Structures prediction of *Plasmodium Falciparum* Signal Peptide Peptidase (PfSPP) and identification of binding Site

Gani Priambodo¹, Agus Kartono¹, Irmanida Batubara²³ and Setyanto Tri Wahyudi¹³*

¹Department of Physics, IPB University, Bogor 16680, Indonesia
²Department of Chemistry, IPB University, Bogor 16680, Indonesia
³Tropical Biopharmaca Research Center, IPB University, Indonesia

*E-mail: stwahyudi@apps.ipb.ac.id

Abstract. Malaria is a tropical parasite disease caused by mosquitoes. *Plasmodium falciparum* is the species that causes the most human deaths compared to other species. *Plasmodium* which infects red blood cells will cause pathology and clinical manifestations in malaria sufferers. *Plasmodium falciparum* Signal Peptide Peptidase (PfSPP) is an important enzyme to infect red blood. I-TASSER (Iterative Threading ASSEmbly Refinement) has been used to predict the 3D structure of the PfSPP enzyme by modeling using proteins in databases. Based on data analysis such as Ramachandran plot, G-Factor, RMSD, Radius of gyration, and NAMD energy, model 5 is the best model of all I-TASSER structure prediction models. AutoLigand was used to predict the binding sites on this model and obtained 5 binding sites points with lowest free energy on each fill point. Binding sites 3, 4 and 5 have the largest volumes 345 Å³, 372 Å³, and 395 Å³ respectively so that they have potential to bind with both ligand mefloquine and primaquine with volume 330 Å³ and 333 Å³ respectively. Three binding sites have the potential to inhibit PfSPP so it cannot function properly and stop the invasion of the parasite’s merozoite to other erythrocytes.

1. Introduction
In the last twenty years, efforts to treat malaria have contributed to a reduction in malaria cases and deaths worldwide, but there has been no significant increase in the past two years. The global tally deaths caused by malaria in 2016 reached 445,000 deaths out of 216 million cases increased by 5 million cases since the epidemic in 2015 [1].

Malaria is a type of tropical parasitic disease. Various types of mosquito vectors carry *Plasmodium* that causes malaria. More than 120 species of Plasmodium can infect mammals, birds, and reptiles, but only six species are known to infect humans [2]. One type of malaria is *Plasmodium falciparum*. *P. falciparum* produces high-level blood-phase parasites with critical organ damage at all ages and can cause severe anemia in children in Africa, the largest area of total malaria deaths.

The intra-erythrocyte phase of malaria parasites can cause various pathological effects and clinical manifestations in the host. Therefore, the merozoites raid on the red blood cells released by the liver or infected erythrocytes is the host route of the disease. *Plasmodium* infects red blood cells by using its components to be able to enter and take over cells. At first, the parasite will disrupt the vacuole home called the parasitophorous vacuole and then damage the membrane of red blood cells. As particle protein in *Plasmodium Falciparum* membrane cells, known as signal peptide peptidase (PfSPP) is one of the crucial enzymes for merozoites raid [3]. Therefore, inhibition of this protease has a vital role as a target.
in resistance to parasite life cycles. This research aims to construct a 3D structure of PfSPP and predict the binding site for a drug target.

2. Materials and Methods

The instrument used in this research was Personal Computer with specification RAM 32 GB, Processor Intel Core i7-8700 3.2 GHz, NVIDIA GEFORCE GTX 1080 Ti, and Operating System Linux Ubuntu 18.04. Software for molecular dynamic (MD) simulation and docking using AMBER 18 and Vina, respectively. Docking analysis using LigPlot 2.1 and Autodock Tools 1.5.6. MD trajectory analysis using cpptraj, NAMD energy, and VMD 1.9.3.

2.1. 3D Structure Construction

Amino acid (AA) sequence of PfSPP was provided by UniProt (https://www.uniprot.org/), with entry id Q81KQ9, from Plasmodium falciparum (isolate 3D7). By using BlastP analysis, there is 25% identity with molecule 2XQB.PDB. Blastp was done by comparing the model data sequence with a sequence that has a 3D structure [4]. Since the similarity was low, the construction of predicted 3D structure was done by I-Tasser (Iterative Threading ASSEmby Refinement) web server.

