	24

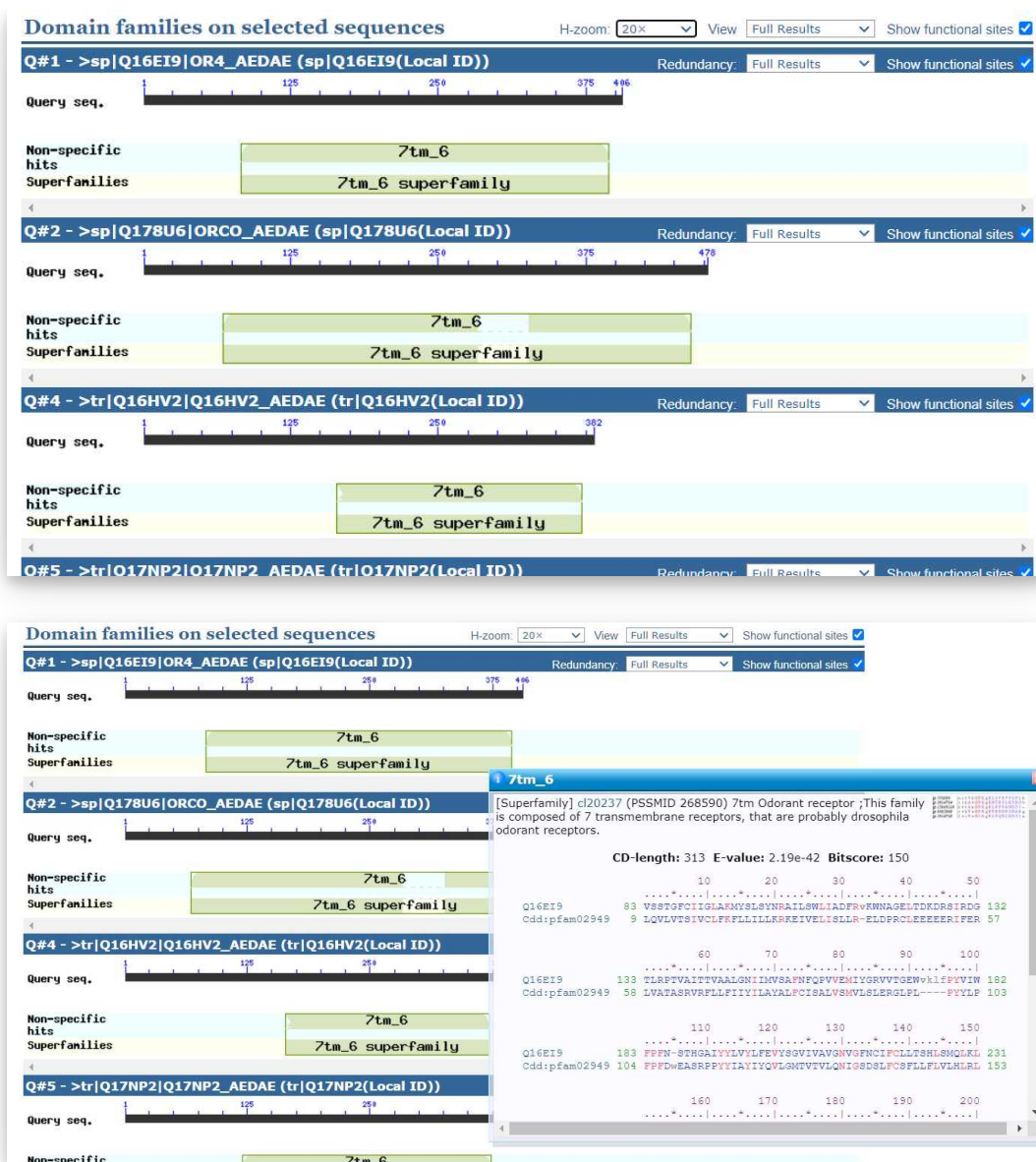
Buttons: ? Help, Reset, X Cancel, OK

Multiple sequence alignment settings using MUSCLE algorithm in MEGA-X tool



### Conserved domain analysis to find the common domains in each protein category

This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**



Sample result showing conserved domains for selected query sequences. CD-length, E-value and score are displayed.

## HOMOLOGY MODELLING OF PROTEINS AND STRUCTURE VALIDATION

- 3 All required protein query sequences from the 20 shortlisted protein targets were modelled first and then their 3-dimensional structures were validated to confirm the accuracy of modelling

### 3.1 Structure modelling of target proteins

Predicting structures of individual protein query sequences could not be performed individually on an online webserver, since there were a large number of proteins whose structures were required. Therefore, a standalone tool for RaptorX protein modelling was installed on the server to perform large-scale structure predictions quickly and accurately. The following tools were installed:

NR (non-redundant) database, template database and PDB files (optional). All required tools such as 309 protein targets were modelled rapidly, using the standalone tool. Unless very few protein sequences existed in a specific protein category, the fragmented sequences were not taken into account for homology modelling.

