International Journal of Mosquito Research

ISSN: **2348-5906**
CODEN: **IJMRK2**
IJMR 2022; 9(6): 01-14
© 2022 IJMR
www.dipterajournal.com
Received: 01-09-2022
Accepted: 05-10-2022

Hitesh Singh
Ph.D. Student, Department of Genetics, Maharshi Dayanand University, Rohtak, Haryana, India

Manisha Kirar
Ph.D. Student, Department of Genetics, Maharshi Dayanand University, Rohtak, Haryana, India

Neelam Sehrawat
Associate Professor, Department of Genetics, Maharshi Dayanand University, Rohtak, Haryana, India

Corresponding Author:
Neelam Sehrawat
Assistant Professor, Department of Genetics, Maharshi Dayanand University, Rohtak, Haryana, India

In-silico characterization and evolutionary analysis of conserved GPI-anchored protein SGU (Secretory Glycoconjugate of Unknownfunction) in *Anopheles gambiae*

Hitesh Singh, Manisha Kirar and Neelam Sehrawat

DOI: https://doi.org/10.22271/23487941.2022.v9.i6a.634

Abstract
To control malaria, researchers are focused on immunogenic protein to develop a vaccine against this life-threatening disease. The target immunogenic proteins against parasite infection and other diseases are mostly focused on GPI-anchor proteins. AgSGU is secretory glycoconjugate, the second most expressed protein in mosquito midgut after blood feeding. Recent studies confirmed that Pfs47 is the key protein for ookinetes invasion in the mosquito midgut and AGAP006398 act as a receptor for Pfs47. Protein interaction pathway analysis using bioinformatics approaches shows that AgSGU makes a complex with AGAP006398 and Pfs47, so here we target AgSGU protein for its functional and molecular characterization. Blastp analysis showing its conservancy in most *Anopheles* mosquito species. Phylogenetic evolutionary analysis shows that AgSGU is an independently evolved protein and closely related to ACON000570 in *Anopheles coluzzii*. Domain architecture analysis results that it is composed of a single MBF2 domain which acts as a transcriptional activator factor. The 3-D structure was prepared by I-TASSER and refined with the Galaxy Refine tool. The calculated RMSD value for the predicted structure was 0.558, suggesting the reliability of the 3-D structure. AgSGU was docked with AGAP006398 and Pfs47 and interacted effectively with both the protein. The antigenicity evaluation confirmed its potential to activate the immune response. The C-ImmSim immune response analysis proved that AgSGU protein activates both humoral and acquired immune responses. Finally, these studies conclude that AgSGU protein is a potential target for developing a transmission-blocking vaccine against malaria.

Keywords: AgSGU, phylogenetics, homologous, GPI-anchor, transmission-blocking, malaria, etc.

Introduction

Even after enormous efforts to eradicate malaria which is a vector-borne illness brought on by the protozoa of the genus "Plasmodium," remains a serious public health concern [31, 12]. In year 2020, according to an assessment made by the World Health Organization [34], there were 229 million cases and 409,000 fatalities (www.who.int/publications). Malaria infection rates are raising due to the high transmission rate, despite the best efforts and international programmes aimed at eradicating the disease [30]. Although antimalarial treatments and medications were thought to be useful in controlling malaria infections, the development of insecticide and drug resistance in mosquitoes decreased the effectiveness of all preventative interventions [27].

Mosquito genome annotation helps in various efforts initiative to control malaria infection [16, 1]. There are numerous studies has been conducted based on the proteome and genome to develop new malaria-fighting technology [28, 29]. Advancement in the sequencing technology enhances the knowledge about pathway and mechanism of targeted diseases. There are various bioinformatics tools have been used to annotate the structure and function of the proteins.

AgSGU (Secretory Glyconjugate of Unknown Function) is a secretory glycoconjugate protein that is highly expressed in the midgut of *Anopheles gambiae* mosquitoes. It is a Glycosyl-phosphatidylinositols (GPI) anchored protein and a crucial component of the plasma membrane [25]. These specific glycolipids are enabling the attachment of soluble proteins to the plasma membranes.

GPI-anchored proteins are present in eukaryotic organisms and play a variety of biological roles, including host defence, protein-ligand recognition, protein-protein interactions, enzymatic activity, and cell-cell communication [21].

Proteomics analysis in *Anopheles gambiae* confirms that it is the second most highly expressed protein in *Anopheles gambiae* peritrophic matrix after blood feeding [6, 5, 4, 3]. The level of SGU protein expression increases during midgut invasion of *Plasmodium falciparum* in *Anopheles gambiae* and *Anopheles dirus*, [20]. AgSGU, a glycosyl-phosphatidylinositol anchored protein, helps in the development of ookinetes during midgut invasion and parasite attachment over the course of sexual development. Presently various bioinformatics tools were used to confirm its evolutionary history, structure predictions, interaction with other proteins, immunogenic response, etc.

Materials and Method

Protein sequence retrieval

The complete genome of the *Anopheles gambiae* has been sequenced from which uncharacterized Secretory Glycoconjugate of Unknown (AgSGU; AGAP000570) function was selected for this study based on a hypothesis of its importance in ookinetes development and parasite invasion. The protein sequence of *Anopheles gambiae* was retrieved from the NCBI data bank (https://www.ncbi.nlm.nih.gov/protein) in FASTA format. It encoded 167 amino acids. The retrieved protein sequence was subjected to the sequence analysis and homology search in insects.

Homologous protein search and evaluation of conserved amino acid residues pattern

The homologous proteins for AgSGU protein were searched by blastp in the vector base web server (https://vectorbase.org/vectorbase/app/search/transcript/UnifiedBlast). Although NCBI databases were also tried for identifying these homologs but vector base shows good query coverage and high sequence identity. For evaluation of sequence identity, the default parameters were optimized for homologous sequence search with a similarity index >30%. All of the homologous sequences were subjected to multiple sequence alignment to find out conserved residues using the online server Clustal W (https://www.ebi.ac.uk/Tools/msa/clustalw2/).

Evaluation of phylogenetic relationship of AgSGU with homologous protein

The phylogenetic tree was constructed by using MEGA 6.06 software [17] to know about its evolutionary history. All ambiguously positioned amino acids were removed from the data work file. The output file was exported in MEGA format for the construction of the phylogenetic tree. All nonaligned sequences were removed from the queries by using BIOEDIT v7.0.5.3 [11], which is a sequences analysis tool used for alignment, editing, and manipulation of sequences. Divergence of the sequence was also studied through MEGA 6.06 [33], and pair-wise deletion was used to remove all ambiguous amino acid positions.

In-silico structural and functional characterization

Functional domain and motif analysis

The functional domains of AgSGU were searched by NCBI-CD (https://www.ncbi.nlm.nih.gov/Structure/cdd/ wrpsb.cgi), Pfam (https://pfam.xfam.org./), and Prosite (https://prosite.expasy.org/). Reverse Position Specific BLAST (RPS-BLAST) is used to compare a query sequence to position-specific score matrices generated from conserved domain alignments as found in the Conserved Domain Database (CDD). Interpro (https://www.ebi.ac.uk-/interpro/search/sequence/) software was used to predict its signal peptide, cytoplasmic domain, and transmembrane helix. TMHMM 2.0 was used for transmembrane helix prediction and the functional protein motif was analysed using the MOTIF (http://www. genome.jp/tools/motif/) server.

Physiochemical properties

Expasy's ProtParam server (https://web.expasy.org/protparam) was used for the analysis of physicochemical properties of the AgSGU, including the number of amino acids, isoelectric point, and molecular weight. Its GRAVY and aliphatic index were also analysed to know about the stability of the protein [8]. The antigenicity and allergenicity of the protein were also checked by Vaxijen 2.0 server (http://www.ddgpharmfac.net/vaxijen/scripts/VaxiJen_scripts/) and Algpred (https://webs.iiitd.edu.in /cgibin/algpred/) respectively. A hybrid approach was used for allergenicity (SVMc+IgEepitope+ARPs BLAST+MAST).

Secondary and 3-D structure prediction validation

SOPMA was used for secondary structure prediction of protein [9, 10]. The secondary structure study of the protein revealed the percentage of coil-coil, alpha-helix, and extended strand. PSIPRED (http://bioinf.cs.ucl.-ac.uk/psipred/) and ENDscript (http://endscript.ibcp.fr/ ESPript/ENDscript/) were used to validate SOPMA results.

The 3-D structure of the protein was analysed with the online tool I-TASSER [32]. I-TASSER is a threading model-based tool to create 3-D coordinates of proteins. The quality of the model structure was evaluated through PROCHECK (https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/), Verify3D (https://servicesn.mbi.ucla.edu/ Verify3D/), and QMEAN (https://swissmodel.expasy.org/qmean/). The structure was validated by RAMPAGE, GalaxyRefine (https://galaxy.seoklab.org/), and Prosa tool (https://prosa.services.came.sbg.ac.at/prosa.php).

Active site prediction

Computed atlas of surface topography of proteins (CASTp) http://sts.bioe.uic.edu/castp/) CASTp and PAR-3D (http://sunserver.cdfd.org.in:8080/protease/PAR_3D/index.html) software were used to predict the active site of the protein. Both comprehensive and qualitative data for locating and measuring active sites on the interior site and surface of the protein were predicted by CASTp.