I-Tasser modeled the structure by using threading modeling. I-Tasser web server is used based on the advantages of better structure results compared to other web servers that have a residual defect. Limit range of protein sequence target that can be received by I-Tasser around 10-1500 residues. I-Tasser modeling process depends on protein size; it requires a maximum of 48 hours (5-10 hours for about 200 residues) [5]. Results from I-Tasser was evaluate using PDBSum and Molprobity [6, 7].

2.2. Energy Minimization and MD Simulation of 3D Model

Next stage is the energy minimization of the resulted model from I-Tasser. PDB file of the model should be prepared by pdb4amber, a module in AMBER18 [8]. Then, titration was made by H++ web server with pH 7.0 to predict the pKa value and configuration of some residue amino acid [9, 10].

After completion of the titration, the PDB-file was prepared by using tleap, to adding hydrogen and solvating with water, and neutralizing the system. Box system was used with dimension 18x18 18 Å from the edge of protein structure. The amount of water used to immerse the protein are 31417 molecules, and 11 ions Cl need to neutralize the system. Since the system ready, the next stage is energy minimization. Minimization consists of 7 stages. Step 1 to 6 was done in 10,000 steps, The 2500 first step done by steepest descent algorithm, and rest with conjugate gradient. Restraining was done gradually, 1000, 500, 200, 100, 50 and 25 kcal/mol, start from 1st to 6th stages. In the 7th stage, there is no restraint and was done in 40,000 steps. The 7th stage was the last stage before heating and equilibration in MD simulation.

Heating dynamics was done in 6 stages to reach 310 K. Start from 0 K; each stage has 30,000 steps and a 50K increase. Periodic Boundary Condition (PBC) starting ON in the heating process. Along the heating dynamics process, the box system was restraint with 10 kcal/mol.Å² dummy spring constant. Equilibration was done after the heating process. The first equilibration was a canonical ensemble, and the five other steps were an isothermal-isobaric ensemble. In this process, we remove the restraint gradually add gamma_In gradually until we reach gamma_In=5. After the equilibration completion, the MD simulation begins. The system in MD simulation was an isothermal-isobaric ensemble and without restraint. MD simulation was done at 310K for 50 ns.

2.3. Binding Site Prediction

After MD simulation, PfSPP prediction model then prepared by Autodock to merge nonpolar hydrogen, add charges, assign appropriate atom types to include only C, HD, and OA. Then Grid options were used to creating a grid box with a spacing of 1 Å to encompasses the entire receptor and save as gpff to be launched in AutoGrid. Moreover, the AutoGrid was used to process the grid box maps for Autoligand to predict the binding site model [11].
Before prediction, AutoLigand commands must be loaded in browse command > select a package-AutodockTools > module-Autoligand command > load > select module. Later, grid box maps results will be used as a domain for scanning protein-ligand high affinity in each binding site and identify the best volume, shape, and type of atom for each binding sites [12]. Additional software USCF-Chimera was used to predicting hydrogen bonds residue as donor or acceptor, charged, and hydrophobic amino acids around each binding site within 3.5 Å.

3. Results and Discussion

3.1. 3D Structure Construction

I-Tasser web server gives five models (Figure. 1). Each model has an internal score from I-Tasser. The confidence of each model is quantitatively measured by C-score that is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-Score has a range of -5 to 2, where C-score of higher value signifies a model with high confidence and vice-versa. Based on the data that has been obtained (Figure.1), model 1 is a model with high confidence. This score was obtained when the model just came out from I-Tasser.

Figure 1. 3D model results of PfSPP from I-Tasser; a. Model 1 (C-Score = -0.75); b. Model 2 (C-Score = -1.48); c. Model 3 (C-Score = -3.12); d. Model 4 (C-Score = -2.99); d. Model 5 (C-Score = -2.04).

