Sequence analysis and comparative modelling of nucleocapsid protein from *pseudomonas stutzeri*

S Omar, F Mohd Tap, K Shameli, R Rasit Ali, NW Che Jusoh and N B Ahmad Khairudin*

Chemical Energy Conversion and Application iKohza, Department of Chemical Process Engineering, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia

*r-bahiah@utm.my

Abstract. Protein structure can be determined by either experimental method or by computational prediction, which commonly known as homology modelling. The objective of this study is to predict the structure of protein from *Pseudomonas stutzeri* using homology modelling. The suitable template was identified with 80.35% of sequence identity determined by sequence alignment. MODELLER was used to predict the model using the method of satisfaction of spatial restraints. The model was then analysed by computational analysis tools such as Ramachandran`s Plot, Errat, Verify3D Profile evaluation, Secondary Structure Prediction Analysis and Protein Binding Site (ProBiS). The results suggest that the model is reliable and has good stereochemical properties.

1. Introduction

The purpose of homology modelling is to predict and develop three-dimensional structure of unknown protein based on the similar protein structure with an accuracy that could be equivalent to the results achieved from the experiment. Since, the experiment methods using NMR analysis and X-ray diffraction are time consuming, costly and hard to perform, it is more convenient to apply comparative modelling. Besides, the structural genomics prediction can use low resolution structures provided by homology modelling methods to find the target protein structure. According to Protein Data Bank (PDB) [1], the total amount of protein structure that can be accessed is about 125,000 structure in 2016, which year by year, the number of new found of protein structure escalate significantly. It is a big leapt in the world of protein in having computational prediction of 3D structure to determine the proteins function and their significant [2]. By the comparative modelling as alternative method, the protein prediction structure become easier and the known structure rapidly progress compare to before this method been introduced. The objective of this study is to generate and predict a tertiary model of nucleocapsid protein by homology modelling using method of spatial restraints.

2. Methods

The potential template was identified using BLAST [3]. The template was chosen based on the homologous of the known protein structure and their similarities must exceed more than 30% to avoid

alignment error. Sequence alignment between the template and the target sequence was performed using UCSF Chimera program [4]. MODELLER program was used to model the structure by using method of spatial restraints [5]. It uses either distances or optimization techniques to satisfy the spatial restrains. The model was analysed based on Ramachandran Plot and protein tertiary structure environment.

3. Results and Discussion

Position-Specific Iterated BLAST (PSI-BLAST) is another method to search for the template. It is an extended program that can search in more wide scope compared to BLAST. Furthermore, it can help to find more distant relative sequences by searching with a custom position-specific scoring matrix (PSSM). Therefore, in this study, PSI-BLAST method is chosen to get a better template for protein *Pseudomonas stutzeri* (R4RRT6). Protein 1EX9 has higher percentage of identities that is 80.35%, compared to 4HS9 which is only 44.44%, it is strongly preferable to use protein 1EX9 as a template to create the 3D homology modelling for the target protein. However, this study employs multiple sequence alignment as shown in figure 1 that involved both template proteins to gain better model.

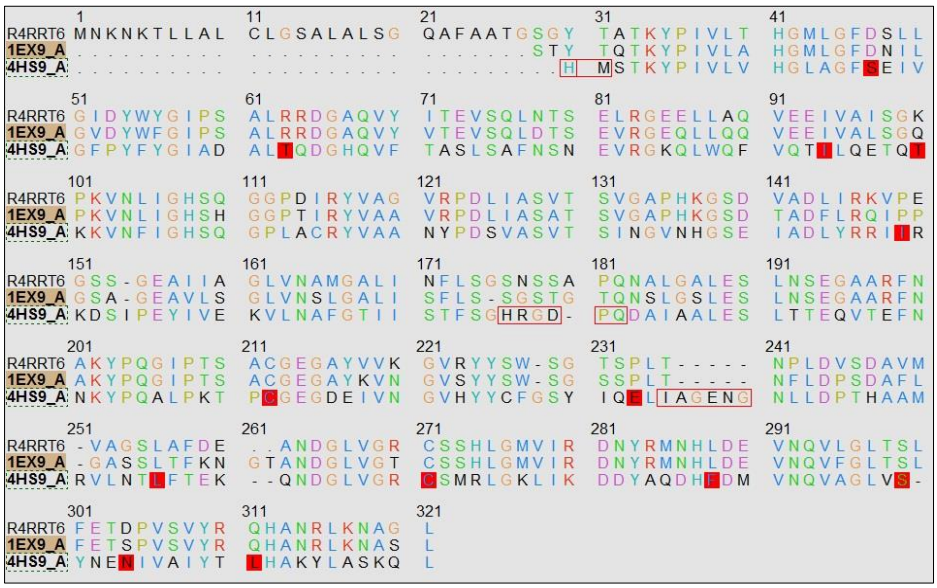

Figure 1. Multiple Sequence Alignment of *Pseudomonas stutzeri* (R4RRT6) with 1EX9 and 4HS9.