Post-translation modification

Post-translational modification occurs in all proteins and plays a significant role in various biological processes by affecting protein structure and function. It happened due to the covalent and enzymatic modification of proteins at a specific position at the time of protein biosynthesis. Most of the modifications are related to the various biological process, so here we study phosphorylation (http://www.cbs.dtu.dk/services/NetPhos/), glycosylation (https://services.healthtech.dtu.dk/service.php?NetNGlyc-1.0), sulfation sites (https://web.expasy.org/sulfinator/) and Myristoylator (https://web.expasy.org/myristoylator/) to predicts N-terminal myristoylation of proteins.

Determination of codon adaptation index

To check the probability of expression of the conserved *AgSGU* protein with the unknown function in the mosquito by using codon adaptation index (CAI) was calculated by the Jcat tool (http://www.jcat.de). The level of expression was compared with the housekeeping gene as a reference.

Analysis of Fold@ index and disorder region of proteins

The local and general probability of the folding pattern of the protein under specified conditions was estimated through the use of the Fold@index tool (https://fold. proteopedia.org). The disorder of the protein was also predicted by ANCHOR [7], PONDR, and Disopred3 (Jones *et al*., 2015) [15]. Here, we study the folding index and disorder of the protein because disordered proteins are mostly related to the disease.

Chromosomal localization and its expression

The chromosomal location of the gene encoding for the AgSGU in *Anopheles gambiae* and its homologous sequence was analysed by tBlASTn against all mosquitoes. The expression of the protein was analysed with the online tool MozAtlas (http://mozatlas.gen.cam.ac.uk).

Molecular docking for protein-protein interaction analysis

STRING software was used for understanding about AgSGU protein interaction pathway. The protein interaction was studied by molecular docking Patchdock (http://bioinfo3d.cs.tau.ac.il/PatchDock) and Firedock tool (http://bioinfo3d.cs.tau.ac.il/FireDock). Molecular docking gives the idea about the binding affinity between proteins.

MD simulation and immune response analysis

To verify the thermo stability and flexibility of AgSGU protein, MD simulation was performed using the iMODS (https://bio.tools/imods) online tool. Immunogenic response against AgSGU protein was also analysed using C-ImmSim (https://www.iac.rm.cnr.it/~filippo/-projects/c-immsim-online.html) tool to confirm it as a target antigen for transmission-blocking vaccine development.

Results

Sequence retrieval of *Anopheles gambiae* AgSGU and its homologous

The sequence of AgSGU was retrieved from the NCBI data bank, it is a single exonic protein encoded for 167 amino acids with sequence identification number AGAP000570. The homologous sequence was searched by Blastp with all non-redundant PDB files. The deep view of its homology search shows that AgSGU was showing homology in most of the *Anopheles* species and show >30% identity with Aedes, *Culex*, and some other insects (Table 1) which are undergoing further analysis.

Conserved amino acid profiling in homologous sequences

The conserved amino acid sequences were analysed (Figure 1) using Clustal W. It revealed that the amino acids Q (128), GG (91-92), T (157), and Y (162) were conserved in all the homologous sequences. It shows SGU protein showing > 50% conservation within most of the Anopheles mosquitoes and conserved sequences are LV, SW, NQSST, T and the conserved repeated sequences are L and VVVV. The functional MBF2 domain is highly conserved in all of the homologous sequences.

Phylogenetic evolutionary relationship analysis

The evolutionary history was analysed by using the Maximum Likelihood method formulated on JTT matrix-based model [14]. The tree constructed with the highest log likelihood (-14649.86) is shown (Figure 2). Initial tree (s) for the heuristic search was obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model and then selecting the topology with a superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 38 amino acid sequences. There were a total of 745 positions in the final dataset. Evolutionary analyses were conducted in MEGA X. In the evolutionary tree (Figure 2) there are mainly two clades from which clade i is unbranched including only single species (ASIS13596), clade ii constructed with twelve independently emerged subclades (a, b, c, d, e, f, g, h, i, j, k, l) from these subclades c, d, f, g, h, i, k,l are branched while rest are branchless. Phylogenetic analysis shows that AGAP000570 is closely related to the ACON000570 in *Anopheles coluzzii* and is highly diverged from ASIS13596.

Divergence evolutions between sequences were also estimated based on the Poisson correction method. In Maximum Likelihood Estimate of Substitution Matrix, the frequency of substitution for each amino acid was 7.69% (A), 5.11% (R), 4.25% (N), 5.13% (D), 2.03% (C), 4.11% (Q), 6.18% (E), 7.47% (G), 2.30% (H), 5.26% (I), 9.11% (L), 5.95% (K), 2.34% (M), 4.05% (F), 5.05% (P), 6.82% (S), 5.85% (T), 1.43% (W), 3.23% (Y), and 6.64% (V).The calculated maximum Log likelihood value for this computation was -14649.856.

Domain Architecture and conserved domain analysis

The AgSGU was composed of a single functional domain MBF2 domain belonging to the IPR031734 family which acts as a transcriptional activation factor for TFII [18] (Figure 3). The Inter Proscan analysis showed that the amino acid sequence from 1-12 is a signal peptide, localized extra-cellularly, and amino acids from 132-157 are embedded in the membrane, which is responsible for pore formation to transducer signals and cellular communication.

Disordered sequence prediction and folding index

There are numerous fundamentally disordered proteins present in the mosquito, so the disorder of AgSGU was analysed. The results were obtained from databases, PONDR, ANCHOR, and Disopred3 for AgSGU protein disorder, there were a total of four disordered region sequences from (1-5), (31-38), (78-111), and (118-148) with average predicted score 0.421 (Figure 4a). Folding @ index showed that there is a 0.306 probability of unfold ability (Figure4b). PONDR result also calculated the hydropathy (Figure 4c) of the protein and net charge on protein.

Post-translation modification plays a very crucial role in biological processes. The major post-translation modifications such as phosphorylation, glycosylation, sulfation, and myristoylation of the target protein AgSGU were analysed.

The phosphorylation site predicted for serine, tyrosine, and threonine residue of the protein (Figure 5a), and these sites predicted for ATM, CKI, CKII, CaMII, DNAPK, EGFR, GSK3, INSR, PK A, PKB, PKC, PKG, RSK, SC, cdc2, cdk5, and p38MAPK enzymes. Two potential glycosylation sites were predicted at amino acid position 50th (NPDL) and 95th (NQQS) (Figure 5b). There was no sulfation modification for AgSGU protein and the N-terminal myristoylation site was also not present because there is no glycine at the end of the protein.

Secondary and 3-D structure prediction and its physiochemical characterization

The secondary structure of the protein was predicted by the PSIPRED, SOPMA, and ENDscript server. From the SOPMA tool, the most prominent form is the random coil (38%) (Figure 6A, B)

It is a single peptide protein containing 169 amino acid residues with molecular weight 18.09 KDa. Its Extinction coefficient is 17420 with an absorbance of 0.963 which represents that in AgSGU all of the cysteines are in reduced form. Its instability was confirmed by the instability index (46.77) and GRAVY Index (0.208).The antigenicity of the protein is 0.572 from Antipro and 0.786 from Vaxijen and proved its potential to induce an immune response.

The 3-D structure of the protein predicted with I-Tessar was shown in Figure (7A). GalaxyRefine was used to fill the gap in the sequence and mission residue from the predicted PDB coordinate. The UCSF chimera calculated RMSD and Mol probity for the predicted structure was 0.558A°and 2.795 confirmed its reliability. The RAMPAGE (Figure 7B) and Prosa Z-score (Figure 7 C) values confirmed that most of the amino acid of AgSGU is in the favourable region showing its stability.

Codon adaptation index for expression level analysis

There is codon biasness in a different organism because of heterogeneity in the usage of codons within the species. The Codon Adaptation Index AgSGU was determined and found to be 0.963 containing 72% GC in comparison with the housekeeping gene 40s ribosome 0.413 with 54.03 GC contains. The CAI index confirmed that the AGAP000570 gene is relatively highly expressed in *Anopheles gambiae* in comparison to the housekeeping gene, so it is a good choice to consider AgSGU protein for further exploration.

Chromosomal localization of AgSGU protein and its homologous sequence

AgSGU gene in *Anopheles gambiae* is present in X chromosome from 10,081,696 to 10,082,373 bp (Figure 8) sequence in a forward direction without any intron. Clustal W analysis revealed that there is no homolog or paralogs present in *Anopheles gambiae* for AgSGU protein. The conserved protein present in other Anopheles mosquitoes (ACOM033520, ACON000570, AQUA010799, AMEM013822, AARA010474, ASTEI04415, AFUN022132, AMAM022625) Aedes (AAEL013885, AALF009866) and with some other insects (LLOJ005584, PPAI010383, LLOJ006207, MDOA016020) theirs chromosomal localization is shown in table-1.

Protein-protein interaction

To understand AgSGU interaction with other proteins, the STRINGS database was used. STRINGS results are based on the experimental data, co-expression, text mining, and its co-occurrence. All parameters are included for protein-protein interaction analysis. There is a total of 23 active binding sites for interaction with other proteins and two major binding sites with an area 101.2 A^0 and volume 80.860 A^0 composed of Ser (84), Ala (85), Ser (86), Glu (87), Leu (89), Leu (107), Leu (109), Gln(110), Tyr (111), Val (115), Gly (117), Val (119), Gln (147) and the other major binding pocket was with 101.69A^0 areas and 42.78 A^0volume formed with amino acid sequence Phe (20), Val (23), Ile (24), Glu (54), Val (66),Ala (68), Gly (69), Gly (70), Ser (71), Thr (72), Gln (74), Gln (75), Ile (76), Val (77) (Figure 9).

