Article

BIOEDUSCIENCE
ISSN: 2614-1558

http://journal.uhamka.ac.id/index.php/bioeduscience

Identification of Uncharacterized *Plasmodium falciparum* Proteins via In-silico Analysis

Vianney Widjaja [1] , Albert Lim [1] , Benedicta Aini [1] , Gabrielle Audrey Gandasasmita [1] , Jeremie Theddy Darmawan [2] , and Arli Aditya Parikesit [2*]

[1] Department of Biomedicine, School of Life Sciences, Indonesia International Institute for Life Sciences, Jl. Pulomas Barat Kav.88, Jakarta, Indonesia, 13210.
[2] Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jl. Pulomas Barat Kav.88, Jakarta, Indonesia, 13210.
[*] Correspondence: arli.parikesit@i3l.ac.id

Abstract

Background: However, it is globally known that Malaria is caused by the Plasmodium parasite, mainly lethally by *Plasmodium falciparum*. This research aims to understand the structure and function of three uncharacterized *P. falciparum* proteins (PF3D7_1468000, PF3D7_1147400, PF3D7_1351100) using bioinformatic methods in hopes of learning more about Malaria. **Methods:** The three uncharacterized *P. falciparum* proteins were inserted into Phyre2 for knowing the protein homology, InterPro, and SUPERFAMILY hidden Markov models for understanding the domain annotation, scanprosite for knowing the post-translational modification, Ramachandran plot for protein validation, and Yasara for visualizing the protein. **Results:** The third protein showed the highest confidence and coverage level of 100%, followed by the second protein, and the lowest was the first protein. Interpro and SUPERFAMILY results identified the first protein as the WD40 repeat superfamily, the second protein as Cytochrome C subunit II-like, and the third protein as the CXXC motif. Scanprosite revealed all sequences possessing protein domains in which the first protein has three protein domains, the second protein has one protein domain, and the third protein has two protein domains. According to the Ramachandran plot, the first and second protein generally has an α-helix structure. In contrast, the third protein's overall β-sheet structure differs somewhat from the protein structure visualization. The three protein visualizations exhibited secondary structures and more than 50 amino acid residues for each protein. **Conclusions:** The second and third uncharacterized proteins (PF3D7_1147400, PF3D7_1351100) could be promising antimalarial drug targets leading to the *P. falciparum* parasite death.

Keywords: Bioinformatics; Malaria; *Plasmodium falciparum*; Uncharacterized proteins

Article history
Received: 14 March 2022
Accepted: 9 May 2022
Published: 31 Aug 2022

Publisher's Note:
BIOEDUSCIENCE stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Citation: Widjaja, V., Lim, A., Aini, B., Gandasasmita, G.A., Darmawan, J.T., & Parkesit, A.A. (2022). Identification of Uncharacterized Plasmodium falciparum Proteins via In-silico Analysis. *BIOEDUSCIENCE, 6*(2), 198-210. doi: 10.22263/j.bes/628770

Introduction

Malaria is one of the most lethal parasitic infections transmitted to humans by the female *Anopheles* mosquito infected with the Plasmodium parasites (WHO, 2021a). Malaria impacted 229 million people globally in 2019; according to the World Health Organization (WHO) (WHO, 2021), the *Anopheles mosquito* and the plasmodium parasite are susceptible to any temperature change; Therefore, the constant climate of Indonesia supports the spread of this disease. In addition, Indonesia has relatively high humidity, and heavy rainfall season that benefits the *anopheles* mosquitoes by prolonging their adult form (Hasyim et al., 2018). Statistically, Indonesia has over 200,000 malaria cases; Nevertheless, South-East Asia, including Indonesia, is the second most affected region in the world by Malaria (Kemenkes RI, 2020; WHO, 2021a, 2021b). *Plasmodium falciparum*, one of five known Plasmodium parasite species that infect humans, is the most severe type of Malaria in humans and accounted for half of all malaria infections in 2018 (Zekar & Sharman, 2020; WHO, 2021a).

Antimalarial drug therapies are expected to be a solution to resolve the disease. However, some factors that originate from the drug, such as side effects and resistance, make malaria eradication difficult. For example, some antimalarial drugs such as artemisinin, chloroquine, and piperaquine have led to the emergence of parasite resistance (Pearson, 2013). The persistence of malaria infection is due to the mosquito vectors evolving to be insecticide-resistant and the lack of an effective malaria vaccine (Ellis et al., 2010; Osier et al., 2014; Hamid et al., 2017; Sumarnrote et al., 2017). The complexity of the life cycle (Gardner et al., 2002), the parasite's ability to do antigenic variation (Scherf et al., 2008), and the poor understanding of the interaction between immune cells and the parasite (Langhorne et al., 2008) are some of the factors affecting the development of drug resistance and lack of an efficient vaccine.

Bioinformatics is a modern principle that combines information science, mathematics, and biology that will help in answering biological questions such as analyzing genome sequence data, gene variation and expression, stimulating environments for whole-cell modeling, prediction and analysis of protein and gene function and structure, analysis and presentation of molecular pathways to learning about gene-disease interactions, and complex modeling of gene regulatory networks and dynamics (Bayat, 2002; Zhang & Liu, 2013). One of the bioinformatics known methods is called protein homology, which is defined by two similar or homologous proteins. Finding a protein's similarity requires sequence alignment, which utilizes tools such as FASTA, the most common tool BLAST, SSEARCH, and HMMER3. Afterward, more specific and accurate alignments can be formed using Multiple Sequence Alignment (MSA) with the established similar protein sequence (Pearson, 2013). Application of protein characterization in biological products aid in determining the safety and efficacy of drugs. The insights obtained can help alleviate the differences in development and production comparability due to variations in natural expression. It is instrumental in drug discovery, development, and manufacturing stages. The data can provide a basis for candidate screening and guidance on the cell line, strain selection, and upstream operating conditions (Kaltashov et al., 2012).

Bioinformatics-based *in-silico* studies are generally inexpensive and can help direct and confirm the numerous existing or future *in-vitro* and *in-vivo* studies on identifying antimalarial drug targets. Even though the three uncharacterized *P. falciparum* proteins are present in Uniprot, no research has been conducted on them. Thus, this is the first *in-silico* study that uses state-of-the-art bioinformatics methods to lay the groundwork for future *in-vitro* and *in-vivo* antimalarial drug target studies on the three uncharacterized *P. falciparum* proteins.