For the structure of protein R4RRT6, it was constructed using MODELLER version 9.18 which generated 20 models. The top three models (Model 1, Model 2 and Model 20) that had the lowest energy values were selected for further analyses. Objective function is the free energy inside of the model structure that had been generated. High objective function means that it has higher free energy which cause a bumping effect that happens between molecules in the model and consequently lowered the stability of a model. So, it is more preferable to choose a model with lower energy value for further analysis as it has high stability and no bumping effect.

Figure 2 shows the tertiary structure prediction for protein R4RRT6 based on the template 1EX9. The secondary structure prediction analysis was done using Polyview program and the result is presented in figure 3. In the figure, the red line represents helices, the blue line represents the coils or loops while the green line represents beta strands. It was shown that the model contains eight helices. Table 1 shows the different percentage for alpha-helix, beta-sheet and random coil for each structures of the model. In this table, Model 2 has been analysed that the number of percentage for secondary protein structure which is the $\boldsymbol{\alpha}$-helix, $\boldsymbol{\beta}$-sheet and random coil is about the same with Model 20 which then followed with Model 1.

Table 1. Percentage of Secondary Protein Structure for the Comparison on Model 1, Model 2 and Model 20 of Protein Structure.

Secondary Protein Structure	Model 1	Model 2	Model 20
α-helix	36.27%	38.38%	38.73%
β- sheet	11.27%	14.44%	14.79%
Random Coil	52.46%	47.18%	46.48%

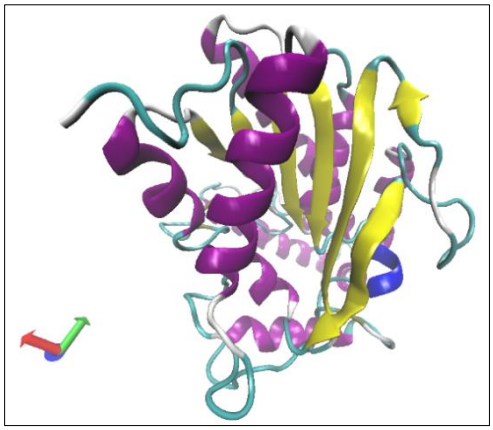

Figure 2. Three Dimensional Structure of Protein for *Pseudomonas stutzeri* (R4RRT6)

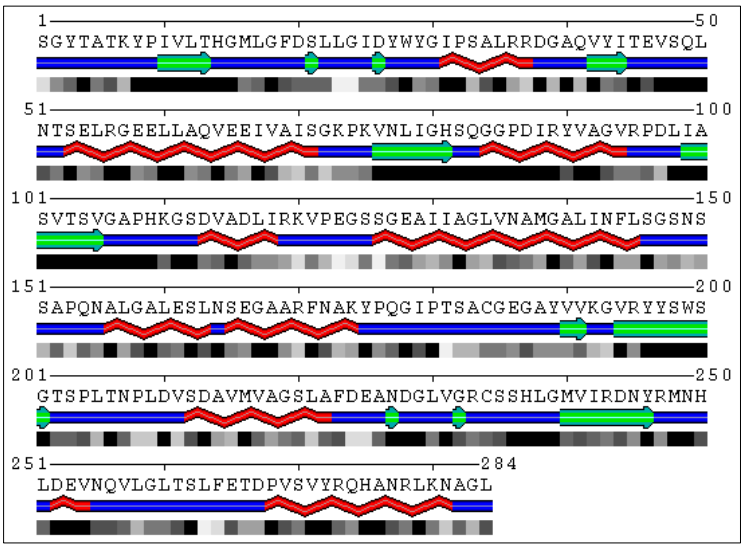

Figure 3. Secondary Structure Prediction

The first analysis on assessing the quality of the model was done using ERRAT [6]. It is beneficial to assess the arrangement of different types of atoms in each protein models. ERRAT is an "overall quality factor" for non-bonded atomic interactions which the higher scores mean higher quality. Normally, the accepted range is more than 50 for a high quality model. In this current case, the Errat score for the model is 94.928. Figure 4 shows Errat plot for Model 2. Since, the overall quality factor is 94.928, it is proven that the tertiary structure of protein R4RRT6 that been generated is a high quality model.