Based on STRINGS databases it was confirmed that there are four interacting proteins AGAP002216 (cation transport regulator-like protein 2), AGAP006398 (unspecified product), AGAP000570 (Secretory glycoconjugate protein) and AGAP007745 (unspecified product) as shown in figure 10. Recently it was confirmed that AGAP006398 acts as key a receptor for ookinetes invasion by interacting with Pfs47 (Molina-Cruz *et al*., 2020) [22]. STRING database shows that AGAP007745 is co-expressed along with the AGAP006398 receptor. After interaction of Pfs47 with AGAP006398, the expression of AGAP000570 was increased and involved with the ookinetes invasion and development inside the mosquito midgut (Mithaes *et al*., 2014) [20]. As the AGAP000570 expressed the permeability of the midgut membrane change and allow the ookinetes inside the midgut with the activation of AGAP002216.

Molecular docking was performed with both proteins to confirm AgSGU interaction directly with Pf47, which is a key protein for ookinetes invasion through bypassing the mosquito immune system or may interact with Pf47 receptor AGAP006398 (Agrec47). Docking analysis confirmed that AGAP000570 (AgSGU) makes a complex with AGAP006398 and Pfs47 shown in figure 11 and helps in ookinetes invasion and developments inside mosquito midgut (Table -2).

MD simulation for determination of thermostability and flexibility of AgSGU protein and its immune response

The deformability index (Figure 12 A) is the measure of flexibility, a higher peak of a sequence showing the flexibility of the residue. B-factor (Figure 12 B) is the average of RMSD value and calculated from NMA obtained from multiply NMA mobility by 8pi^2[]. The covariance matrix (Figure E) identifies the degree of pairing between residues and the related residues are shown in red and unrelated residues in white color. The elastic network model (Figure 12 C) shows the pairs of atoms connected with strings; in graph one spring is representing the pair of atoms. Eigenvalue associated (Figure 12 D) with the motion stiffness, the higher the Eigenvalue more is the stiffness.

Immune response

Molecular docking confirmed that AgSGU protein binds efficiently with the recAg47, which is act as a key receptor for ookinetes invasion inside mosquito midgut, so here the immunogenic response of the target protein AgSGU was analyzed and shown in Figure 13. IgG1 is highly expressed and reaches a maximum concentration 7-10 days of immunization. T helper cells activated after 5 days of immunization and remain constant for up to 35 days. Tcytotoxic cell activated after 10 days of immunization. C-Imm Simm analysis confirmed that our target protein produces an effective immune response and becomes the target candidate for the transmission-blocking vaccine.

Table 1: Homology search of AgSGU with *Anopheles* species and Aedes, *Culex* and some other insects.

Protein Id	Species	Amino acid residues	Exon No	Mol. Weight	Isoelectric point	e-value	Identity	Location
ACOM033520	Anopheles_coluzzii	274	2	29095	8.87	2.00E-114	100%	EQ090156:42352-44016(+)
ACON000570	Anopheles_coluzzii	169	1	18096	7.16	5.00E-116	100%	RWKB01000025:401391-401900(+)
AQUA010799	Anopheles_quadriannulatus	169	1	18075	8	2E-113	98%	KB668177:614389-614898(-)
AMEM013822	Anopheles_merus	169	1	18047	7.16	5E-106	98%	KI915274:318299-318808(+)
AARA010474	Anopheles_arabiensis	169	1	18022	7.13	5E-105	98%	KB704784:4788330-4788839(-)
AMEC002131	Anopheles_melas	232	5	25027	8.9	2E-96	80%	AXCO02016608:5249-6244(+)
ACUA014121	Anopheles_culicifacies	168	1	17777	5.81	8E-73	65%	KI422624:73771-74277(+)
ASTE001198	Anopheles_stephensi	167	1	17915	5.96	4E-72	64%	KB665176:1767600-1768103(-)
ASTEI04415	Anopheles_stephensi	167	1	17915	5.96	4E-72	64%	KE388912:366804-367307(-)
AFUN022132	Anopheles_funestus	168	1	17892	6.28	2E-68	66%	AfunF3_X:12954728-12955234(-)
AEPI009084	Anopheles_epiroticus	164	1	17506	5.08	7E-68	77%	KB671479:92787-93281(+)
AMAM022625	Anopheles_maculatus	167	1	17870	6.21	1E-66	64%	AXCL01026025:493-996(+)
AMIN005326	Anopheles_minimus	168	1	18000	5.66	3E-64	65%	KB664054:3578717-3579223(-)
AFAF004320	Anopheles_farauti	168	1	17809	5.1	6E-57	62%	KI915062:259855-260361(+)
ASIC003926	Anopheles_sinensis	167	1	18022	4.48	8E-54	61%	KE524793:155751-156254(-)
ASIS013596	Anopheles_sinensis	747	8	81446	7.27	8E-49	61%	KI916604:71205-75724(-)
CPIJ018795	Culex_quinquefasciatus	166	2	18116	7.3	2E-41	50%	DS233096:60718-61274(+)
AAEL013885	Aedes_aegypti	158	2	17136	5.66	2E-39	52%	AaegL5_3:303464102-303464644(+)
AALF009866	Aedes_albopictus	158	2	17179	4.78	1E-38	53%	JXUM01S000020:668893-669432(-)
CPIJ018793	Culex_quinquefasciatus	153	1	16448	5.79	7E-38	55%	DS233096:57094-57555(-)
CPIJ018794	Culex_quinquefasciatus	153	1	16490	5.02	9E-38	55%	DS233096:59928-60389(+)
AALF004765	Aedes_albopictus	172	2	17294	8.03	2E-35	50%	JXUM01S001398:148569-149087(-)
AAEL004809	Aedes_albopictus	155	2	16550	9.58	2E-31	57%	AaegL5_1:17498094-17498626(+)
LLOJ005584	Lutzomyia_longipalpis	154	1	16512	7.87	1E-31	47%	JH689677:26348-26812(-)
PPAI010383	Phlebotomus_papatasi	161	2	18110	6.72	2E-27	37%	JH661024.1:114614-115164(+)
LLOJ006207	Lutzomyia_longipalpis	167	2	18693	6.98	2E-26	42%	JH689742:26336-26898(+)
MDOA005682	Musca_domestica	167	2	18527	6.01	1E-19	32%	KB855323:394791-398879(+)
SCAU004971	Stomoxys_calcitrans	177	2	19533	9.27	2E-19	35%	KQ079927:374578-378238(-)
GBRI012435	Glossina_brevipalpis	159	2	18309	5.24	2E-17	34%	KK351028:284102-286507(+)
GFUI027230	Glossina_fuscipes	169	4	19224	6.23	3E-17	34%	KK351819:666499-672971(+)
GPPI018114	Glossina_palpalis	168	2	19051	7.14	2E-16	34%	KN796263:78647-80289(-)
GMOY007637	Glossina_morsitan	168	2	18926	6.64	2E-15	32%	GmorY1_scf7180000650567:72602-73962(-)
PPAI008801	Phlebotomus_papatasi	151	1	16816	8.38	9E-15	31%	JH666053.1:1490-1945(-)
GAUT033493	Glossina_austeni	168	1	18944	6.48	2E-14	30%	KK502441:187633-189050(+)
GPAI024588	Glossina_pallidipes	168	4	18801	6.48	6E-14	30%	KK500059:301641-303073(-)
LLOJ008878	Lutzomyia_longipalpis	158	1	17438	5.88	1E-12	33%	JH690093:15378-15854(+)
MDOA016020	Musca_domestica	147	2	16333	6.44	3E-10	30%	KB855891:48154-48656(+)

Table 2: AgSGU protein interaction with AGAP006398 and Pfs47 showing its binding energy.

