Ab-Initio Protein Structure Prediction of *Leucosporidium antarcticum* Antifreeze Proteins Using I-TASSER Simulations

MOHD BASYARUDDIN ABDUL RAHMAN[1]*, MOHAMMAD FAIRUZ ZULKIFLI[1], ABDUL MUNIR ABDUL MURAD[2], NOR MUHAMMAD MAHADI[2], MAHIRAN BASRI[1], RAJA NOOR ZALIHA ABDUL RAHMAN[3] and ABU BAKAR SALLEH[3].

[1]Department of Chemistry, Faculty of Science,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, MALAYSIA
[2]Malaysia Genome Institute,
43600 UKM Bangi, Selangor Darul Ehsan, MALAYSIA
[3]Laboratory of Industrial Biotechnology, Institute of Bioscience,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, MALAYSIA

Abstract: - Organisms living in cold environment produce some Antifreeze Protein (AFP) which exhibit special functions as a result of cold adaption. AFP is currently being identified in many organisms such as bacteria, plants, fish, and fungi that exposed to freezing stress. Due to the limited structural information from fold library, it gave a big challenge in its structure prediction. Therefore, this study seeks to predict the three-dimensional (3D) model of the *Leucosporidium antarcticum* antifreeze protein by using homology modeling, threading and *ab-initio* methods. As low of percentage of sequence identity, not more than 25% ('twilight zone') and poor results in threading methods, the search proceeded with *ab-initio* method by using I-TASSER simulations, where 5 predicted models were obtained. All the models were then evaluated with PROCHECK and Verify3D servers. Ramachandran Plot showed that the residues in most favored regions were 75.2% with only 4 residues in disallowed regions (Ser21, Phe29, Ala100 and Ala114). For the Verify3D, the structurally and functionally important residues in AFP have scored from 0.30-0.60. These results suggest that *ab-initio* methods as I-TASSER may soon become useful for low-resolution structure prediction for proteins that lack of close homologue of known structure.

Key-Words: - antifreeze proteins (AFP), Leucosporidium, homology modeling, fold recognition/threading, ab-initio, I-TASSER

1 Introduction

Bioactivity screening of antarctic microflora has yielded numerous enzymes that active at low temperatures. It was found that organisms living in extreme environment produce some antifreeze protein which exhibit special functions as a result of cold-adaptation. The protein was first found in species of fish that have adapted to extremely cold temperatures by using the protein [1]. AFP is currently being identified in many organisms exposed to freezing stress. They are found in bacteria, insects, plants and fish, but only the eukaryotes are biochemically well characterized. Psychrophilic bacteria were also found to bear such AFP albeit of a different type. The exact mechanism is still unsure, but the protein interacts with ice in a way that makes the organism less sensitive to cold temperatures. This helps the organism to minimize the internal damage caused by the cold temperature [2]. AFP is believed to interact complementary with the prism plane of ice crystal so as to depress the freezing pointy non-colligatively. Because of its

unique function, AFP has been regarded to possess high potential for industrial applications (e.g. food, artificial rain). However, the detailed molecular basis of these functions has not been provided so that it does not lead to a new material production.

Functional characterization of a protein sequence is one the most frequent problems in biology [3]. This task is usually facilitated by accurate three-dimensional (3D) structure of the studied protein. In the absence of an experimentally determined structure, comparative or homology modeling can sometimes provide a useful 3D model for a protein that is related to at least one known protein structure [4]. Comparative modeling predicts the 3D structure of a given protein sequence (target) based primarily on its alignment to one or more proteins of known structure (templates). The prediction process consists of fold assignment, target-template alignment, model building, and evaluation [5]. The number of protein sequences that can be modeled and the accuracy of the predictions are increasing steadily because of the growth in the number of known protein structures and also because of the improvements in the modeling software. It is currently possible to model with useful accuracy significant parts of approximately one half of all known protein sequences [6].