RaptorX standalone tool comprises of five main programs that are all used during structure predictions. The buildFeature program was used to generate the features of the protein query sequence. Aligning the target protein sequence to the template selected from existing databases was then carried out using CNFalign. Similar templates for the query were searched for in the database, using the CNFsearch package. Using build3Dmodel, a prediction was carried out for the query protein, followed by generating 3D models for the same by buildTopModels package. The instructions provided in the readme file (<http://raptorx.uchicago.edu/>) of RaptorX tool were followed to generate the 3D models of the query protein sequences.

### 3.2 Code for building features

```
./buildFeature -i seq_file [-o tgt_file ] [-c cpu_num]
```

### 3.3 Code for aligning target sequence to template

```
./CNFalign_lite -t template_name -q target_name [-l  
tpl_root] [-g tgt_root] [-d output_root]
```

OR

```
./CNFalign_fast -t template_name -q target_name [-l  
tpl_root] [-g tgt_root] [-d output_root]
```

OR

```
./CNFalign_normal -t template_name -q target_name [-l  
tpl_root] [-g tgt_root] [-d output_root]
```

### 3.4 Code for building 3D models

Model building from single template

```
./build3Dmodel -i align_file -q query_name [-d pdb_root] [  
-m mod_bin ] [ -n mod_num ]
```



```
./buildTopModels -i rank_file [-k TopK] [-d pdb_root] [-m mod_bin]
```

Once the structures were predicted successfully, the generated PDB files were validated.

### 3.5 Structure validation

The predicted structures were validated using Ramachandran plots, generated from SAVES v6.0 (<https://saves.mbi.ucla.edu/>), from ProCheck structure analysis tool. The total number of amino acids in the favorable regions, allowed and disallowed regions were noted to ensure the predicted structures could be taken forward for molecular docking studies.

## MOLECULAR DOCKING

### 4 Molecular docking studies for selected protein targets against natural bioactives

Preliminary docking studies against all 309 predicted proteins evaluated their binding energies. The software used for docking was POAP (Parallelized Open Babel and AutoDock suite) for performing multiple protein-ligand docking concurrently and rapidly. This software combines several tools such as AutoDock 4.2.6, AutoDock Vina, MGL Tools 1.5.6, Open Babel and GNU parallel, all essential tools for high-performance computing, merged into one. Each of these tools were installed individually to be able to access the final POAP docking software. The preliminary docking scores revealed negative energies during the binding event of protein and ligands (natural small molecules, understood to have good insect repelling properties). A representative protein that seemed to provide the best binding with the natural active, was docked again to comprehend the binding at a molecular level. For this purpose, the preparation of ligand was initially carried out. To assess and compare the effects of natural active binding with that of existing controls, ligands that have insect repelling properties and that are synthetic in nature, such as DEET (N,N-diethyl-meta-toluamide), icardin, IR3535 and permethrin were used as the positive controls, while ethanol and acetone were used as the negative controls.

#### 4.1 Preparation of the ligands and controls

The 3D structures in .sdf format of the natural bioactives were retrieved from NCBI PubChem (<https://pubchem.ncbi.nlm.nih.gov/>). Preparation of the ligands for purposes of docking was carried out in POAP command line using the interactive mode via the following code:

```
bash POAP_lig.bash -s
```

The path for the folders containing the ligands were provided, along with providing the maximum number of jobs that were run parallelly, in this case, 64. Ligand optimization was carried out using the Merck molecular force field, and the 10 ligand conformations were generated using the weighted rotor search

method. From among these 10 conformers, only the best ligand conformer was attained and minimized via the steepest descent algorithm. This algorithm minimizes the conformer generated by having the direction of the largest gradient opposite to the direction of the first minimized conformer. When the first direction provides a minimization, another one is carried out in other steepest descent direction, and the process is continued till a minimized value is attained from all directions. In the present case, 100 minimization steps were carried out, with other fixed default parameters such as vander waal's cut-off distance (6 nm), criteria of convergence ( $1e-6$ ), electrostatic cut-off distance (10 Å) and non-bonded pair frequency (10) unchanged. Hydrogen atoms were appended and when the command was run, the final prepared ligands (natural actives + controls) were obtained in 3D conformers in .pdbqt format, feasible for docking.