Protein	Global energy	Attractive VdW	Repulsive VdW	ACE	HB
AGAP006398	-54.4	-36.67	30.21	-6.36	-4.62
Pfs47	-42.56	-14.29	20.3	-4.31	-3.4

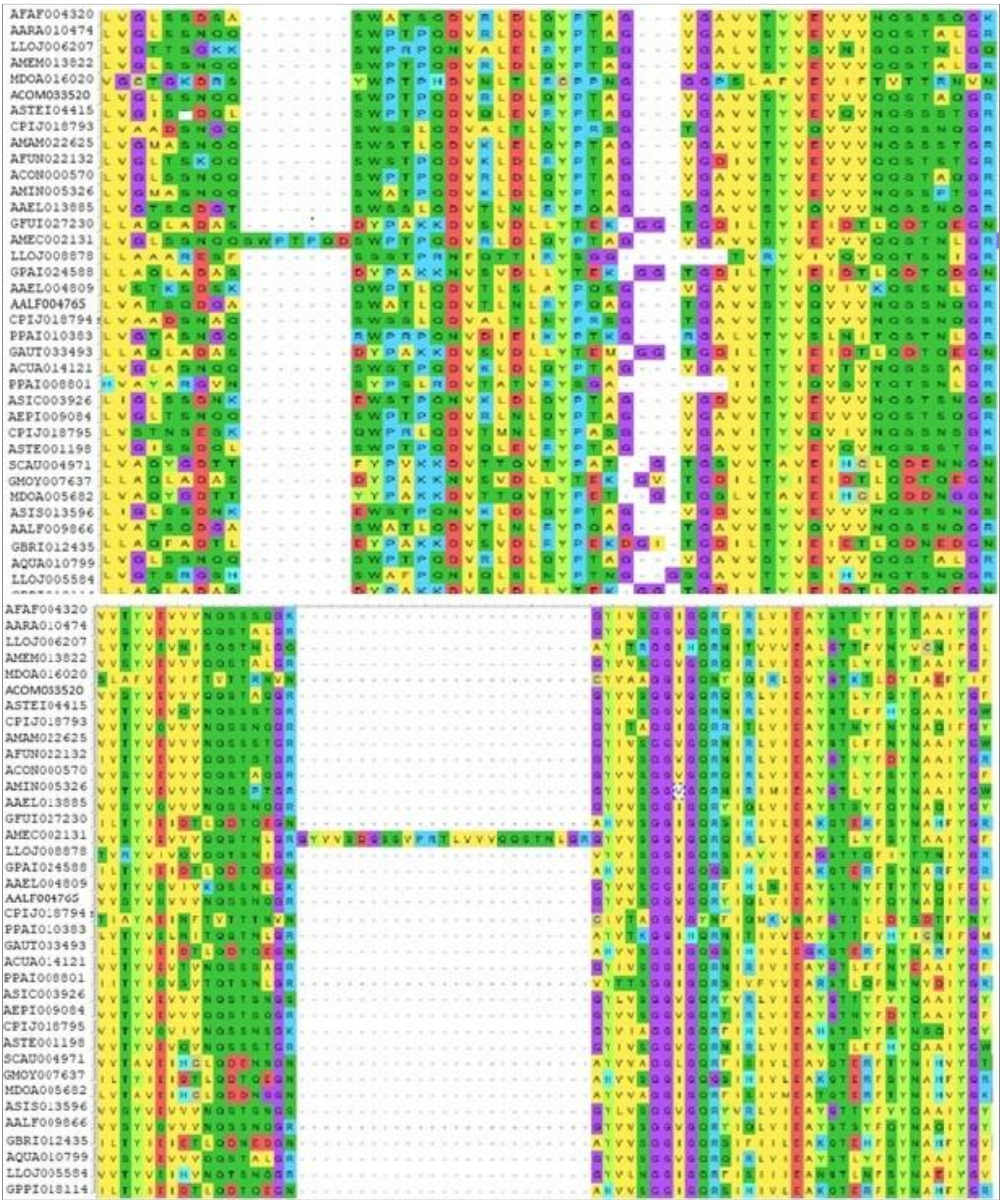

Fig 1: Clustal W result analysis showing conserved amino residues throughout all selected species.

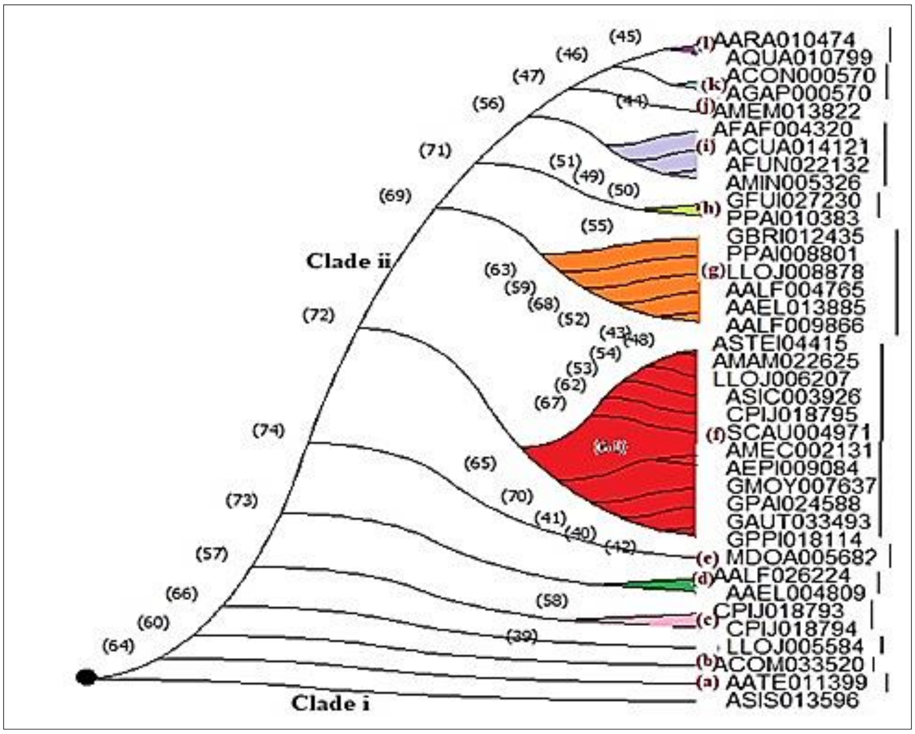

Fig 2: Phylogenetic analysis of SGU protein in mosquito and some other insects species.

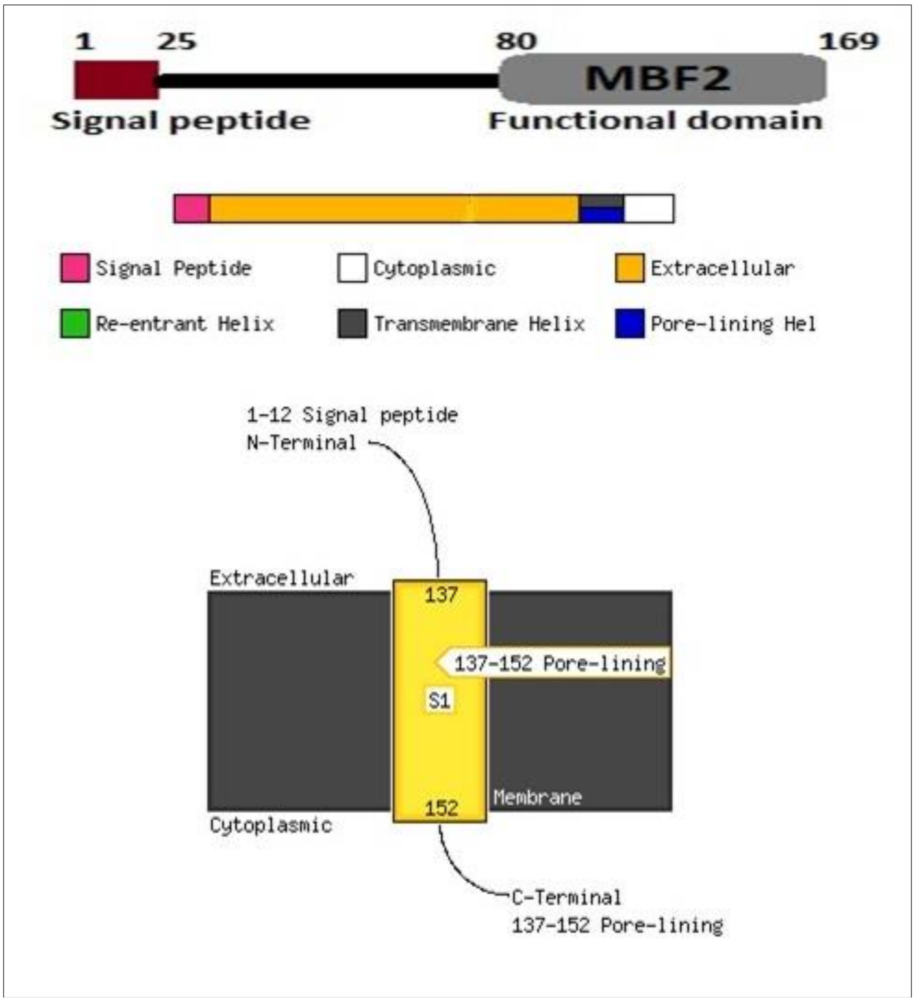

Fig 3: The domain architecture analysed with Pfam for AgSGU protein.