Although the *homology modeling* method seems the most reliable, it can be applied only when 3D structure of a similar sequence is already known. In order to overcome this drawback, Bowie, Luthy and Eisenberg proposed a threading method [7]. In this method, given an amino acid sequence and a set of protein structures (or structural patterns), a structure into which the sequence is most likely to fold is computed. An *alignment* between amino acids of a sequence and spatial positions of a 3D structure is computed using a suitable *score function* in order to test whether or not a sequence is likely to old into a structure. The process of computing an alignment between a sequence and a structure is called *protein threading*, and its alignment, a *threading*.

The *ab-initio* prediction methods consist in modeling all the energetic involved in the process of folding, and then in finding the structure with lowest free energy. This approach is based on the 'thermodynamic hypothesis', which states that the native structure of a protein is the one for which the free energy achieves the global minimum. While *ab-initio* prediction is clearly the most difficult, it is arguably the most useful approach. Even models with errors may be useful, because some aspects of the function can be predicted from just coarse structural features of a model [8], [4]. I-TASSER simulation results showed that it can consistently predict the correct folds and sometimes high resolution models for small single-domain proteins. Compared with other *ab-initio* modeling methods such as ROSETTA and TOUCHSTONE II, the average performance of I-TASSER either much better or is similar within a lower computational time. These data, together with the significant performance of automated I-TASSER server (the Zhang-Server) in the 'free modeling' section of the recent Critical Assessment of Structure Prediction (CASP)7 experiment, demonstrate new progresses in automated *ab-initio* model generation [9], [10], [11].

2 Methods

Data Mining and Sequence Analysis

The linear chain of AFP containing 177 residues was subjected to various sequence analysis on SWISS-PROT [12], PDB [13], BLAST [14] and PSI-BLAST [15].

```
MRSNFHPLAASFIVRCAFLHSRRFTDSLFQLLSSLISLTSAATAIDLGV
AGQYDVVARSAITLGALAEITGNVGLSPGLSTALTGFTLVPVEDHGT
FCSAGVKYCGADSLSTSATSLLVKGRIDAPDFPSSPAILGQAATDVV
AAWKSAFSQELSPADYTKRDFAGGLLSDLTLAPG
```

Fig.1: Sequences of amino acids of *Leucosporidium antarcticum* AFP.

Pair-wise and multiple sequence alignment between AFP sequence and the templates were carried out using CLUSTALW [16]. Superfamily HMM [17] and PSI-BLAST [15] were used to identify any conserved domains or families found in the protein.

Model Development and Evaluation

Five secondary structure prediction methods were used in this work to obtain the information on the secondary structure: Sspro [18], PHD [19], GOR4 [20], FI-Pred [21] and Jnet [22]. The amino acids sequence was then threaded to the library that contained already known protein fold by using mGenThreader [23], 3DPSSM [24] and FUGUE [25]. *Ab-initio* method was used to develop the protein 3D model and I-TASSER [9], [10], [11] as the web server. The resulted model was evaluated using SWISS MODEL [26], [27], [28], [29], [30] (PROCHECK [31] and Verify3D [32]).

3 Results and Discussion

3.1 Data Mining and Sequence Analysis

The AFP residues(Fig.1) was subjected to sequence analysis on PDB using PSI-BLAST but there was no similarity because of lack of PDB files in the protein-type of fungi. In the templates searching, PDB web server was used and 'antifreeze protein' was used as the keyword. As a result, 54 templates of antifreeze protein were found. In sequence alignment, multiple alignments of ClustalW were used as the web server. From ClustalW, the results showed that all of the templates contained low percentage of identity which means they were all in the 'twilight zone' [33]. 'Twilight zone' is a zone where the template contained percentage of identity that less than 30% and it is not suitable to be use as a template in the homology /comparative modeling. As the similarity between the target and the templates decreases, alignments contain an increasingly large number of gaps and alignment errors. Errors in comparative models can be divided into five categories; errors in sidechain packing, distortions and shifts in correctly aligned regions, errors in regions without template, errors due to misalignments and incorrect templates.