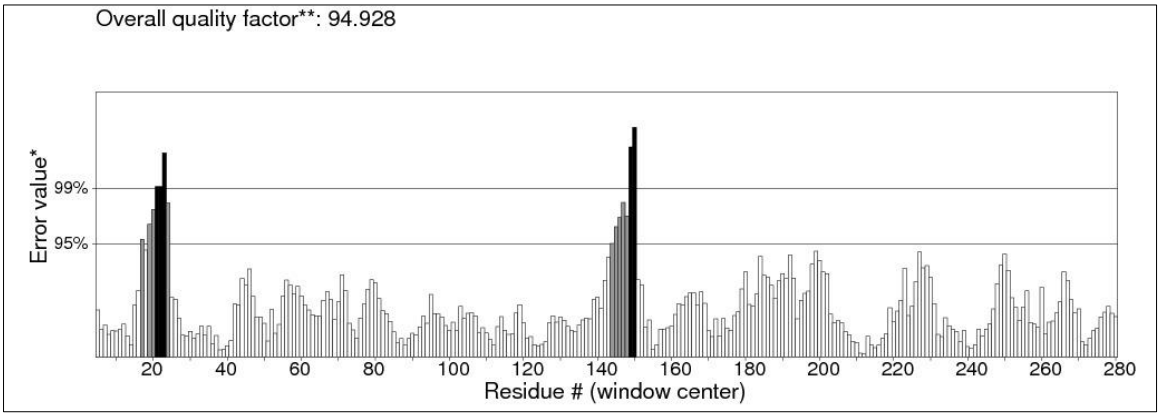

Figure 4. Errat Plot for Model 2

Results from Ramachandran Plot analysis shows that 96.1% of the residues were located in the most favorable core region and 0.0% was in the disallowed region. This result can be considered as significant because of the high percentage of residues in favoured region (>90%). Table 2 shows the summary of the result when comparing the tertiary structure between Model 1, Model 2 and Model 20. Figure 5 shows the Ramachandran Plot for the best model which is Model 2.

Table 2. Percentage of Amino Acid Residues in Each Region for Ramachandran Plot

Model	Ramachandran Plot (%)			
	Core	Allowed	Generous	Disallowed
1	96.5	2.8	0.7	0.0
2	96.1	3.2	0.7	0.0
20	95.4	3.9	0.7	0.0

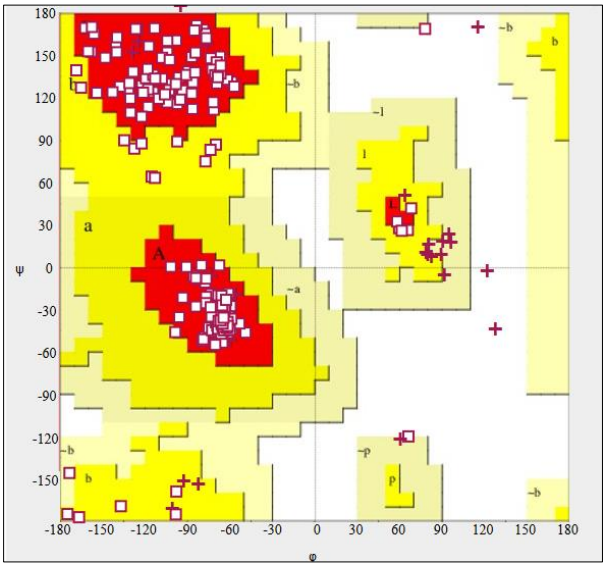

Figure 5. Ramachandran Plot of Predicted Structure of R4RRT6 (model 2)

Figure 6 shows the result of Verify 3D Profile for Model 2 [7]. In the figure, all of the plotted lines were in the allowed range. Therefore, it was indicated that each of the model structures obtained represents a high quality model. The score was represented by vertical axis while horizontal axis represents the protein`s residues [Rosetti et al., 2008]. Verify 3D Profile assessment is used to analyse the compatibility of an atomic model (3D) with their own amino acid sequences. Each residue is assigned a structural class which based on their location and environment (alpha, beta, loop, polar, non-polar and others). Furthermore, in the structural class, a high resolution structure is used to obtain score for each of the 20 amino acids. The scores range from -1 (bad score) to +1 (good score).