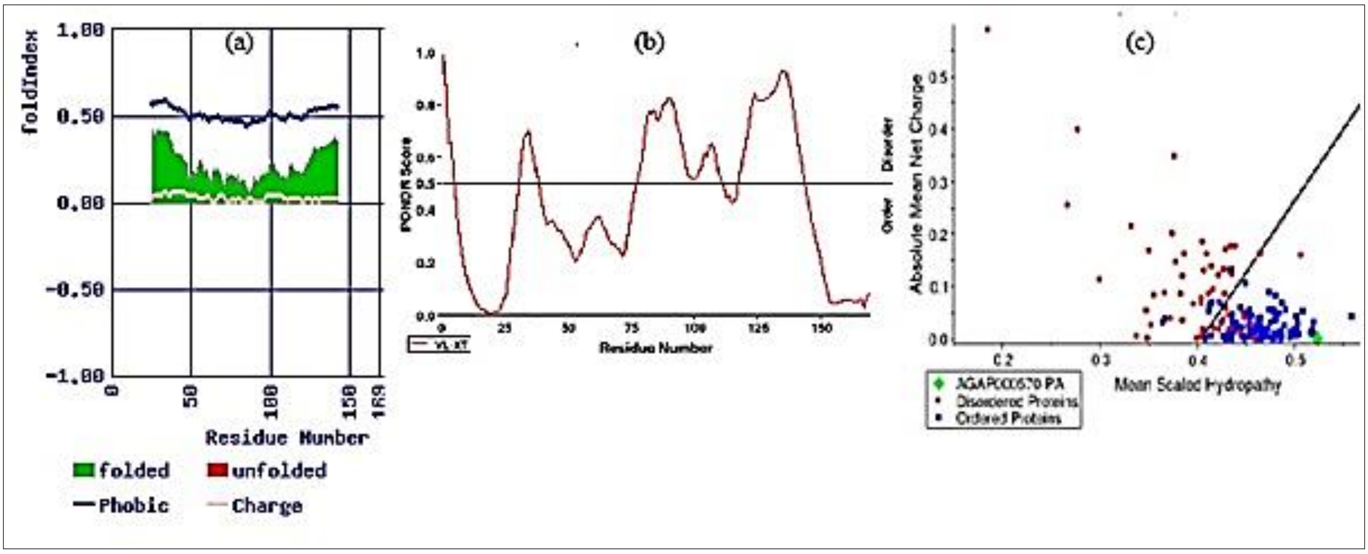

Fig 4: Disorderness and folding index prediction. In figure 4a folding pattern of protein was shown in green colour. The figure 4b is showing the disordered region of the protein and protein hydropathy and net charge of the protein showing in figure 4c. Phophorylation, Glycosylation and other post-translation modification site prediction.

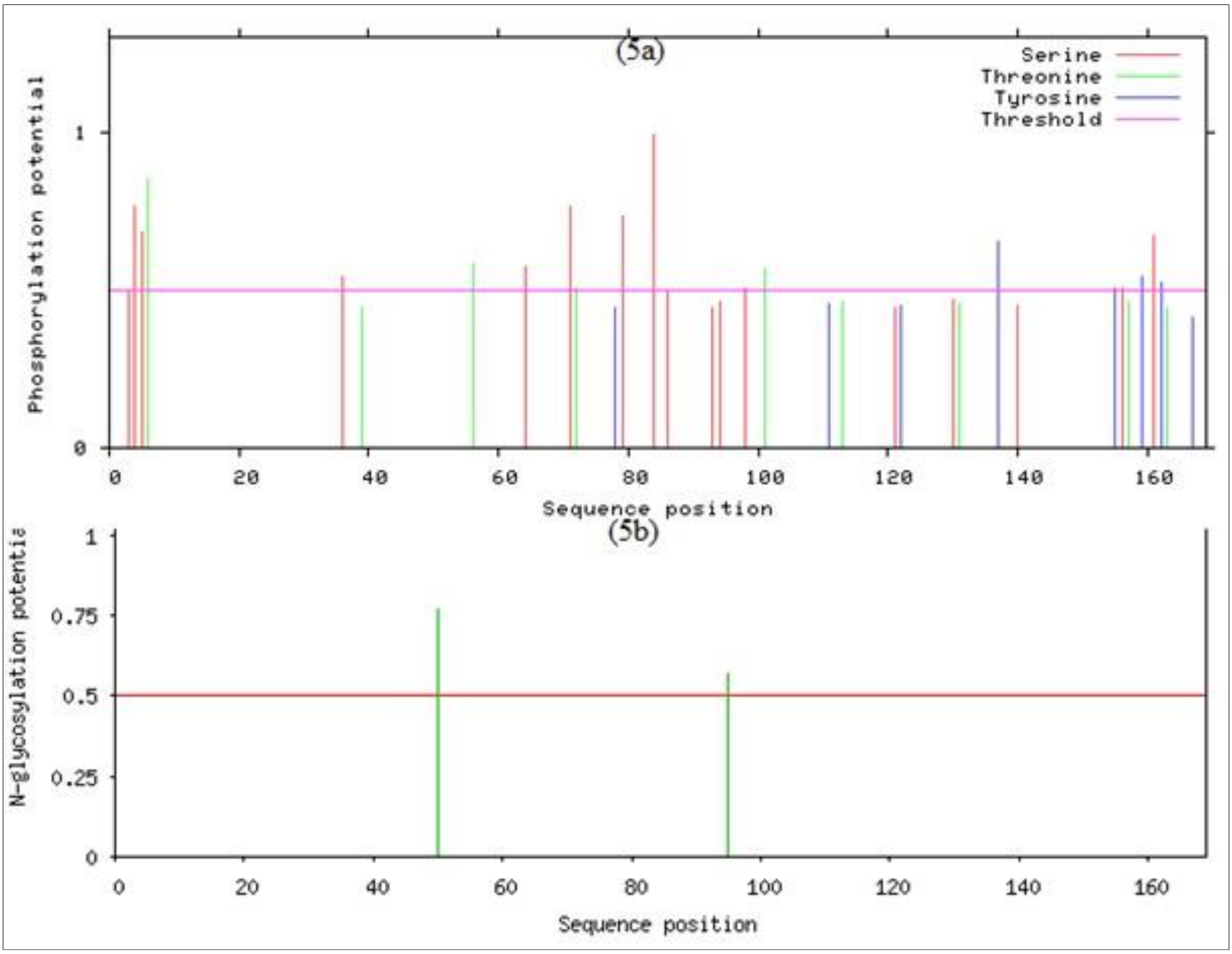

Fig 5: Post-translation modification prediction. Figure 5a showing phosphorylation site at serine, theonine and tyrosine residue of the protein, Figure 5b representing glycosylation site of the protein.

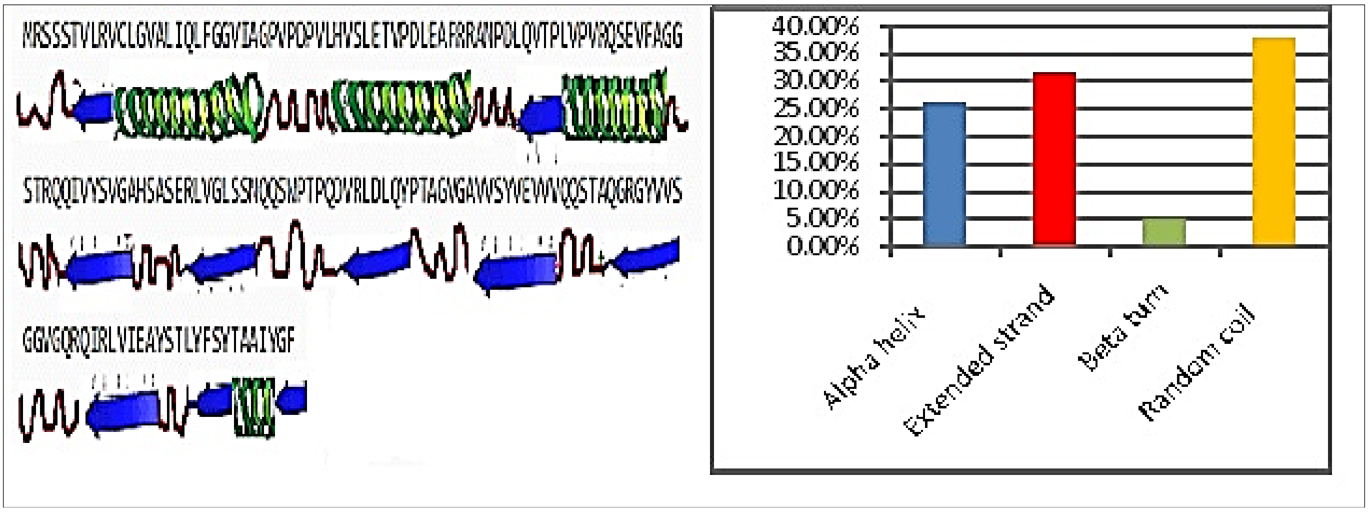

Fig 6: SOPMA secondary structure prediction for SGU protein in *Anopheles gambiae* to study the amino acid sequence in Alpha helix region, beta turn and extended strand region (A) the figure (B) showing the percentage of regions graphically.