Beside internal assessment from I-Tasser using C-Score, we also evaluate the model using PDBSum and Molprobity. Evaluation from PDBSum involves Ramachandran plot and G-Factor while evaluation from Molprobity involves clash score, rotamer, and Ramachandran.
Table 1. Ramachandran Plot Statistics of all models from I-TASSER results.

| Non-glycine and non-proline residues | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|--------------------------------------|---------|---------|---------|---------|---------|
| Residues in most favored regions     | 70.4%   | 68.8%   | 74.3%   | 73.3%   | 81.0%   |
| Residues in additional allowed regions | 23.3% | 24.9% | 16.9% | 17.7% | 15.3% |
| Residues in generously allowed regions | 3.2% | 3.2% | 5.0% | 6.3% | 2.1% |
| Residues in disallowed regions      | 3.2%    | 3.2%    | 3.7%    | 2.6%    | 1.6%    |

All of the Molprobity evaluation was combine to be the Molprobity Score. Ramachandran plot is used to determine the structure quality, which modeled with the amino acid torsion angle distribution. Model 5 (Table 1) has good structural quality because the amount of non-glycine and non-proline residues in disallowed regions is low (1.6%) compared to model 1 and other models.

Table 2. 3D Structure Evaluation of all models from I-TASSER results

| Models  | G-Factor* | Molprobity Score^ |
|---------|-----------|-------------------|
| Model 1 | -0.69     | 3.35              |
| Model 2 | -0.64     | 3.13              |
| Model 3 | -1.32     | 4.12              |
| Model 4 | -0.71     | 3.32              |
| Model 5 | -0.54     | 3.14              |

*Values below -0.5 – unusual
Values below -1.0 – highly unusual
^ MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized

Low G-Factor is closely related to low conformational possibilities. The more residues in the disallowed regions will decrease the G-Factor value. Model 5 has a value of -0.54 (Table 2), so it has the best possible conformation among other models even though it is included an unusual 3D structure range. The model evaluation score using Molprobity was normalized to be on the same scale as the X-Ray Diffraction resolution. Therefore, the lowest resolution is the best resolution. Model 2 and model 5 have the best resolution with 3.13 and 3.15 (Table 2).

3.2. 3D Structure After Minimization

Table 3. Ramachandran Plot Statistics for all models after minimization

| Non-glycine and non-proline residues | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|--------------------------------------|---------|---------|---------|---------|---------|
| Residues in most favored regions     | 74.1%   | 75.1%   | 70.6%   | 77.8%   | 79.4%   |
| Residues in additional allowed regions | 18.3% | 19.3% | 23.3% | 16.4% | 17.2% |
| Residues in generously allowed regions | 3.7% | 2.6% | 2.4% | 2.6% | 2.1% |
| Residues in disallowed regions      | 4.0%    | 2.9%    | 3.7%    | 3.2%    | 1.3%    |
All model produced by I-Tasser undergoes minimization energy. Energy minimum reach -433850, -474000, -414170, -407450 kcal/mol for model 1, 2, 3, 4, and 5 respectively. Model 1 has the lowest minimum energy, but according to Table 1, the scoring number for Non-glycine and non-proline residues were lower than model 5. Model 5 has good structural quality because the amount of non-glycine and non-proline residues in disallowed regions is low (1.3%) and high amount residues in most favored regions (79.4%) (Table 3) compared to model 1 and other models.

One of the score to depict good 3D structure is G-Factor. The range of G-factor from positive to negative number. More positive indicated the better structure, so getting closer to zero also indicated the better structure. There is various G-factor number from -0.54 to -0.37 (Table 4). The G-Factor value in model 5 is also the highest compared to other models, and therefore, model 5 is the best model. There is a significant decrease at the Molprobity score for each model after minimization. Model 2 and 4 have the lowest Molprobity score, followed by model 3 and 5 (Table 4).

Table 4. 3D Structure Evaluation after minimization

| Models | G-Factor* | Molprobity^ |
|--------|-----------|-------------|
| Model 1 | -0.48 | 2.35 |
| Model 2 | -0.42 | 2.08 |
| Model 3 | -0.54 | 2.24 |
| Model 4 | -0.44 | 2.16 |
| Model 5 | -0.37 | 2.29 |

*Values below -0.5 – unusual
*Values below -1.0 – highly unusual
^ MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized to be on the same scale as X-ray resolution.