Plasmodium has a large number of stage-specific proteins that are involved in motility, maturation, metabolic adjustments, and infectivity. However, there are numerous proteins still uncharacterized until this day. Identifying the unknown proteins in plasmodium can be the beginning of solving the complications of this disease by serving as future therapeutic targets. For example, a paper studied a conserved with unknown function Plasmodium protein PF3D7_0406000 (PFD0300w) found to be localized in the cytoplasm of asexual and sexual stage Plasmodium parasites (Pandey et al., 2021). Thus, each uncharacterized Plasmodium protein has unique functions and benefits that potentially contribute to parasitic death and elimination in the future. In this research, three uncharacterized *P. falciparum* proteins were identified and further analyzed to understand their metabolic activity and potential as drug targets for malaria eradication. This study aims to use *in-silico* research to determine the three uncharacterized *P. falciparum* proteins as potential antimalarial drug targets.

Method

In this research, an *in-silico* study type was used. The methods utilized were Phyre2, InterPro, SUPERFAMILY hidden Markov models, scanprosite, Ramachandran plot, and

Yasara. The unknown protein sequences of *P. falciparum* were identified via the Microsoft Bing search engine with the search term "Uncharacterized protein *Plasmodium falciparum*." Then, three of the four unknown *P. falciparum* protein sequences (PF3D7_1468000, PF3D7_1147400, PF3D7_1351100) were downloaded in pdb format from Uniprot.

Research Procedure

The FASTA sequences of the three uncharacterized *P. falciparum* proteins were obtained from Uniprot. They were then inputted into Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) in normal modeling mode. After 12-24 hours, the resultant 3D structure and protein homology model of all the uncharacterized proteins were obtained.

To obtain information about the uncharacterized protein domains, InterPro was used by accessing the website (https://www.ebi.ac.uk/interpro/). Interpro provides protein functional analysis by categorizing it into families and predicting the environment and essential sites. Firstly, the FASTA format of the three uncharacterized proteins was inserted into the provided box on the home page of InterPro. After the search, the URL link showed the result page that detailed information about the possible family and domains.

For SUPERFAMILY, the sequence search facility allows users to submit 1000 protein sequences and obtain the corresponding domain assignment. This domain annotation search can be accessed through (http://supfam.org/sequence/search), where FASTA sequences can be inserted raw or in a file format. Once inserted, the domain annotation process began. The results displayed the annotated superfamily and the corresponding E-values. More detail on each superfamily, such as the domain, was obtained by clicking on the assigned superfamily followed by the family.

The FASTA sequence of uncharacterized *P. falciparum* protein was obtained from Uniprot. The sequences were then inputted into the FASTA box on the Scanprosite web server (https://prosite.expasy.org/scanprosite/). For the parameter, "Exclude motifs with a high probability of occurrence" was set on an "Exclude profile from the scan," and "Run the scan at high sensitivity" were set off. Afterward, the "graphic view" was selected for the output format, then click on "start to scan" for the result.

The three uncharacterized protein sequences in pdb format were loaded into the Ramachandran plot server (https://zlab.umassmed.edu/bu/rama/) with a default setting of not including the glycine residue. The residue clustering and protein secondary structures in the Ramachandran plot were analyzed, and a valid Ramachandran plot should have outliers of less than 10%.

The visualization of the uncharacterized proteins was performed using the downloaded YASARA software (http://www.yasara.org/). Each uncharacterized protein was downloaded in pdb format and loaded into YASARA for detailed analysis of the amino acid residues, protein structure and motif, and many more.

Result

Protein homology modeling in 3D structure

The resultant 3D structure and protein homology model of all the uncharacterized proteins were obtained. All sequences of the three uncharacterized proteins showed high confidence and some similar protein homology. The first, second, and third sequences displayed 99% confidence with 23% coverage, 99.1% confidence with 53% coverage, and 100% with 100% coverage, respectively (Kelley et al., 2015). Based on Phyre2, the first sequence is similar to myosin ii heavy chain in Smooth muscle and classified as contractile protein. The second sequence is identical to ubiquinol oxidase protein and is classified as an oxidoreductase. Lastly, the third sequence is similar to the 30S ribosomal protein S27E.

Phyre2 was used for protein homology modeling in 3D structures. From the result of Phyre2, the first sequence is similar to myosin ii heavy chain in Smooth muscle and classified as a contractile protein with a confidence level of 99% and coverage of 23%. The second sequence is similar to ubiquinol oxidase protein. It is classified as oxidoreductase with a confidence level of 99.1% and a range of 53%. The third sequence is similar to the 30S ribosomal protein S27E, with a confidence level and coverage of 100%.

The result will display detailed information on the possible family and domains, the annotated superfamily and family, and the corresponding E-values. More detail on each superfamily, such as the domain name, region, and function.

The InterPro result of the first sequence showed two homologous superfamilies which are WD40 repeats and WD40 repeat-like at residue 262-538. The result page also showed two unintegrated proteins, which are WD40 repeat protein and cilia and flagella-associated protein 57 at position 84-1932. The InterPro result of the second sequence showed a superfamily in the residue region 17-76. The sequence has homology with the cytochrome C oxidase subunit II, the transmembrane domain superfamily. The InterPro result of the third sequence showed three families of the sequence; however, only one family has been identified: the CXXC motif-containing zinc-binding protein at residue 1-155. The result also showed one unintegrated sequence that is MAL13P1.257-like at residue 1-155.

Three superfamilies were identified for the first sequence from the SUPERFAMILY results, as shown in Table 1. The residue regions are 262-538, 137-330, and 865-996 in order of lowest to highest E-value, and all of these identified domains are from WD40 repeat-like superfamily and WD40 family (Gough et al., 2001). For the second sequence, the annotation identified one superfamily in the residue region starting from 28 to 73, which can be observed in Table 1. It belongs to the superfamily and family of Cytochrome c oxidase subunit II-like, transmembrane region. The superfamily E-value is 0.000000409, while the family E-value is 0.0071 (Gough et al., 2001). For the third sequence, the domain annotation generated 1 identified superfamily and family, as shown in Table 1. It belongs to the superfamily and family of MAL13P1.257-like, with the associated E-value at 4.32e-53 and 0.0000000223, respectively. This domain stretches for 154 residues from position 1 to 155 (Gough et al., 2001).

Identifying post-translational protein modification using scanprosite

The result of each protein sequence from the scanprosite will display on the graphic view showing the post-translational modification in each protein sequence along with the residue regions. According to the first protein sequence in Figure 1., four distinctive profiles were identified. The residue regions are located in 73-85, 179-189, 861-1442, and 1552-1966. According to the second protein sequence in Figure 2., one distinctive profile was identified. The residue regions are located in 34-77 and represent the cytochrome c oxidase subunit II. According to the third protein sequence in Figure 3., two distinctive profiles were identified. The residue regions are 44-122 and 150-156.