## 4.2 Preparation of the proteins and molecular docking by virtual screening

AutoDock 4.2.6 was used individually to prepare the proteins for docking. This was carried out by adding appropriate Gasteiger charges to each atom of the macromolecules, with the non-polar hydrogens all merged. Moreover, it is also essential to re-distribute the partial charges of the atoms in AutoDock 4.2.6, to prepare and save the final macromolecules in .pdbqt format, suitable for docking. The prepared proteins were saved in the docking library, created for providing easy access to the protein targets while using POAP. The grid boxes were generated for the prepared proteins in AutoDock 4.2.6, and the x, y, and z coordinates of the grid box that covered the entire protein molecule, were noted to make the configuration file in .txt, which is significant for docking. The configuration file also mentioned the number of modes for docking, which was set to 9. Grid box was prepared in a manner to allow the ligands to scan the entire protein for binding pockets and to bind to the best binding pocket.

Virtual screening was performed in interactive mode in POAP software (<https://pubmed.ncbi.nlm.nih.gov/29533817/>), using the AutoDock Vina option for multiple protein-multiple ligand docking. The following code was employed in the terminal opened in the scripts file of POAP:

```
bash POAP_vs.bash -s
```

The prepared ligands and protein paths were provided as per the prompts, and exhaustiveness for docking was set to 50, to ensure the docking scores were reliable, reproducible, stable and more robust. The top docked complexes, identified via least negative energies of binding were then carried forward for molecular dynamic simulations to ensure docking stability and validate the results.

## MOLECULAR DYNAMIC SIMULATIONS

### 5 Molecular dynamic simulations for best docked complexes

Protein-ligand binding induces alterations in the structure of the proteins and may cause

changes in the stability of the docked complexes. At times, the ligand bound to the protein may dissociate from it during simulation due to its binding instability. To assess the stability of the docked complexes, a molecular dynamic simulation was carried out at 100ns for the top three complexes with least binding energies. This was performed using Maestro workspace in Desmond, by Schrodinger, that is known for its high reliability, scalability, accuracy, precision and rapidity. The simulations are performed in a realistic manner, mimicking the real-world scenarios/environments to evaluate the docked complex stability at specific conditions.

Top three best docked complexes were imported to the workspace after setting the working directory as necessary and pre-processing of the protein was carried out by gauging the structural errors and refining them. The interaction complexes were optimized by minimizing via steepest descent algorithm for 500 times, and removing the waters as required. The environment for 100ns simulation was then set up using the system builder option in Maestro. Solvent model used was TIP3P and the boundary box for the complexes was defined by utilizing an orthorhombic box that constricts the complexes to 5 Å at each x, y and z axis to simulate the complexes within the containment box. The entire system was then neutralized by the appending of chlorine and sodium ions, but also any other ions depending on the total charge the system possesses.

From literature study, it was identified that the approximate pH of the mid-gut of *A. aegypti* prior to blood feeding was found to be 6.0. Since it is at this point that the mosquito must be interceded from carrying out further operations such as feeding, the simulation was set to run at this pH, by altering the PropKA value in Desmond software. Since 50ns simulations are considered to provide good results, but not very conclusive on the complex stability, simulations were carried out at 100ns in the present project to ensure the robustness of the simulations. The trajectory recording interval was set to 0.1ns and normal pressure temperature (NPT) was selected with temperature set to 310K as the thermodynamic parameters for simulation.

## 5.1 Analysis of 100ns simulations studies

A simulation interaction diagram tool was used to scrutinize and analyze the results of simulations. Results obtained in the form of RMSD (root mean square deviation), protein and ligand RMSF (root mean square fluctuations), protein-ligand contacts, the interaction of the ligand with the amino acids of the protein during the 100ns simulation and the time frame of the amino acid residue interactions were all noted and analyzed.

## 5.2 RMSD

Modifications in the values of RMSD help analyze the stability of the docked complex and the mode of binding. Root mean square deviation is computed as the change in the displacement of a certain selection of atoms for a specific time frame with respect to the reference. This was estimated and analyzed for all 1000 frames within the set trajectory. The formula for RMSD is:

The RMSD for frame  $x$  is:

Where,  
 $N$  = total number of atoms  
 $r'$  = particular position of atoms in the  $x$  frame after superimposition to the frame of reference,  
 $t_{ref}$  = time of reference  
 frame  $x$  = recorded at a time of  $t_x$

### 5.3 RMSF

The root mean square fluctuation calculates the average particle deviation, such as the amino acid residue of the protein, over time, from a position of reference, which is typically the time averaged position of the particle. Comprehending protein and ligand RMSFs help in analyzing the conformational changes of the protein and/or ligand during the simulation event.