3.3. Molecular Dynamics Simulation

3.3.1. Root Mean Square Deviation (RMSD) analysis. Displacement of the 3D structure starts from 0 ns to 50 ns was depict by Root Mean Square Deviation (RMSD) showed in Figure 2. Graph of RMSD has been refined with a moving average of every 10 data (Figure 2). Changes in model 1 ranged from 2.43 Å to 9.79 Å with average 8.49 Å, for model 2 ranged from 2.46 Å to 7.51 Å with average 6.84 Å, for model 3 ranging from 3.04 Å to 6.77 Å with average 6.02 Å, for model 4 ranged from 2.58 Å to 8.59 Å with average 7.45 Å and for model 5 ranged from 2.46 Å to 6.49 Å with average 5.83 Å. Model 1, 2, and 4 have a bit broaden RMSD value; it was indicated to have high dynamic movement. While model 5, the conformation changes is more stable than other models.

Figure 2. RMSD values of all models for 50 ns simulation (yellow = model 1; red = model 2; green = model 3; purple = model 4; blue = model 5).
3.3.2. The radius of gyration analysis. There are significant changes in the radius of gyration for each model, but model 3 and 5 have the most constant value compared to other models (Figure 3). The average value of model 1 is 24.18 Å, for model 2 is 22.13 Å, for model 3 is 22.36 Å, for model 4 is 22.37 Å, and for model 5 is 22.83 Å. Model 3 and 5 have a high average value compared than the other models except model 1, but model 5 does not have a significant change in radius gyration. This indicates that the model 5 has stable in a whole globular shape.

![Figure 3](image1.png)

**Figure 3.** Radius of gyration of all models for 50 ns simulation (yellow = model 1; red = model 2; green = model 3; purple = model 4; blue = model 5).

3.3.3. NAMD energy analysis. Conformation energy is the sum of bond energy, bond angle, and dihedral angle. Throughout the simulation, the energy required tends to decrease and then stabilizes. The data indicated that proteins tend to be stable (Figure 4). Figure 11 shows total energy for all models. Based on Figure 3, an average energy value was 1597.58, 1162.61, 3 is 905.97, 1108.72, and 1057.77 for model 1, 2, 3, 4, and 5, respectively. The highest average total energy is in model 1, and the lowest is models 3 followed by model 5. The lowest of the conformational energy, the structure more natural.

Based on conformational energy, models 3 and 5 have the best and constant analysis values compared to other models. But the RMSD, Radius of gyration, and Ramachandran plot of model 5 showed that model 5 is better than model 3. Therefore model 5 was chosen as a model for PiSPP and will be used for the next simulation and analysis.

![Figure 4](image2.png)

**Figure 4.** Conformational energy of all models for 50 ns simulation (yellow = model 1; red = model 2; green = model 3; purple = model 4; blue = model 5).
We try to summarize the evaluation of model 5, start from “just came out” structure, minimize structure and structure after 50 ns MD simulation (Table 5). There was an increase in the number of residues in the most favored region, from 81.0% to 83.3% after MD simulation. The residue in additional allowed regions increased from 15.3% to 17.2% after minimization but then decreased to 14.8% after 50 ns molecular dynamics simulation. The interesting one is the residues in disallowed regions decreased from every evaluation, 1.6% in “just came out” structure, decreased to be 1.3% after minimization and finally 1.1% after 50 ns MD simulation.

Table 5. Evaluation of Model 5 PfSPP

| Ramachandran plot | I-Tasser Results | After minimization | After 50 ns MD simulation |
|-------------------|------------------|-------------------|--------------------------|
| Residues in most favored regions | 81.0% | 79.4% | 83.3% |
| Residues in additional allowed regions | 15.3% | 17.2% | 14.8% |
| Residues in generously allowed regions | 2.1% | 2.1% | 0.8% |
| Residues in disallowed regions | 1.6% | 1.3% | 1.1% |
| G-Factor | -0.54 | -0.37 | -0.99 |
| Molprobity score | 3.14 | 2.29 | 2.06 |

G-Factor and Molprobity Score also has improved after MD simulation. Model 5 has -0.99, getting closer to zero, of G-factor and 2.06 Molprobity score after 50 ns MD simulation. This fact was showed that molecular dynamics simulation has a significant effect on structure improvement or refinement. Molecular dynamics simulation was to mimic the natural condition in the reaction tube. So, the structure that undergoes molecular dynamics simulation has minimum energy and natural structure.
Figure 5. (a) Prediction of free energy for 5 groups of different fill point with 5 lowest free energies (red circle). This prediction was done on the surface of model 5 by AutoLigand (Blue=100 fill points; Purple=110 fill points; Green=120 fill points; Orange=130 fill points; Yellow=140 fill points). (b) 5 lowest free energy binding sites identified on the surface of model 5 by AutoLigand