Protein validation

All three uncharacterized proteins displayed good Ramachandran plots of outliers less than 10%, as shown in Figure 4A-C below. Figure 4A. Displayed 445 amino acid residues with only two unshown amino acid residues due to being either a glycine or proline residue. Highly Preferred observations are displayed in green crosses whose value is 441 amino acid residues (99.101%). Preferred observations are displayed in brown triangles with 3 (0.674%) values. Questionable observations such as outliers are displayed in red circles whose value is 1 (0.225%).

Figure 4B displayed 37 amino acid residues with only one unshown amino acid residue due to either glycine or proline residue. Highly Preferred observations are displayed in green crosses whose value is 35 amino acid residues (94.595%). Preferred observations are displayed in brown triangles whose value is one amino acid residue (2.703%). Questionable observations such as outliers are displayed in red circles whose value is one amino acid residue (2.703%).

Table 1. SUPERFAMILY results of the three protein sequences

Protein Sequence	SUPERFAMILY results					
	Sequence	Start to End residue	Superfamily	Superfamily E-value	Family	Family E-value
1	tr|Q8IKF7|Q8IKF7,PLAF7	282-538	WD40 repeat-like	0.0000000256	WD40-repeat	0.023
	tr|Q8IKF7|Q8IKF7_PLAF7	137-330		0.000897		0.032
	tr|Q8IKF7|Q8IKF7PLAF7	865-996		0.0238		0.043
2	tr|C6S3G8|C6S3G8_PLAF7	28-73		0.000000409		0.0071
3	tr|Q8IDI8|Q8IDI8_PLAF7	1-115		4.32e-53		0.0000000223

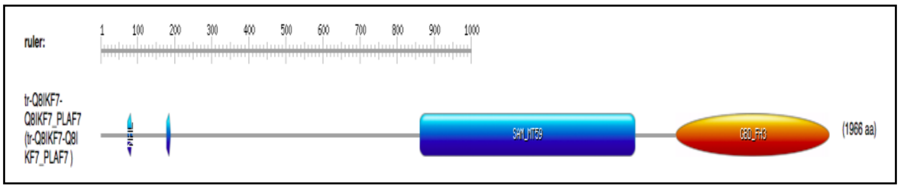

Figure 1. Illustration of the first sequence protein's domains and residue length using scanprosite. Residue proteins 73-85 and 179-189 represent the NHL repeat; residue 861-1442 represent the cytochrome c lysine N-methyltransferase 1; and residue 1552-1966 represent the formin homology 3 (FH3).

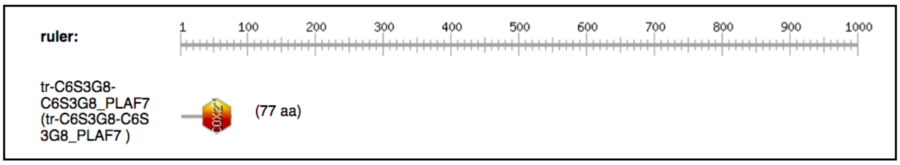

Figure 2. Illustration of the second sequence protein's protein domains and residue length using scanprosite. Residue regions 34-77 represent the cytochrome c oxidase subunit II.

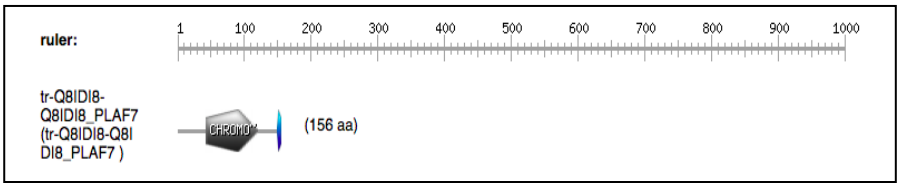

Figure 3. Illustration of the third sequence protein's protein domains and residue length using scanprosite. Residues 44-122 represent the chromo or chromo shadow domain, while residue 150-156 represents the protein prenyltransferases alpha subunit repeat.

Figure 4C displayed a total of 146 amino acid residues with only 2 unshown amino acid residues due to either a glycine or proline residue. Highly Preferred observations are displayed in green crosses whose value is 142 amino acid residues (97.260%). Preferred observations are displayed in brown triangles whose value is three amino acid residues (2.055%). Questionable observations such as outliers are displayed in red circles whose value is one amino acid residue (0.685%).

Protein visualization and analysis

The three uncharacterized proteins were analyzed for their amino acid residues, protein structure, and motif through Yasara and Uniprot. Figure 5 shows that the first uncharacterized sequence is revealed to have 1966 amino acid residues and a summed mass of 221287.791 g/mol. The protein sequence is composed of secondary structures and 34.5% helix represented in the dark blue, 10.7% of sheet presented in red, 7.4% of turn presented in green, and 47.4% in light blue. As shown in Figure 6., the second uncharacterized sequence is revealed to have 156 amino acid residues and a summed mass of 17434.923 g/mol. The protein sequence is composed of secondary structures, and 55.8% of the sheet is presented in red, 17.3% in green, and 26.9% in light blue. As shown in

Figure 7., the third uncharacterized sequence is revealed to have 77 amino acid residues and a summed mass of 8397.669 g/mol. The protein sequence is composed of secondary structures, and 66.2% of the helix is represented in dark blue, 15.6% of turn presented in green, and 18.2% in light blue.