3.2 Secondary Structure Prediction

In the secondary structure predictions, there were five web servers had been used; Sspro, PHD, GOR4, FI-Pred and Jnet. Each of these web servers had their own accuracy and the overall result can be seen in the sequence consensus. As a result, 45.20% of the sequence was random coils (C), 44.63% was alpha helix (H), 6.78% were beta strand (E) and the rest was unknown (?). Unknown result means that the region had the same amount amino acids. All of these data will be used as a reference to the AFP model that will be built.

3.3 Template Selection

Because of the low percentage of identity in the sequence alignment, the first two steps in the homology modeling had been substituted by fold recognition/threading method. In this method, even with no homologue of known 3D structure, it may be possible to find a suitable fold for the protein among 3D structures. It had been recognized that proteins often adopt similar folds despite (without being affected by) no significant sequence of functional similarity. There were 3 web servers that had been used in this method; mGenthreader, 3DPSSM and FUGUE.

Table 1: Summary of mGenthreader e-value results in threading method.

Confidence levels	e-value	PDB ID	Percentage of identity
GUESS	0.181	1wd7A0	13.0
GUESS	0.197	2bm8A0	12.4
GUESS	0.202	1oj7A0	10.7
GUESS	0.234	1e3hA0	15.8
GUESS	0.259	1ezwA0	10.2
GUESS	0.289	1h12A0	8.5
GUESS	0.294	1woqA0	22.0
GUESS	0.302	1zvwA0	22.6
GUESS	0.308	1rrmAo	15.8
GUESS	0.309	1wn1A0	11.3

Table 2: Summary of FUGUE z-score results in threading method.

Profile hit (PDB ID)	Z-score	Confidence levels
hs2coua	2.02	GUESS
hs1u6hb	2.00	GUESS
hsd1fcda3	1.97	UNCERTAIN
hs1pq1b	1.97	UNCERTAIN
hs2hina	1.96	UNCERTAIN
VPR	1.93	UNCERTAIN
hs1ckkb	1.93	UNCERTAIN
hs1zvzb	1.90	UNCERTAIN
hs1xsza	1.90	UNCERTAIN

Table 3: Summary of 3DPSSM e-value results in threading method.

PDB ID (SCOP Code)	e-value
d1kkea1	95
d1pgl22	1.2e+02
d1ntha	1.4e+02
d1puga	1.4e+02
c1o8ta	1.5e+02
d1c8ba	1.6e+02
c1ybxa	1.6e+02
d1hjra	1.8e+02
c1o12b	1.9e+02
dv10a2	1.9e+02

mGenThreader showed that the lowest e-value for the threading is 0.181 and it is in the GUESS confidence level. In FUGUE, recommended cutoff for z-score is above 6.0 (certain 99% confidence level) but the highest scoreis only 2.02 which means GUESS 50% confidence level. For 3DPSSM, the lowest e-value is 1.2e+02. All the threading results showed low confidence level and we can conclude that homology modeling method is not suitable anymore to be used in AFP structure prediction.

3.4 Model Development and Evaluation

Because of low percentage of sequence identity in the sequence alignment and poor result in the threading methods, we proceed to the *ab-initio* methods by using I-TASSER [9], [10], [11]. Simulation results show that I-TASSER can consistently predict the correct folds and sometimes high-resolution models for small single-domain proteins. Compared with other *ab-initio* modeling methods such as ROSETTA and TOUCHSTONE II, the average performance of I-TASSER is either much better or similar within a lower computational time [9]. As a result, there were five predicted AFP models (Fig. 2). Each model had it own accuracy and the most accurate model was model (a). All the models were built based on PPA (profile-profile alignment) threading alignments and iterative TASSER (threading, assembly, refinement) simulations. In PPA method, a computer program forces the sequences to adopt every known protein in turn, and for each case a scoring function is calculated. This measures the suitability of the sequence for that the particular fold. The model (a) was picked and evaluated using PROCHECK [31] and Verify3D servers [32].