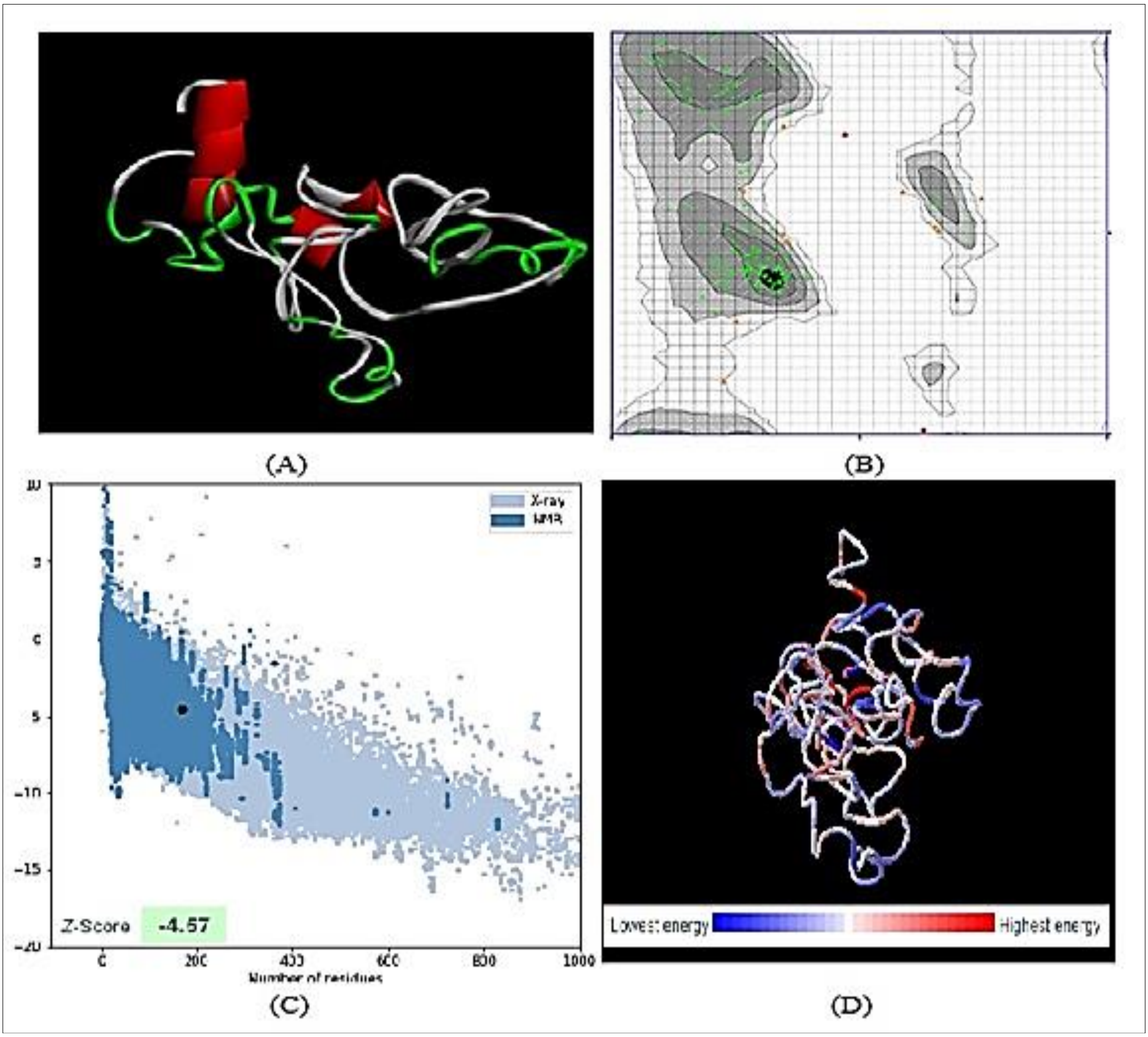

Fig 7: 3-D structure prediction and validation. (A) 3-D structure predicted from I-Tessar and 3-Dpro (B) RAMPAGE analysis for ramachandaran plot showing green colour for highly favourable region, brown for preferred observable region and red for unfavourable region (C) Prosa result for Z-score value (D) Energy graph based on protein amino acid sequence.

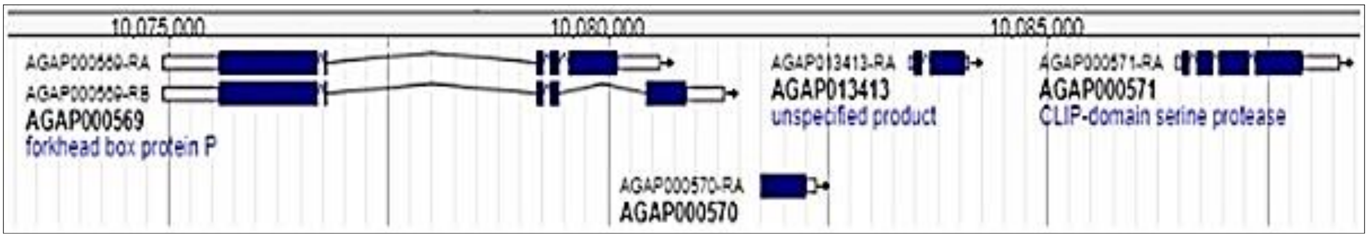

Fig 8: Chromosomal localization of AgSGU gene in *Anopheles gambiae* genomic supercontig in respective of their neighbour gene. In this figure blue box representing the exonic region and white box represent UTR’S and introns are shown with the black line.

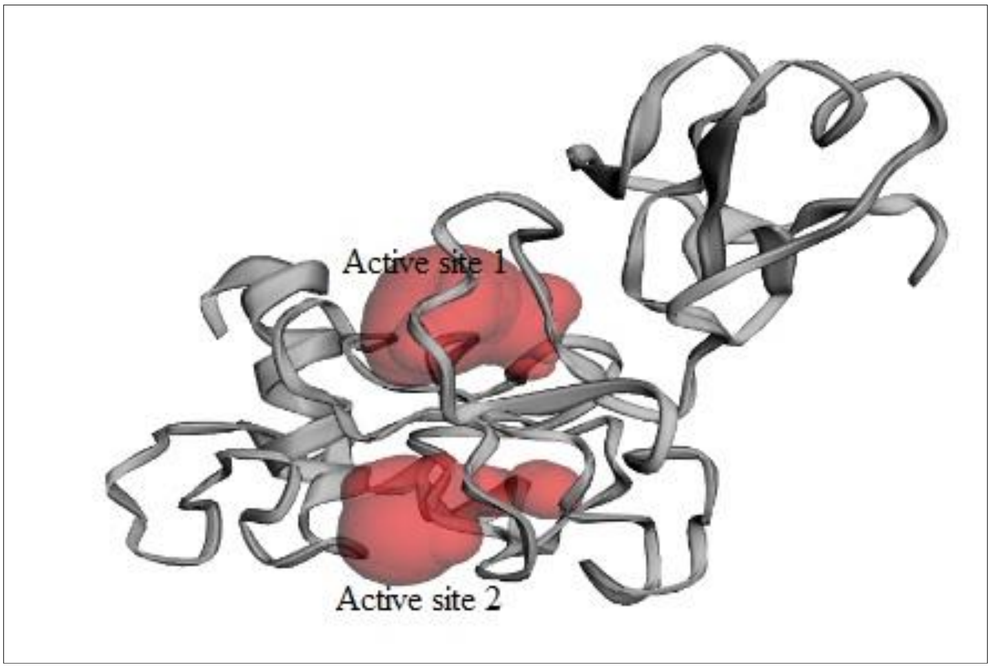

Fig 9: CASTp binding pocket analysis results showing that in AgSGU there are two binding pockets showing in red colour, major binding pockets (1) and minor binding pocket (2).

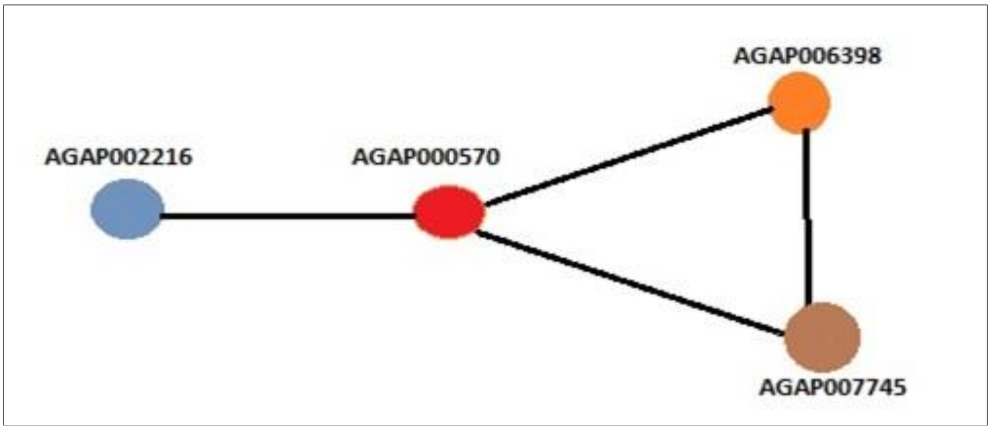

Fig 10: Protein-Protein interaction of four proteins AGAP002216, AGAP000570 AGAP006398 and AGAP007745.