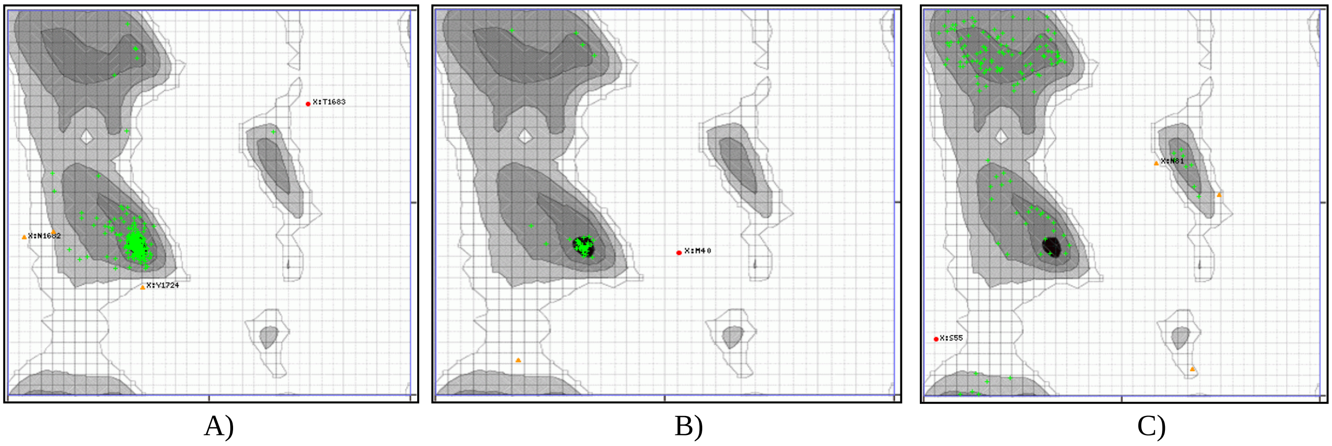

Figure 4. Ramachandran plot of the uncharacterized proteins. Highly preferred conformations are represented by black-dark grey-grey-light grey (Delta >= -2). Preferred conformations are represented by white with black grid (-2 > Delta >= -4) while questionable conformations are represented by white with grey grid (Delta < -4). Highly Preferred observations are shown as green crosses. Preferred observations are shown as brown triangles. Questionable observations are shown as red circles. A) Ramachandran plot result of the first uncharacterized protein. B) Ramachandran plot result of the second uncharacterized protein. C) Ramachandran plot result of the third uncharacterized protein. They are adapted from (Weng & Jiang, 2004). Copyright 2004 by Indonesia International Institute for Life sciences (i3L) students. They are adapted with permission.

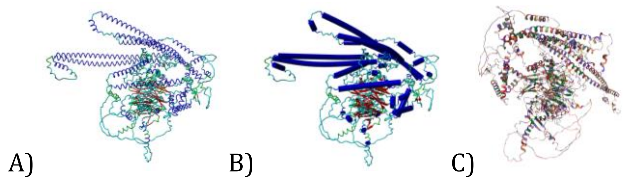

Figure 5. The model of the first uncharacterized protein. A) 3D structural model of first uncharacterized protein visualized through Yasara. B) Illustration of the protein secondary structure in the color blue visualized through Yasara. C) Illustration of the amino acids present in the protein shown in different colors visualized through UniProt.

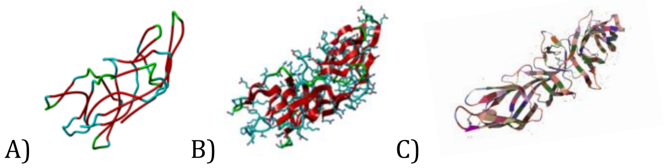

Figure 6. The model of the second uncharacterized protein. A) 3D structural model of second uncharacterized protein visualized through Yasara. B) Illustration of the protein secondary structure in the colors pink and yellow visualized through Yasara. C) Illustration of the amino acids present in the protein shown in different colors visualized through UniProt.

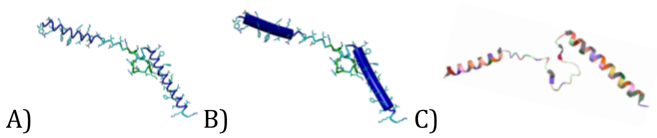

Figure 7. The model of the third uncharacterized protein. A) 3D structural model of third uncharacterized protein visualized through Yasara. B) Illustration of the protein secondary

structure in the color pink through Yasara. C) Illustration of the amino acids present in the protein shown in different colors visualized through UniProt.

Discussion

The 3D structure of the uncharacterized protein sequences was predicted using Phyre2, a state-of-the-art, publicly available, and user-friendly tool that uses advanced remote homology detection methods to predict and analyze protein function, 3D structures, and mutation. However, Phyre2 has a known limitation: modeling is impossible or unreliable if homology between a user-specific sequence and a sequence of known structures cannot be demonstrated. This reflects the broader and persistent difficulty of protein folding problems, and there is still no reliable method for predicting protein structures from sequences alone without reference to known structures. As illustrated in Figure 8, in stage 2, the sequence is compared to a HMMs database of known protein structures, and the top-scoring alignments from this search are used to construct a simple backbone-only model. This search's top-scoring alignments are utilized to build a basic backbone-only model. In stage 3, loop modeling rectifies model insertions and deletions. In stage 4, the amino acid side chains are added to create the final Phyre2 model (Kelley et al., 2015).

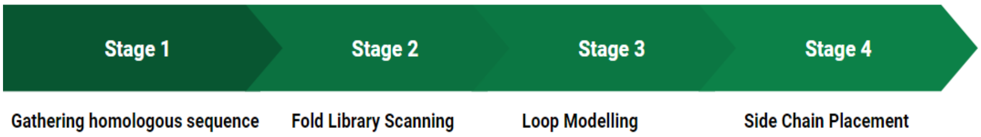

Figure 8. How Phyre2 works. It was adapted from (Kelley et al., 2015).

According to the Phyre2 results, the third sequence is the best because it has the highest confidence level and coverage (100%). The third sequence is similar to 30s ribosomal protein S27E on Archaeoglobus fulgidus. The second best is the second sequence, with a confidence level of 99.1% and coverage of 53%, and it is similar to the ubiquinol oxidase protein on Escherichia coli. The last is the first sequence, with a 99% confidence level and 23% coverage, which is similar to the myosin ii heavy chain in smooth muscle and is classified as contractile protein. The high confidence level proved that results from the Phyre2 server are accurate.

The high confidence level proved that results from the Phyre2 server are accurate, emphasizing the significance of these Phyre2 findings; some studies have shown that Phyre2 has the versatile ability to predict resultant valid proteins from various pathogens. An *in-silico* study on hepatitis C virus (HCV)-core protein domain 1 uses the Phyre2 server to predict the secondary and tertiary structures of domain one and the core of the unknown HCV sequence (Dehghani et al., 2020). Another study predicted the protein structure of neuraminidase of the influenza virus using the Phyre2 server (Thayer, 2016). Another research also utilized Phyre2 to observe the tertiary structure of Treponema pallidum (syphilis) and to study its protein function on a proteome-wide scale (Houston et al., 2018). Lastly, a study utilized Phyre2 to do a phylogenetic analysis and structural modeling of the SARS-CoV-2 spike protein (Jaimes et al., 2020).