**The ligand RMSF for frame  $x$  is:**

Where,  
 $T$  = trajectory time over which the RMSF is calculated  
 $t_{ref}$  = the reference time (usually for the first frame, and is regarded as the zero of time)  
 $r$  = position of atom  $i$  in the reference at time  $t_{ref}$   
 $r'$  = position of atom  $i$  at time  $t$  after superposition on the reference frame

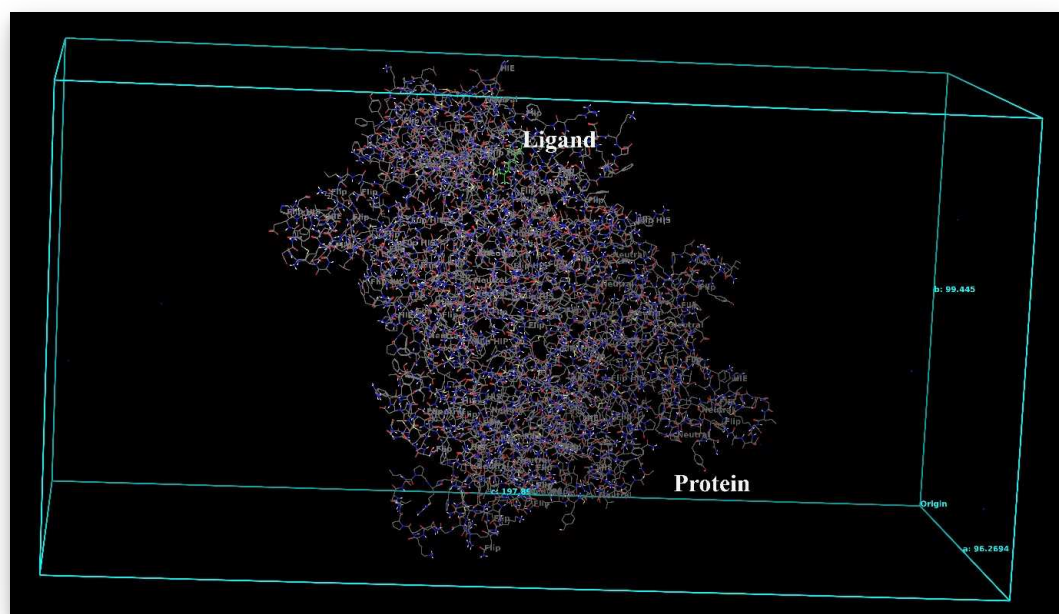
**The protein RMSF for frame  $x$  is:**

Where,  
 $T$  = trajectory time over which the RMSF is calculated  
 $t_{ref}$  = the reference time (usually for the first frame, and is regarded as the zero of time)  
 $r$  = position of atom  $i$  in the reference at time  $t_{ref}$   
 $r'$  = position of atom  $i$  at time  $t$  after superposition on the reference frame

Apart from these, the ligand properties, hydrophobic bond interactions,  $g$ =hydrogen bond interactions, water bridges and ionic bonds between the ligand and residues of the proteins were also identified to study the interactions at a molecular and structural level. Therefore, the present study used this entire



protocol to screen the whole proteome of *A. aegypti*, and used several natural plant-derived bioactives to study the efficacy of their docking and interactions against potential protein targets in the organism to use these naturals as alternative repellents to harmful synthetic ones.



**Molecular dynamic simulation environment containing the boundary box and an example interactive complex**

## CONCLUSION AND SCOPE

### 6 Conclusion and future scope

This protocol was standardized after vigilant observations and trials to screen the whole proteome of an organism to arrive at specific potential targets. Identification of the reference proteome is important since it serves as the representative proteome for related other species of an organism. Selection of the mosquito species was also carried out based on the fact that the proteome of *A. aegypti* encompasses the proteome of other important species of mosquitoes such as *Anopheles* and *Culex* sp, providing a wider scale for searching for important targets. Whole proteome screening revealed that the most important proteins in the organism were those that were involved in the circadian rhythm of the mosquito, expressed between 0-44 hour period of its lifecycle. Hence, a microarray data analysis revealed the expression of these, which along with literature survey helped identify twenty major protein targets. The conserved domain analysis showed important domains common to most of the complete query sequences identified from each of the twenty protein categories, which were then selected for modelling. 309 proteins were modelled and their structures validated. Preliminary docking studies between the natural bioactives, selected controls and all proteins showed negative binding energies. The best ones were docked as a representative protein for each protein category, and the top three docked complexes were simulated at 100ns to unveil stability of docked complexes, indicating high potentiality of the natural molecules to be used as alternatives to synthetic repellents. Additional work such as predicting ligand toxicity and calculating free energy of binding using MMGBSA studies are potential future studies.