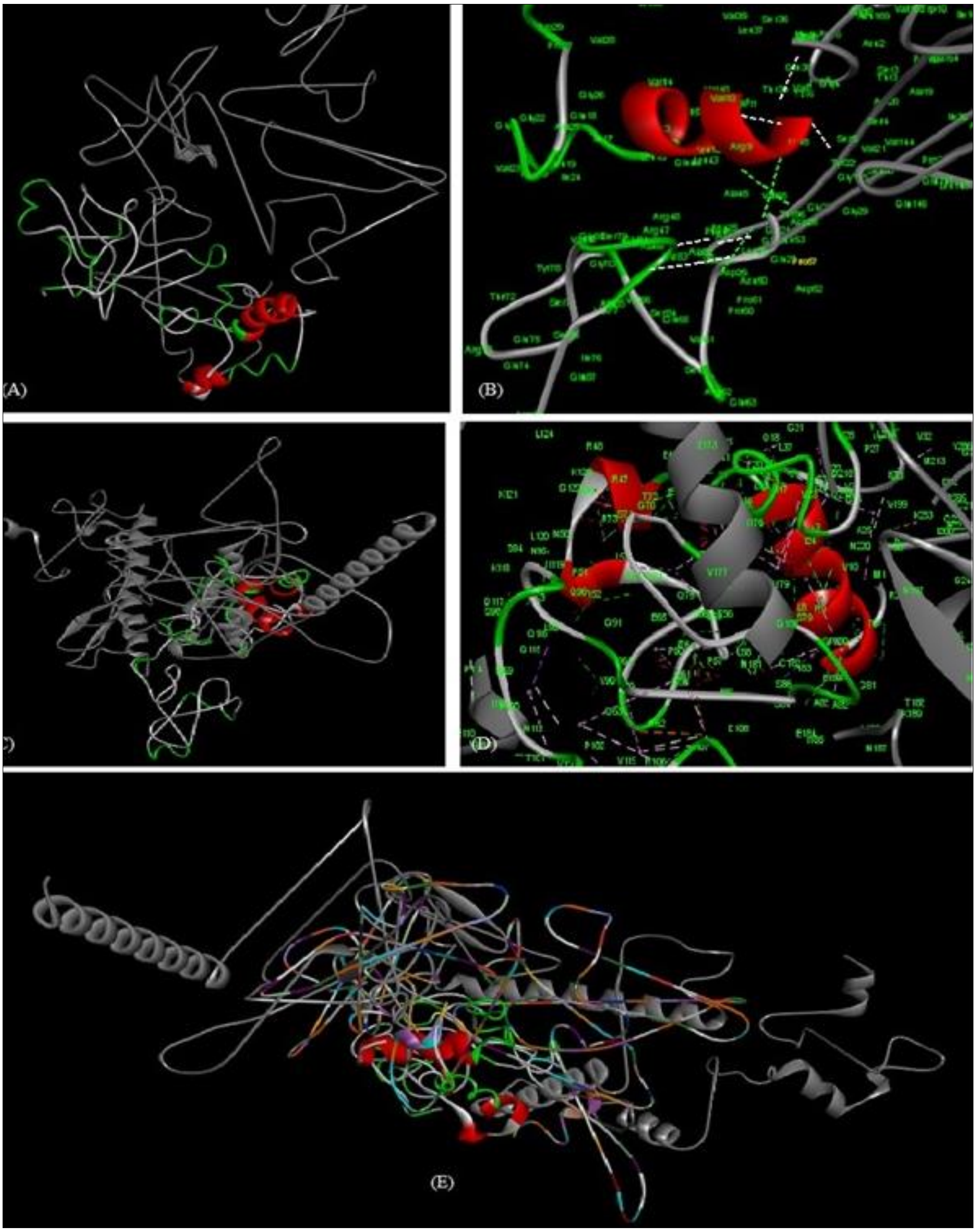

Fig 11: AgSGU interaction with AGAP006398 (Agrec47). In figure (A) AgSGU showing in red and grey colour and AGAP006398 in grey. Figure (B) bonding pattern between AgSGU and AGAP006398 was shown green line showing hydrogen bonding. Figure (C) and (D) is protein interaction between AgSGU and Pfs47. In Figure (E) there is graphical representation superimposing of AgSGU, AGAP006398 and Pfs47.

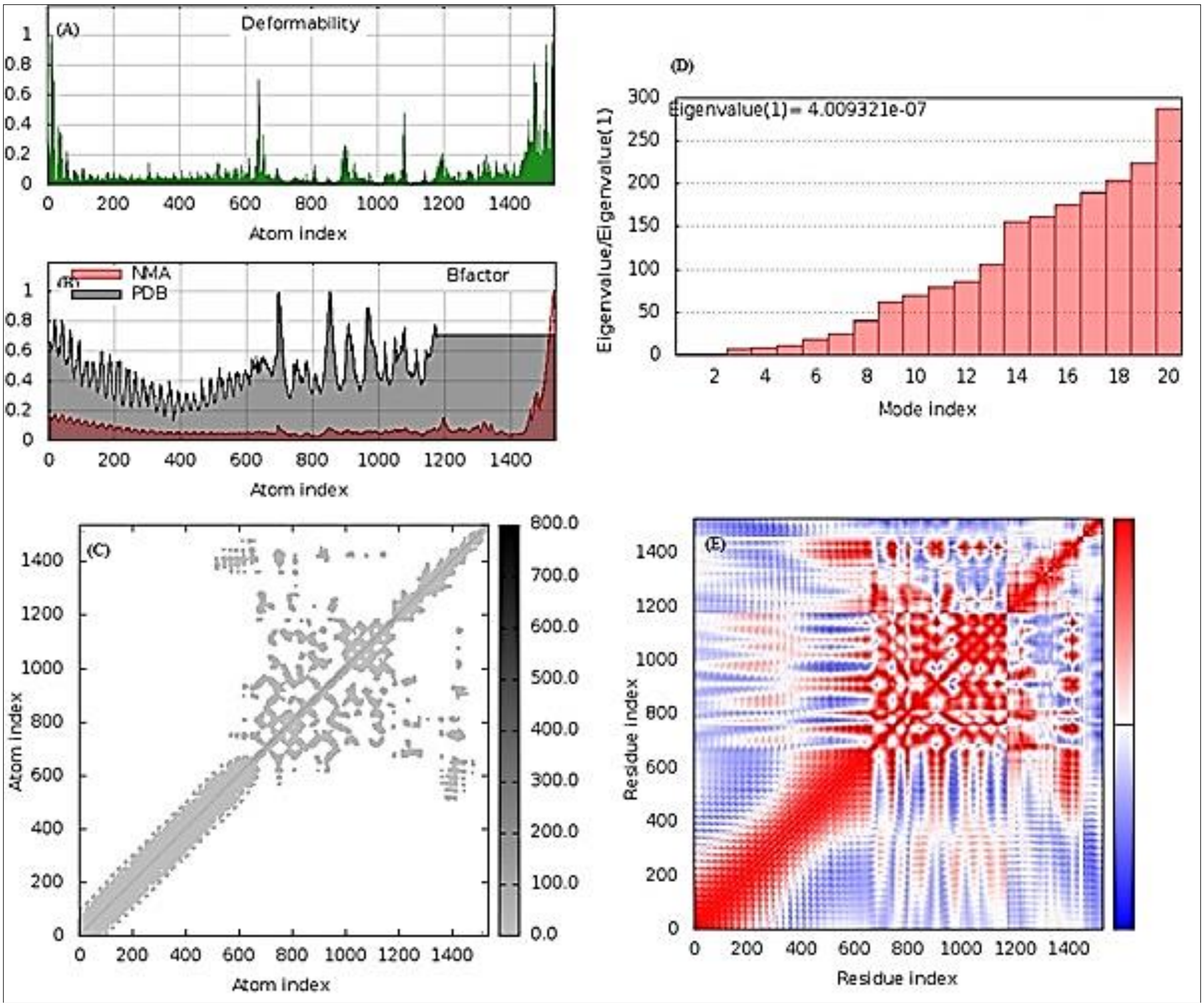

Fig 12: MD simulation analysis from iMODS. Figure 12A showing the deformability index of the protein representing highly deform region in the form of hinges. Here, calculated B-factor showing in Figure 12B. Elastic network represented in Figure 12C deep grey dot for pair of linked residues. Eigen value representing in figure 12 D. Figure 12 E representing covariance between the residue, the residue in red colour are related with each other, white for unrelated residues and blue colour showing anti-related residues.