Domain annotation is used to identify various regions of a protein sequence and subsequently discover the functions of each annotated domain. The SUPERFAMILY web server features a Hidden Markov Model profile and domain sequences obtained from the Structural Classification of Protein database (SCOP), CATH, PDB, and ECOD, which help classify deep into the level of class, fold, superfamily, and family. It also has additional analysis tools, including identifying domain representation between genomes, constructing the phylogenetic tree, analyzing the domain distribution of superfamily and families, and ontology-based annotations. The most obvious limitation of SUPERFAMILY is the completeness of the reference databases used to annotate the domains (Pandurangan et al., 2018).

On the other hand, InterPro is a similar web server that annotates domains of sequences based on reference databases. Still, it is considered complete than SUPERFAMILY because it integrates 13 protein signature databases, including SUPERFAMILY and Pfam, into a central resource, making it a comprehensive resource for annotating protein families, domains, and functional sites. InterPro combines all the resources into one entry, removing redundancies while keeping the relevant information from all of them in the results. The web server includes a wide range of information from its annotation result, including a unique name and accession number, Gene Ontology terms, and entry types that consist of family, domain, repeat, site, and homologous superfamily (Blum et al., 2021). As such, SUPERFAMILY and InterPro annotate the domains more comprehensively and accurately.

The first sequence's annotation from SUPERFAMILY and InterPro servers identified the exact WD-40 repeat superfamily domains in 262 to 538 (Gough et al., 2001). SUPERFAMILY discovered two additional WD-40 repeat superfamily domains on locations 137-330 and 865-996 with higher E-values, which may be why InterPro did not identify it. WD-40 superfamily can be found in most eukaryotes and is involved in various functions, including signal transduction, transcription regulation, cell cycle regulation, and apoptosis. It is an assembly point for protein complexes and mediators for transient interaction between other proteins (Jain & Pandey, 2018).

The same superfamily was identified from the second sequence annotated on both SUPERFAMILY and InterPro servers. It is identified as a cytochrome C subunit II-like, transmembrane region superfamily and was located in the region starting from 28 and ending in 73 (Gough et al., 2001). The annotation obtained was completed with a low E-value of 0.000000409, which is desirable. This superfamily is an oligomeric enzymatic complex that is involved in the respiratory pathway of both eukaryotes and aerobic prokaryotes, where it is involved in the transfer of electrons from cytochrome c to oxygen (Capaldi, R., Malatesta, F., & Darley-Usmar, 1983; García-Horsman et al., 1994). The location of this superfamily varies in eukaryotes and aerobic prokaryotes, with it being located in the mitochondrial inner membrane in the former and plasma membrane in the latter.

The annotated third sequence produced very different results. Although the domain was identified in the same region of 1-155, SUPERFAMILY failed to label the domain and identified it as a hypothetical protein MAL13P1.257-like that belongs explicitly to the malarial parasite *P. falciparum* (Gough et al., 2001). However, InterPro identified the superfamily as CXXC motif-containing zinc-binding protein, eukaryotes that is described with function to assist the binding of Zn2+ ion with four Cys residues from two CxxC motifs (Furukawa et al., 2018).

From the scanprosite results, the first sequence represented four distinctive profiles, which include the NHL repeat located in residues 73-85 and 179-189; the cytochrome c lysine N-methyltransferase 1 located in residue 861-1442; and Rho GTPase-binding/formin homology 3 (GBD/FH3) in residue 1552-1966. The NHL repeats are conserved homologies of amino acid sequences and are identified in a large family of growth regulators in eukaryotic and prokaryotic proteins (Jeong et al., 2009). The NHL repeats contain many protein domains, which involve protein-protein interaction (Edwards et al., 2003). Cytochrome c lysine N-methyltransferase 1 belongs to class 1 methyltransferase enzymes that are essential and present in all life forms. The methyltransferase enzymes are classified into 1, 2, and 3. Class 1 methyltransferases are enzymes consisting of seven β-sheet proteins associated with twisted strand structure. Class 2 methyltransferases are enzymes that exemplify the SET protein associated with histone methyltransferases. Class 3 methyltransferases are surface membrane enzyme proteins (Katz et al., 2003).

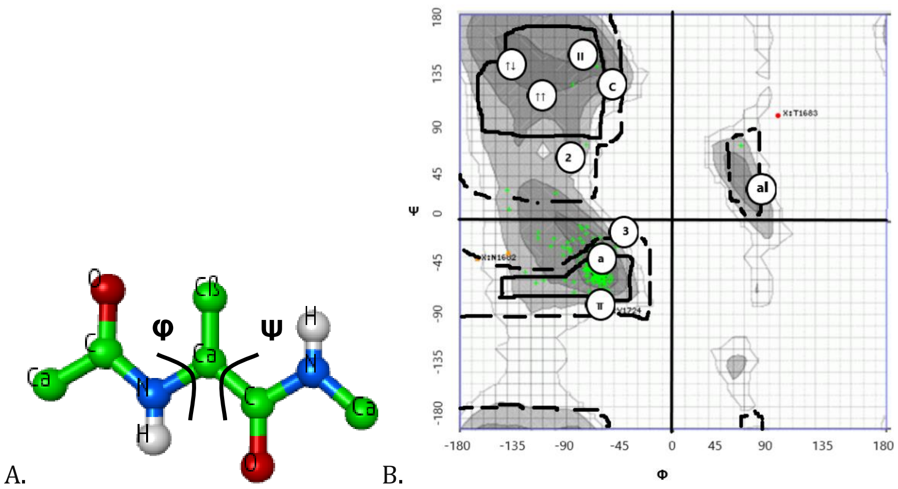

Figure 9. A dipeptide with torsion angles and a Ramachandran plot with labeled regions. **A.** A ball and stick dipeptide model with a central alanine residue showing rotations by the torsion angles ψ and φ; In which ψ generated from the Ni-Cαi-Ci-Ni+1 torsion angle and φ generated from the Ci-1-Ni-Cαi-Ci torsion angle. **B.** A classic Ramachandran plot of the first uncharacterized protein with outlines displaying the regions of classically permitted (dashed lines), core permitted (solid lines), and extreme-limit permitted (dotted lines) for the alanine dipeptide. The linear group locations are presented as α-helix (α), 3_{10}-helix (3), π-helix (π), left-handed α-helix(α_L), 2.2_7 ribbon (2), collagen (C), polyproline-II (II), parallel β-sheet (↑↑), and antiparallel β-sheet (↑↓) (Ramachandran et al., 1963; Ramachandran & Sasisekharan, 1968; Venkatachalam, 1968; Hollingsworth & Karplus, 2010).