The 3rd, 4th, and 5th binding sites have the large volume of other 5 binding sites 345 Å³, 372 Å³ dan 395 Å³ (Fig 8) so that it requires higher free energy than the other binding sites. Calculation of the volume of 5 commercial ligands Chloroquine, Doxycycline, Halofantrine, Mefloquine and Primaquine using 3v (Voss Volume Voxelator) website [13]. Mefloquine and primaquine have the smallest volume among the other 5 ligands, 330 Å³, and 333 Å³. Since the volume of mefloquine and primaquine is smaller than the 3rd, 4th, and 5th binding sites, therefore these two ligands have the potential to bind with these binding sites. Further research in molecular dynamics simulations on the docking of both ligands and the three binding sites is a need as a benchmark for the stability of PfSPP proteins and commercial ligands.
Figure 6. The three binding sites predicted by AutoLigand in the structure of model 5 as potential drug sites for commercial ligands. a. Green (-0.191 kcal/mol Å³), b. Yellow (-0.185 kcal/mol Å³), c. Red (-184 kcal/mol Å³)

4. Conclusion
The prediction of the PfSPP 3D structure model of protein sequences by the I-Tasser automatic web server produces 5 prediction models that approach the I-Tasser database template. The results of the analysis of C-Score, G-factor, Molprobility score, RMSD, the radius of gyration, and NAMD energy in all models in the early stages, minimization and molecular dynamics steps for 50 ns and 300K indicated that model 5 is the best structure.

The results of binding site analysis on fill points 100, 110, 120, 130 and 140 are 5 binding sites with the lowest free energy to the highest, i.e., fill point 100, 110, 120, 130, and 140. Binding sites 3, 4 and 5 have the enormous volumes 345 Å³, 372 Å³, and 395 Å³ respectively so that they have potential to bind with both ligand mefloquine and primaquine with volume 330 Å³ and 333 Å³ respectively. It can be concluded that all three of binding sites have the potential to inhibit merozoites raid of Plasmodium falciparum signal peptide peptidase in resistance to parasite life cycles.
Acknowledgment
The authors thank The Ministry of Research, Technology and Higher Education of the Republic of Indonesia for research funding through scheme Penelitian Dasar Unggulan Perguruan Tinggi to Setyanto Tri Wahyudi with contract number: 129/SP2H/PTNBH/DRPM/2018

References
[1] Akanmori BD, Okeibunor J and Moeti MR 2018 J. Rare Dis. Res. Treat. 3 11-14
[2] Ashley EA, Phyo AP and Woodrow CJ 2018 The Lancet 391 1608-21.
[3] Alam A, Goyal M, Iqbal MS, Pal C, Dey S, Bindu S, Maity P and Bandyopadhyay U 2009 Expert Rev. Clin. Pharmacol. 2 469-489.
[4] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ 1997 Nucleic Acids Res. 25 3389-402.
[5] Yang J, Yan R, Roy A, Xu D, Poisson J and Zhang Y 2015 Nature Methods 12 7-8.
[6] Laskowski RA, Chistyakov VV and Thornton JM 2005 Nucleic Acids Res. 33 266-88.
[7] Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS and Richardson DC 2009 Acta Crystallogr. Section D Bio. Crystallogr. 66 12–21.
[8] Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B and Woods R 2005 J. Comput. Chem. 26 1668-88.
[9] Anandakrishnan R, Aguilar B and Onufriev AV 2012 Nucleic Acids Res. 40 537-41.
[10] Gordon JC, Myers JB, Folta T, Shoja V, Heath LS and Onufriev A 2005 Nucleic Acids Res. 33 368-71.
[11] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS and Olson AJ 2009 J. Comput. Chem. 30 2785-91.
[12] Forli S, Huey R, Pique ME, Sanner M, Goodsell DS and Olson AJ 2016 Nat. Protoc. 11 905–19.
[13] Voss NR and Gerstein M 2010 Nucleic Acids Res. 38 555-62.