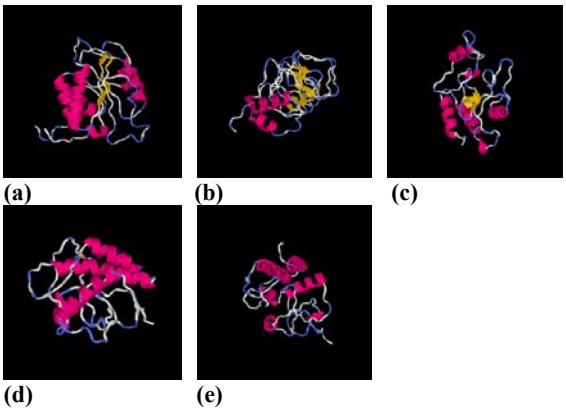

Fig.2: Results of predicted models of AFP (a,b.c,d,e) from I-TASSER simulations.

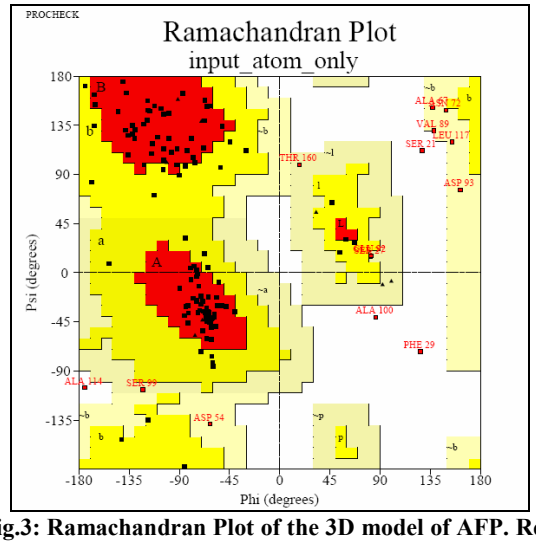

Fig.3: Ramachandran Plot of the 3D model of AFP. Red region represents the most favored region, yellow = allowed region, light yellow = generously allowed region, white = disallowed region.

PROCHECK analyses (Fig.3) showed that only 4 residues were located in the disallowed region of the Ramachandran Plot [31] and 75.2% of the residues were located in the most favored region **(**red region**)** The other residues were found to reside in the additional and generously allowed regions. The four residues that were located in the disallowed region were Ser21, Phe29, Ala100 and Ala114. From the Verify3D [32] analysis, it showed that the structurally and functionally important residues in AFP have score from 0.30-0.60 and it indicates that the quality of the model is not good as in high resolution crystal structures but satisfying. Verify-3D scores below or near 0.0 reflect structures that are almost certainly incorrect. Verify-3D scores near 1.0 reflect scores similar to that expected for a valid protein of the same size. It was also found that 73.47% of the residues scored more than 0.2, meaning that ~73% of the residues complemented with the 1D-3D model. For the quality of the predicted model to be considered satisfactory, it is expected to have the Verify3D score more than 80%. But for this model, the score it satisfying for a sequence that only had percentage of sequence identity in the 'twilight zone' [33].

4 Conclusion

We have performed sequence analysis and attempted to predict the 3D structure of AFP using the method of *ab-initio* due to very low similarity to any available experimentally solved 3D protein structures. A series of molecular modeling and computational methods were combined in order to gain insight into the 3D protein structure of AFP. With the 3D model, perhaps it can be used to seek the expected binding sites (α-helix) of the antifreeze protein for further research.

References:

[1] Davies P.L., Baardnes J., Kuiper M.J. and Walker V.K. 2002. Structure and function of antifreeze proteins. *Phil. Trans. R. Soc. Lond. B.* 357:927-935.