Over 230 different methyltransferase enzymes have their functions identified in biosynthesis, signal transduction, gene silencing, and chromatin regulation (Kozbial & Mushegian, 2005; Wlodarski et al., 2011). At the same time, the cytochrome c lysine N-methyltransferases participate in the degradation of the lysine group from its substrate, cytochrome c-L-lysine, to produce S-adenosylhomocysteine and cytochrome c-N6-methyl-L-lysine (Srinivasan et al., 2013). Additionally, the S-adenosylhomocysteine can be catalyzed further through the adenosyl-homocysteinase enzyme to production for homocysteine and adenosine, which are essential to induce parasitic pathogenicity and redox stress for triggering gametocytogenesis in Malaria (Beri et al., 2017; Lima et al., 2017). Formin protein is a regulatory multi-domain protein that is conserved in vertebrates and invertebrates. Additionally, the formin protein has been identified in various regulatory processes for synthesis adherens junctions, cytokinesis, cell polarity, and actin cytoskeleton formation (Rivero et al., 2005). In vitro studies have also demonstrated the formin homology of two domains in actin polymerization and localizing to the barbed end during filament elongation (Baum et al., 2008).

In the second sequence, one distinctive profile was identified. The residue regions were located in 34-77 and represent the cytochrome c oxidase subunit II. Cytochrome c oxidase is a membrane enzyme protein found in most prokaryotic and eukaryotic cells (Castresana et al., 1994). Cytochrome c oxidase participates in the electron transport chain, particularly in removing an electron from the cytochrome c molecule through the oxidizing process and transferring two electrons into a dioxygen molecule to produce a water molecule. Additionally, the process of the oxidizing cytochrome c molecule involves protons translocated across the membranes, and the concentration difference will trigger in ATP synthesis process assisted by ATP synthase (Fontanesi et al., 2008).

In the third sequence, two distinctive profiles were identified: the chromo/chromo shadow domain located in residue 44-122 and protein prenyl-transferases alpha subunit repeat located in residue 150-156. The Chromo or chromo shadow domain is a conserved amino acid sequence constituting approximately 60 amino acids and involves alteration in the chromatin, which crucially represses gene expression (Flueck et al., 2009). The heterochromatin protein 1 (HP1) chromodomain was first identified in drosophila and participates in methyl-binding histone, which is essential for the repression of gene expression (Eissenberg, 2006). Protein prenyltransferases alpha subunits belong to prenyltransferase enzymes. The prenyltransferases consist of two subunits, α, and β. The subunits α of prenyltransferases participate in binding to isoprenyl substrates, and subunits β of prenyltransferases participate in binding to peptide substrates. Three types of subunits α are expressed in prenyltransferases: farnesyltransferase and geranylgeranyltransferase I, which bind to the same cysteine substrates, and geranylgeranyl II, which binds to a non-cysteine substrate (Maurer-Stroh et al., 2003).

As a part of protein structure validation, utilizing the Ramachandran plot provides information on the protein structure, the amino acid conformation, and the direct contacts between atoms. Ramachandran plot is defined as an easy-to-understand visualization plot showing the psi (ψ)-phi (φ) torsion angle pair of the polypeptide chain in a protein structure (Sheik et al., 2002), as illustrated in Figure 9A. Thus, the Ramachandran plot depicts the distribution of torsion angles in a protein structure and its secondary structures, which are comprehensibly clustered into distinct areas (Sheik et al., 2002), as portrayed in Figure 9B when comparing Figure 9B with Figure 4, the first uncharacterized protein roughly has an overall α-helix structure with some π-helix, 3_{10}-helix, polyproline-II, and collagen, which matches its visualization of having mostly helix structures in Figure 5. When comparing Figure 9B with Figure 4B, the second uncharacterized protein roughly has an overall concentrated α-helix structure with an antiparallel β-sheet along with some collagen and polyproline-II, which matches its visualization of being essentially sheet structured to an extent in Figure 6. When comparing Figure 9B with Figure 4C, the third uncharacterized protein roughly has the following structures: α-helix, 3_{10}-helix, π-helix, left-handed α-helix, along with the more concentrated collagen (C), polyproline-II (II), parallel β-sheet (↑↑), antiparallel β-sheet structures; Hence, making it an overall β-sheet structure, which is slightly different with its visualization of mostly composting of helix structures in Figure 7. As the Ramachandran plot clearly outlines the inside of the protein while YASARA functions to visualize the outside of the protein, YASARA cannot correctly visualize the insides of a protein structure due to the visuals of overlapping residues; Hence, it is understandable that both results may differ.

Unfortunately, little information was obtained on the first uncharacterized protein to elucidate whether or not its function is critical and whether or not its inhibition promotes the death of the *P. falciparum* parasite. However, the second uncharacterized protein may function as a transmembrane farnesylation in the mitochondria that accommodates a subunit of intracellular proteins (Hall et al., 2002). Notably, some studies stated inhibiting the protein farnesylation of *P. falciparum* is essential to the parasite's survival (Eastman et al., 2005) and to further understand its mitochondria function (Ha & Lee, 2012). Following the third uncharacterized protein being a protein prenyltransferase, this protein may be required to develop and replicate harmful eukaryotic microorganisms such as the malaria protozoan *P. falciparum*. This protein also plays an essential role in eukaryotic signal transduction (Hast & Beese, 2011), specifically in the post-translational modification of proteins involved in signal transduction pathways, cell cycling, and DNA replication regulation. In essence, the potential drugs targeting this protein can then inhibit the differentiation and division of the malarial parasite (Chakrabarti et al., 2002). Hence, the second and third uncharacterized proteins may be potential drug targets affecting the survival of the *P. falciparum* parasite. However, further *in-vitro* and *in-vivo* studies must be conducted to clarify their essential role in the *P. falciparum* parasite.

Conclusion

In conclusion, the second and third uncharacterized proteins may be essential drug target candidates which directly affect *P. falciparum* survival. However, not much information is known about the first uncharacterized protein.

Acknowledgment

The authors would like to thank The Research and Community Service Department (LPPM) of Indonesia International Institute for Life Sciences (i3L) for supporting this research.

Declaration statement

The authors reported no potential conflict of interest.

References

Baum, J., Tonkin, C. J., Paul, A. S., Rug, M., Smith, B. J., Gould, S. B., & Cowman, A. F. (2008). A malaria parasite formin regulates actin polymerization and localizes to the parasite-erythrocyte moving junction during invasion. *Cell Host & Microbe*, *3*(3), 188–198. https://doi.org/10.1016/j.chom.2008.02.006

Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ: British Medical Journal*, *324*(7344), 1018.