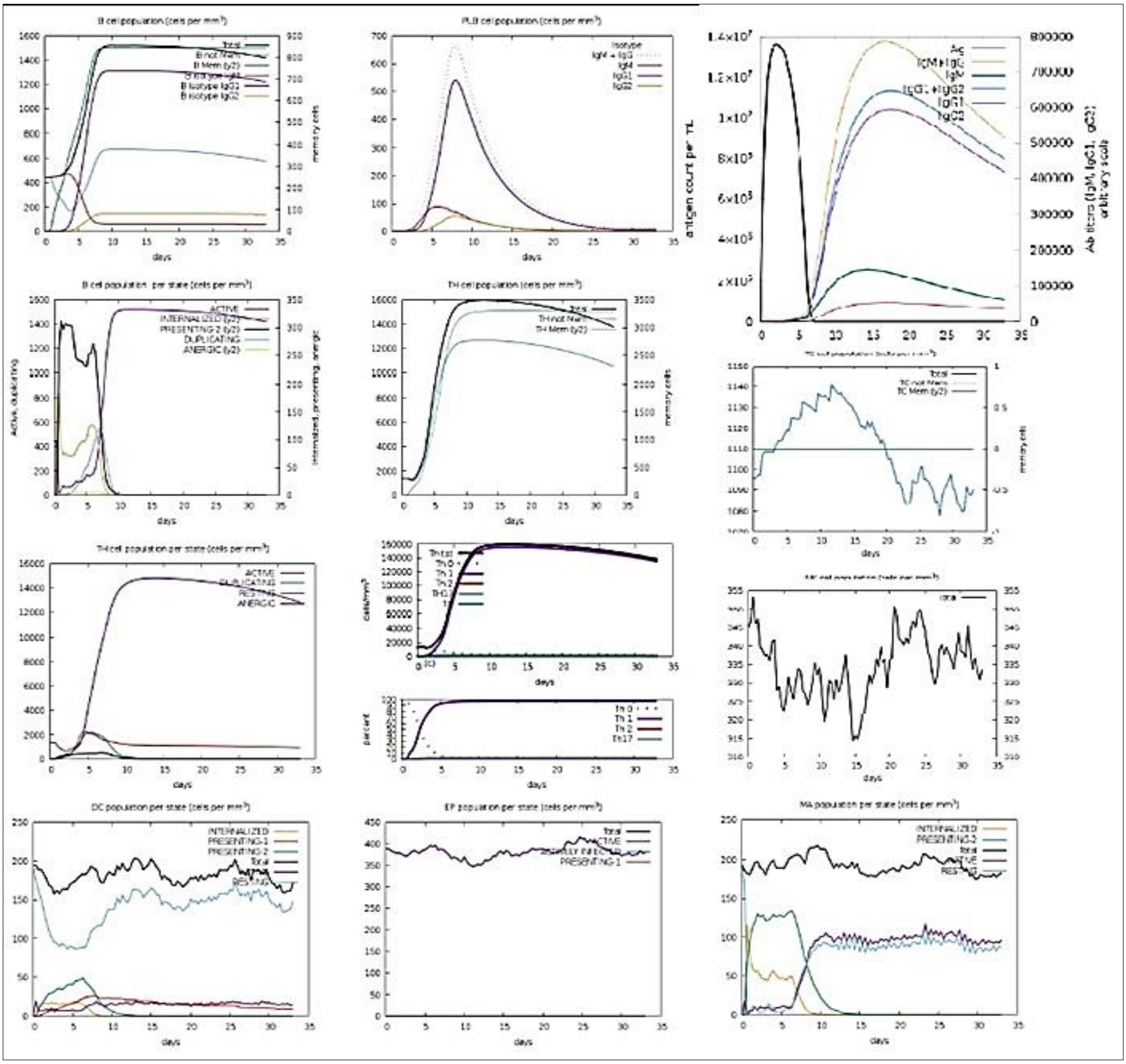

Fig 13: Immune response against AgSGU protein calculated with C-Imm Simm tool.

Conclusion

Anopheles gambiae is the primary malaria vector across the world. AgSGU is a GPI-anchored protein expressed in the midgut specifically after blood feeding. GPI- anchored proteins currently become the target for vaccine development to control various diseases. *In-silico* analysis shows that mosquito SGU protein interacts with ookinete receptor recAg47 and helps in parasite midgut invasion. To confirm its validity as a target antigen for the transmission-blocking vaccine, AgSGU protein is subjected to its evolutionary relationship and functional and structural characteristics. Clustal W analysis shows that AgSGU protein is conserved in most of mosquitoes. The physiochemical study confirmed stability from the GRAVY index of the protein its half-life is more > 20hr. The secondary structure of the protein is mainly composed of random coil, Beta turn, extended strand, and alpha helix.3-D structure of the protein predicted from I-Tessar and 3 Dpro and validated from Ramachandran plot, it confirmed those 91.6% amino acids are present in the favourable region. Molecular docking confirmed that AgSGU binds effectively to recAg47 with free energy -54.6 Kcal/mol. Comprehensive analysis of AgSGU protein confirmed it as a target antigen for the transmission-blocking vaccine.

Declaration

Authors declare that there are no conflicts between authors.

References

1. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, *et al*. Plasmo DB: The Plasmodium Genome Resource. A Database Integrating Experimental and Computational Data. Nucleic Acids Res. 2003; 31(2):212-215.
2. Choi HP, Juarez S, Ciordia S, *et al*. Biochemical characterization of hypothetical proteins from *Helicobacter pylori*. PLoS ONE. 2013; 8(6):e66605.
3. Dinglasan RR, Jacobs-Lorena M. Flipping the paradigm on malaria transmission-blocking vaccines. Trends in parasitology. 2008; 24(8):364-370.
4. Dinglasan RR, Fields I, Shahabuddin M, Azad AF, Sacci

JB. Monoclonal antibody MG96 completely blocks *Plasmodium yoelii* development in *Anopheles stephensi*. Infection and immunity. 2003;71(12):6995-7001.

5. Dinglasan RR, Kalume DE, Kanzok SM, Ghosh AK, Muratova O, Pandey A, *et al*. Disruption of Plasmodium falciparum development by antibodies against a conserved mosquito midgut antigen. Proceedings of the National Academy of Sciences. 2007;104(33):13461-13466.
6. Dinglasan RR, Valenzuela JG, Azad AF. Sugar epitopes as potential universal disease transmission blocking targets. Insect biochemistry and molecular biology. 2005;35(1):1-10
7. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics. 2009;25(20):2745-2746.
8. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. In The proteomics protocols handbook. Humana press; c2005. p. 571-607.
9. Geourjon C, Deleage G. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics. 1995;11(6):681-684.
10. Guermeur Y, Geourjon C, Gallinari P, Delage G. Improved performance in protein secondary structure prediction by inhomogeneous score combination. Bioinformatics (Oxford, England). 1999;15(5):413-421.
11. Hall TA. Bio Edit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 1999;41:95-98.
12. Hema K, Ahamad S, Joon HK, Pandey R, Gupta D. Atomic Resolution Homology Models and Molecular Dynamics Simulations of Plasmodium Falciparum Tubulins. ACS Omega. 2021;6(27):17510-17522.
13. Idrees SNS, Kanwal S, Ehsan B, Yousaf A, Nadeem SMIR. *In silico* sequence analysis, homology modeling and function annotation of Ocimum basilicum hypothetical protein G1CT28_OCIBA. Int J Bioautomation. 2012;16(2):111-118.
14. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 1992 Jun 1;8(3):275-82.
15. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015 Mar 15;31(6):857-63.
16. Kissinger JC, Brunk BP, Crabtree J, Fraunholz MJ, Gajria B, Milgram AJ, *et al*. The Plasmodium Genome Database. Nature. 2002;419(6906):490-492.
17. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X. Molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution. 2018;35(6):1547-1549.
18. Liu D, Ishima R, Tong KI, Bagby S, Kokubo T, Muhandiram DR, *et al*. Solution structure of a TBP–TAFII230 complex: protein mimicry of the minor groove surface of the TATA box unwound by TBP. Cell. 1998;94(5):573-583.
19. Lubec G, Afjehi-Sadat L, Yang JW, John JP. Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol. 2005;77(1-2):90-127.
20. Mathias DK, Jardim JG, Parish LA, Armistead JS, Trinh HV, Kumpitak C, *et al*. Differential roles of an Anopheline midgut GPI-anchored protein in mediating *Plasmodium falciparum* and *Plasmodium vivax* ookinete invasion. Infection, Genetics and Evolution.2014; 28:635-647.
21. McConville MJ, Ferguson MA. The structure, biosynthesis and function of glycosylated phosphatidy linositols in the parasitic protozoa and higher eukaryotes. Biochemical Journal. 1993 Sep 1;294(Pt 2):305.
22. Molina-Cruz A, Canepa GE, e Silva TLA, Williams AE, Nagyal S, Yenkoidiok-Douti L, *et al*. Plasmodium falciparum evades immunity of anopheline mosquitoes by interacting with a Pfs47 midgut receptor. Proceedings of the National Academy of Sciences. 2020;117(5):2597-2605.
23. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008;92(5):255-264.
24. Nimrod G, Schushan M, Steinberg DM, Ben-Tal N. Detection of functionally important regions in hypothetical proteins of known structure. Structure (London, England: 1993). 2008;16(12):1755-1763.
25. Paulick MG, Bertozzi CR. The glycosylphosphatidylinositol anchor: A complex membrane-anchoring structure for proteins. Biochemistry. 2008;47(27):6991-7000.
26. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. PNAS. 1999;96(8):4285-4288.
27. Ramasamy G, Gupta D, Mohmmed A, Chauhan VS. Characterization and Localization of Plasmodium Falciparum Homolog of Prokaryotic ClpQ/HslV Protease. Mol. Biochem. Parasitol. 2007;152(2):139-148.
28. Sardar R, Katyal N, Ahamad S, Jade DD, Ali S, Gupta D. In-silico Profiling and Structural Insights into the Impact of nSNPs in the P. Falciparum Acetyl-CoA Transporter Gene to Understand the Mechanism of Drug Resistance in Malaria. J. Biomol. Struct. Dyn. 2021;39(2):558-569.
29. Sourabh S, Chauhan M, Yasmin R, Shehzad S, Gupta D, Tuteja R. *Plasmodium falciparum* DDX17 Is an RNA Helicase Crucial for Parasite Development. Biochem. Biophys. Rep. 2021;26:101000.
30. Sumner KM, Freedman E, Abel L, Obala A, Pence BW, Wesolowski A, *et al*. Genotyping Cognate *Plasmodium Falciparum* in Humans and Mosquitoes to Estimate Onward Transmission of Asymptomatic Infections. *Nat.* Commun. 2021;12(1):1-12.
31. Uwimana A, Legrand E, Stokes BH, Ndikumana JLM, Warsame M, Umulisa N, *et al*. Emergence and Clonal Expansion of *In vitro* Artemisinin-Resistant *Plasmodium falciparum* Kelch13 R561H Mutant Parasites in Rwanda. Nat. Med. 2020;26(10):1602-1608.
32. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature methods. 2015;12(1):7-8.
33. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In Evolving genes and proteins. Academic Press; c1965. p. 97-166.
34. World Health Organization; c2020.