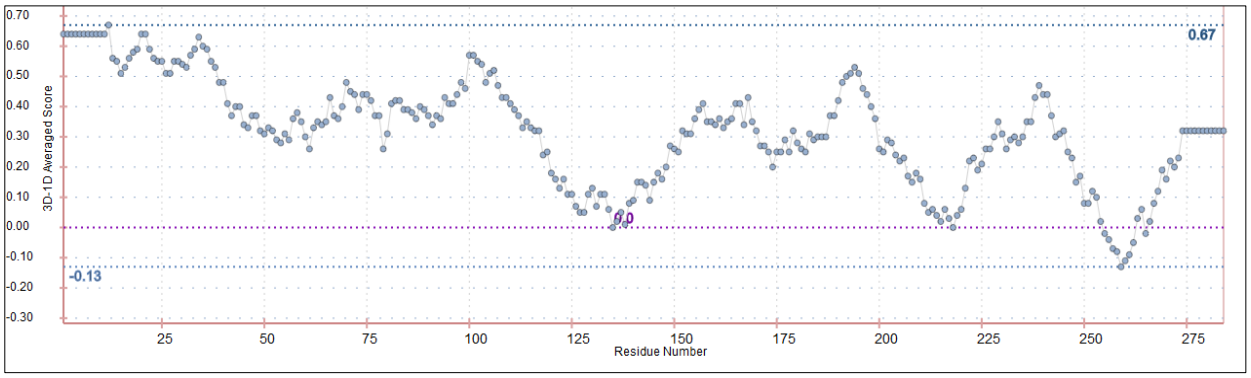

Figure 6. Verify 3D Result for Model 2

The protein binding site analysis was completed using ProBiS which is commonly use to detect structures which have similar binding sites in proteins and local pairwise alignment of crystallographically or NMR determined protein structure from the PDB. Binding site is a location where the protein binds to a ligand. Since only a few residues actually participate in binding the ligand, the other residues usually act as a framework to provide correct conformation and orientation. Figure 7 shows the proposed binding site for the nucleocapsid protein from *Pseudomonas stutzeri* (R4RRT6).

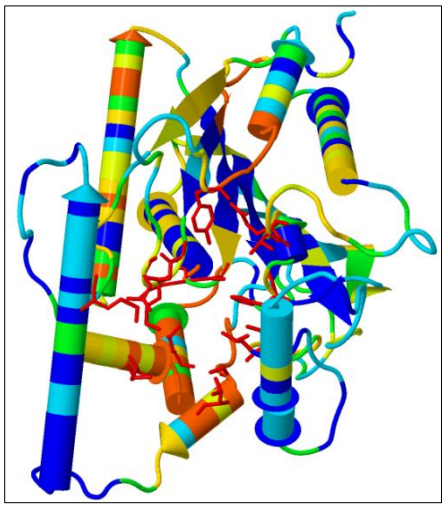

Figure 7. The Predicted Binding Site (Red Region) for Nucleocapsid Protein from *Pseudomonas stutzeri* (R4RRT6)

4. Conclusion

In this study, homology modelling is used to generate 3D structure of protein R4RRT6 by satisfying the spatial restraints of MODELLER program. The suitable template for protein R4RRT6 was selected from the result of multiple sequence alignment which was 1EX9 with sequence identity of 80.35%. Various tools were used to refine and assess the quality of the predicted model.

5. Acknowledgement
The authors would like to thank Universiti Teknologi Malaysia for providing the scholarship and financial support for this project (GUP Vot No: 19H73).

6. References

[1] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE 2000 *Nuc. Acids Res* **28** 235

[2] Schoenrock A, Samanfar B, Pitre S, Hooshyar M, Jin K, Phillips CA, Wang H, Phanse S, Omidi K and Gui Y 2014 *BMC Bioinformatics* **15** 383.

[3] Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ 1990 *J. Mol. Biol.* **215** 403

[4] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC and Ferrin TE, 2004 *J. Comput. Chem.* **25** 1605

[5] Sali A and Blundell TL1993 J. *Mol. Biol.* **234** 779

[6] MacArthur A 1994 Curr. *Opin. Struct. Biol.* **4** 731

[7] Bowie JU, Lüthy R and Eisenberg D 1991 *Science* **12** 164