Beri, D., Balan, B., Chaubey, S., Subramaniam, S., Surendra, B., & Tatu, U. (2017). A disrupted transsulphuration pathway results in accumulation of redox metabolites and induction of gametocytogenesis in malaria. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep40213

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., & Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*(D1), D344–D354. https://doi.org/10.1093/nar/gkaa977

Capaldi, R., Malatesta, F., & Darley-Usmar, V. (1983). Structure of cytochrome c oxidase. Biochimica Et Biophysica Acta (BBA). *Reviews On Bioenergetics*, *726*(2), 135–148. https://doi.org/10.1016/0304-4173(83)90003-4.

Castresana, J., Lübben, M., Saraste, M., & Higgins, D. G. (1994). Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. *The EMBO Journal*, *13*(11), 2516–2525. https://doi.org/10.1002%2Fj.1460-2075.1994.tb06541.x

Chakrabarti, D., Da Silva, T., Barger, J., Paquette, S., Patel, H., Patterson, S., & Allen, C. M. (2002). Protein Farnesyltransferase and Protein Prenylation in Plasmodium falciparum. *Journal of Biological Chemistry*, *277*(44), 42066–42073. https://doi.org/10.1074/jbc.M202860200.

Dehghani, B., Hashempour, T., Hasanshahi, Z., & Moayedi, J. (2020). Bioinformatics analysis of domain 1 of HCV-core protein: Iran. *International Journal of Peptide Research and Therapeutics*, *26*(1), 303–320. https://doi.org/10.1007%2Fs10989-019-09838-y

Eastman, R. T., White, J., Hucke, O., Bauer, K., Yokoyama, K., Nallan, L., & Van Voorhis, W. C. (2005). Resistance to a protein farnesyltransferase inhibitor in Plasmodium falciparum. *Journal of Biological Chemistry*, *280*(14), 13554–13559. https://doi.org/10.1074/jbc.M413556200

Edwards, T., Wilkinson, B., Wharton, R., & Aggarwal, A. (2003). Model of the Brain Tumor–Pumilio translation repressor complex. *Genes & Development*, *17*(20), 2508–2513. https://doi.org/10.1101/gad.1119403

Eissenberg, J. (2006). Molecular biology of the chromo domain: an ancient chromatin module comes of age. *Gene*, *275*(1), 19–29.

Ellis, R. D., Sagara, I., Doumbo, O., & Wu, Y. (2010). Blood stage vaccines for Plasmodium falciparum: current status and the way forward. *Human Vaccines*, *6*(8), 627–634. https://doi.org/10.4161%2Fhv.6.8.11446

Flueck, C., Bartfai, R., Volz, J., Niederwieser, I., Salcedo-Amaya, A. M., Alako, B. T., & Voss, T. S. (2009). Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathogens*, *5*(9), e1000569. https://doi.org/10.1371/journal.ppat.1000569

Fontanesi, F., Soto, I. C., & Barrientos, A. (2008). Cytochrome c oxidase biogenesis: new levels of regulation. *IUBMB Life*, *60*(9), 557–568. https://doi.org/10.1002/iub.86

Furukawa, Y., Lim, C., Tosha, T., Yoshida, K., Hagai, T., Akiyama, S., & Shiro, Y. (2018). Identification of a novel zinc-binding protein, C1orf123, as an interactor with a heavy metal-associated domain. *Plos One*, *13*(9), e0204355. https://doi.org/10.1371/journal.pone.0204355

García-Horsman, J., Barquera, B., Rumbley, J., Ma, J., & Gennis, R. (1994). The superfamily of heme-copper respiratory oxidases. *Journal Of Bacteriology*, *176*(18), 5587–5600. https://doi.org/10.1128/jb.176.18.5587-5600.1994

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., & Barrell, B. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, *419*(6906), 498–511. https://doi.org/10.1038/nature01097

Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal Of Molecular Biology*, *313*(4), 903–919.

Ha, Y. R., & Lee, S. J. (2012). Effect of farnesyltransferase inhibitor on the function of mitochondria of Plasmodium falciparum. *Malaria Journal*, *11*(1), 1–2. https://doi.org/10.1186/1475-2875-11-S1-P62

Hall, N., Pain, A., Berriman, M., Churcher, C., Harris, B., Harris, D., Mungall, K., Bowman, S., Atkin, R., Baker, S., Barron, A., Brooks, K., Buckee, C. O., Burrows, C., Cherevach, I., Chillingworth, C., Chillingworth, T., Christodoulou, Z., Clark, L., … Barrell, B. G. (2002). Sequence of Plasmodium falciparum chromosomes 1, 3–9 and 13. *Nature*, *419*(6906), 527–531. https://doi.org/10.1038/nature01095

Hamid, P. H., Prastowo, J., Ghiffari, A., Taubert, A., & Hermosilla, C. (2017). Aedes aegypti resistance development to commonly used insecticides in Jakarta, Indonesia. *PLoS One*, *12*(12), e0189680. https://doi.org/10.1371/journal.pone.0189680

Hast, M. A., & Beese, L. S. (2011). Structural Biochemistry of CaaX Protein Prenyltransferases. *In The Enzymes*, *29*, 235–257.

Hasyim, H., Nursafingi, A., Haque, U., Montag, D., Groneberg, D. A., Dhimal, M., Kuch, U., & Müller, R. (2018). Spatial modelling of malaria cases associated with environmental factors in South Sumatra, Indonesia. *Malaria Journal*, *17*(1), 1–15. https://doi.org/10.1186/s12936-018-2230-8

Hollingsworth, S. A., & Karplus, P. A. (2010). A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts*, *1*(3–4), 271–283. https://doi.org/10.1515/BMC.2010.022

Houston, S., Lithgow, K. V., Osbak, K. K., Kenyon, C. R., & Cameron, C. E. (2018). Functional insights from proteome-wide structural modeling of Treponema pallidum subspecies pallidum, the causative agent of syphilis. *BMC Structural Biology*, *18*(1), 1–18. https://doi.org/10.1186/s12900-018-0086-3

Jaimes, J. A., André, N. M., Chappie, J. S., Millet, J. K., & Whittaker, G. R. (2020). Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically sensitive activation loop. *Journal of Molecular Biology*, *432*(10), 3309–3325. https://doi.org/10.1016/j.jmb.2020.04.009

Jain, B., & Pandey, S. (2018). WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. *The Protein Journal*, *37*(5), 391–406. https://doi.org/10.1007/s10930-018-9785-7