[2] Graether S.P., GagneS.M., Spyracopoulos L., Jia Z., Davies P.L. and Sykes B.D. 2003. Spruce Budworm Antifreeze Protein: Changes in Structure and Dynamics at Low Temperature. *Journal of Molecular Biology.* 327:1155-1168

[3] Baker, D. 2000. A suprising simplicity to protein folding. *Nature* 405:39-42.

[4] Marti-Renom M.A., Stuart A., Fiser A., Sanchez R., Melo F., Sali A. 2000. Comparative protein structure modeling of genes and genomes *Annu Rev Biophys Biomol Struct* 29:291-325.

[5] Marti-Renom M.A., Yerkovich, B., Sali, A. 2002. Comparative protein structure prediction. *Current Protocols in Protein Science 2.*

[6] Pieper U., Eswar N., Ilyn V.A., Stuart A., Sali A. 2002. Mosbase, A database of annotated comparative protein structure models. *Nucleic Acids Res* 30:255-259.

[7] Bowie, J.U., Luthy, R., and Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional structures, *Science*, 253:164{170, 1991.

[8] Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* 294: 93-96.

[9] Sitao Wu, Jeffrey Skolnick, Yang Zhang. 2007. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology,* vol 5, 17.

[10] Yang Zhang. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, vol 9, 40.

[11] Yang Zhang. 2007. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins, Suppl* vol 8, 108-117.

[12] Bairoch A. and Boekmann B. 1992. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* 20:2019-2022.

[13] Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., weissig H., Shindyalov

I.N. and Bourne P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235-242.

[14] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol*. 215: 403-410.

[15] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search program. *Nucleic Acids Res.* 25: 3389-3402.

[16] Thompson, J.D., Higgins, D.G., Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-spesific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.

[17] Madera M., Vogel C., Kummerfield S.K., Chotia C. and Gough J. 2004. The SUPERFAMILY Database in 2004 :Additions and Improvements. *Nucleic Acids Res.*32:235-239.

[18] Pollastri G., Pryzbylski D., Rost B. and Baldi P. 2002. Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins*. 47:228-235.

[19] Rost B. and Liu J. 2003. The Predictprotein Server. *Nucl Acids Res.* 31:3300-3304.

[20] Garnier J., Gibrat J.F. and Robson B. 1996. In: R.F. Doolittle, Editor, *Methods in enzymology* **266**, Academic Press, New York pp. 540–553.

[21] McGuffin L.J., Bryson K. and Jones D.T. 2000. The PSIPRED Protein Structure Prediction Server. *Bioinformatics.* 15:404-405.

[22] Cuff J.A. and Barton G.J. 1999. Application of Enhanced Multiple Sequence Alignment Profiles To Improve Protein Secondary Structure Prediction. *Proteins.* 40:502-511.

[23] Jones D.T. 1999. An Efficient And Reliable Protein Fold Recognition Method For Genomic Sequences. *J Mol Biol.* 287:797-815

[24] Kelley L.A., MacCallum R.M. and Sternberg M.J.E. 2000. Enhanced Genome Annotation Using Structural Profiles in The Program 3DPSSM. *J Mol Biol.* 299:499-520.

[25] Shi J., Blundell T.L. and Mizuguchi K. 2001. Sequence-structure Homology Recognition Using Environment-spesific Substitution Tables and Structure-dependent Gap Penalties. *J Mol Biol.* 310:243-257.

[26] Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. *Bioinformatics,* 22,195-201.

[27] Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. *Electrophoresis* 18: 2714-2723.

[28] Kopp J. and Schwede T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models Nucleic Acids Research 32, D230-D234.

[29] Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.

[30] Schwede T, Kopp J, Guex N, and Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* 31: 3381-3385.

[31] Laskowski R A, MacArthur M W, Moss D, Thornton J M (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst*., 26, 283-291.

[32] Luthy R, Bowie JU, Eisenberg D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356:83-85.

[33] Rost B. 1999. Twilight Zone of Protein Sequence Alignments. *Protein Eng.* 12:85-94.