Jeong, J. K., Kwon, O., Lee, Y. M., Oh, D. B., Lee, J. M., Kim, S., Kim, E. H., Le, T. N., Rhee, D. K., & Kang, H. A. (2009). Characterization of the Streptococcus pneumoniae BgaC protein as a novel surface beta-galactosidase with specific hydrolysis activity for the Galbeta1-3GlcNAc moiety of oligosaccharides. *Journal of Bacteriology*, *191*(9), 3011–3023. https://doi.org/10.1128/JB.01601-08

Kaltashov, I., Bobst, C., Abzalimov, R., Wang, G., Baykal, B., & Wang, S. (2012). Advances and challenges in analytical characterization of biotechnology products: Mass spectrometry-based approaches to study properties and behavior of protein therapeutics. *Biotechnology Advances*, *30*(1), 210–222. https://doi.org/10.1016/j.biotechadv.2011.05.006

Katz, J., Dlakić, M., & Clarke, S. (2003). Automated Identification of Putative Methyltransferases from Genomic Open Reading Frames. *Molecular & Cellular Proteomics*, *2*(8), 525–540. https://doi.org/10.1074/mcp.m300037-mcp200

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Kemenkes RI. (2020). *Informasi Malaria Resmi Indonesia*. https://www.malaria.id/en. (Accessed on 13th August 2022)

Kozbial, P., & Mushegian, A. (2005). Natural history of S-adenosylmethionine-binding proteins. *BMC Structural Biology*, *5*(1). https://doi.org/10.1186/1472-6807-5-19

Langhorne, J., Ndungu, F. M., Sponaas, A. M., & Marsh, K. (2008). Immunity to malaria: more questions than answers. *Nature Immunology*, *9*(7), 725–732. https://doi.org/10.1038/ni.f.205

Lima, M., Sacramento, L., Quirino, G., Ferreira, M., Benevides, L., & Santana, A. (2017). Leishmania infantum Parasites Subvert the Host Inflammatory Response through the Adenosine A2A Receptor to Promote the Establishment of Infection. *Frontiers In Immunology*, *8*. https://doi.org/10.3389/fimmu.2017.00815

Maurer-Stroh, S., Washietl, S., & Eisenhaber, F. (2003). Protein prenyltransferases. *Genome Biology*, *4*(4). https://doi.org/10.1186/gb-2003-4-4-212

Osier, F. H., Mackinnon, M. J., Crosnier, C., Fegan, G., Kamuyu, G., Wanaguru, M., & Marsh, K. (2014). New antigens for a multicomponent blood-stage malaria vaccine. *Science Translational Medicine*, *30*(6), 247ra102. https://doi.org/10.1126/scitranslmed.3008705

Pandey, I., Quadiri, A., Wadi, I., Pillai, C. R., Singh, A. P., & Das, A. (2021). Conserved Plasmodium Protein (PF3D7_0406000) of Unknown Function: In-silico Analysis and Cellular Localization. *Infection, Genetics and Evolution*, *92*(104848). https://doi.org/10.1016/j.meegid.2021.104848

Pandurangan, A., Stahlhacke, J., Oates, M., Smithers, B., & Gough, J. (2018). The SUPERFAMILY 2.0 database: A significant proteome update and a new webserver. *Nucleic Acids Research*, *47*(D1), D490–D494.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *42*(1), 3–1. https://doi.org/10.1002/0471250953.bi0301s42

Ramachandran, G. N. T., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol*, *7*, 95–99. https://doi.org/10.1016/s0022-2836(63)80023-6

Ramachandran, G. N. T., & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, *23*, 283–437. https://doi.org/10.1016/s0065-3233(08)60402-7

Rivero, F., Muramoto, T., Meyer, A., Urushihara, H., Uyeda, T., & Kitayama, C. (2005). A comparative sequence analysis reveals a common GBD/FH3-FH1-FH2-DAD architecture in formins from Dictyostelium, fungi and metazoa. *BMC Genomics*, *6*(1). https://doi.org/10.1186/1471-2164-6-28

Scherf, A., Lopez-Rubio, J. J., & Riviere, L. (2008). Antigenic variation in Plasmodium falciparum. *Annu. Rev. Microbiol*, *62*, 445–470. https://doi.org/10.1146/annurev.micro.61.080706.093134

Sheik, S. S., Sundararajan, P., Hussain, A. S. Z., & Sekar, K. (2002). Ramachandran plot on the web. *Bioinformatics*, *18*(11), 1548–1549. https://doi.org/10.1093/bioinformatics/18.11.1548

Srinivasan, S., Spear, J., Chandran, K., Joseph, J., Kalyanaraman, B., & Avadhani, N. G. (2013). Oxidative stress induced mitochondrial protein kinase A mediates cytochrome c oxidase dysfunction. *PloS One*, *8*(10), e77129. https://doi.org/10.1371/journal.pone.0077129

Sumarnrote, A., Overgaard, H. J., Marasri, N., Fustec, B., Thanispong, K., Chareonviriyaphap, T., & Corbel, V. (2017). Status of insecticide resistance in Anopheles mosquitoes in Ubon Ratchathani province, Northeastern Thailand. *Malaria Journal*, *16*(1), 1–13. https://doi.org/10.1186/s12936-017-1948-z

Thayer, K. M. (2016). Structure prediction and analysis of neuraminidase sequence variants. *Biochemistry and Molecular Biology Education*, *44*(4), 361–376. https://doi.org/10.1002/bmb.20963

Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers: Original Research on Biomolecules*, *6*(10), 1425–1436.

Weng, A., & Jiang, C. (2004). *Ramachandran Plot Server*. https://zlab.umassmed.edu/bu/rama/(Accessed on 10th November 2022)

WHO. (2021a). *Fact sheet about Malaria*. https://www.who.int/news-room/fact-sheets/detail/malaria. (Accessed on 25th July 2022)

WHO. (2021b). *Malaria*. https://www.who.int/indonesia/health-topics/malaria. (Accessed on 25th July 2022)

Wlodarski, T., Kutner, J., Towpik, J., Knizewski, L., Rychlewski, L., Kudlicki, A., & Ginalski, K. (2011). Comprehensive structural and substrate specificity classification of the Saccharomyces cerevisiae methyltransferome. *PloS One*, *6*(8), e23168. https://doi.org/10.1371/journal.pone.0023168

Zekar, L., & Sharman, T. (2020). *Plasmodium Falciparum Malaria*.

Zhang, S., & Liu, S. (2013). Bioinformatics. *Brenner's Encyclopedia of Genetics*, 338–340. https://doi.org/10.1016/b978-0-12-374984-0.00155-8