



Comparative Modeling, Quality Assessment and Validation of HYD1

Cephas Mawere¹, Suresh Kumar G², Shumbeyi Muzondo³

M.Tech. Student, Department of Bioinformatics, School of IT, Jawaharlal Nehru Technological University, Hyderabad,
India¹

Lecturer, Department of Bioinformatics, School of IT, Jawaharlal Nehru Technological University, Hyderabad, India²

M.Tech. Student, Department of Biotechnology, IST, Jawaharlal Nehru Technological University, Hyderabad, India³

ABSTRACT: HYD1 is a small secreted fungal protein that belongs to class I hydrophobins of *G. moniliformis*. It is able to spontaneously self-assemble at hydrophilic: hydrophobic interfaces into a 10 nm amphipathic rodlet monolayer. Based on this unique property, HYD1 is involved in signaling and attachment of aerial structures (like mycelia, hyphae, and spores) to hydrophobic surfaces of host plant during fungal pathogenesis. This paper seeks to construct a 3D structure of HYD1 based on multiple templates using the novel algorithm available on the M4T online server. PROCHECK, ProSA, Verify3D and 3dss server are used to analyze the quality and consistency of the generated model. The M4T_HYD1 model thus had a compatible DOPE score value of -5547.2 and showed 93.0% amino acid residues under the most favored region [A, B, L]. Evaluation of the 3D model and two templates (2f26C and 2pl6A) with ProSA revealed a compatible Z score value of -3.43, -4.31, and -4.10 respectively, indicating a comparable surface energy affinity of M4T_HYD1 with the templates. A low RMSD value between the target and the template structures upon superimposition reflects high structural connection. Hence, the predicted model proved to be well ratified in terms of geometry and energy contours, indicating that the model is reasonable and reliable for future molecular docking studies.

KEYWORDS: HYD1, Hydrophobins, Self-assemble, Amphipathic, M4T algorithm, DOPE, Superimposition

I. INTRODUCTION

Hydrophobins are small (+/- 100 aa) surface active molecules [1] uniquely found in filamentous fungi (where the destructive plant pathogen *G. moniliformis* is an example) [18]. They are characterized by eight conserved cysteine residues and a conserved spacing of hydrophilic and hydrophobic regions [2]. They constitute 10% of the total cell wall as structural components and they are usually secreted into liquid media by fungi and localized on the surface of aerial mycelia [3]. They are involved in the formation of hydrophobic aerial structures like aerial hyphae, spores and fruiting bodies [1], and play a vital role in fungal development, environmental adaptation and pathogenic interactions. In pathogenesis, for example, hydrophobins mediate attachment of hyphae to each other [3] and to the hydrophobic surface of the host plant before penetration and infection commences [4].

Based on hydrophobicity pattern, solubility and the type of layer they form, hydrophobins are divided into two groups-class I and class II [2]. HYD1 is a class I hydrophobin that is actively involved in fungal growth and development. Like other class I hydrophobins, HYD1 can spontaneously self-assemble into a 10 nm wide amphipathic rodlet film in the presence of a hydrophilic: hydrophobic interface [5]. Thus, the hydrophobic side is exposed to the exterior whilst the hydrophilic surface is bound to the cell wall. In this way, the HYD1 layer is also able to lower the surface tension of the medium or substratum in/on which *G. moniliformis* will grow [6]. This prevents water-logging while maintaining permeability to gaseous exchange, and facilitates spore dispersal. The rodlet layer is covered by a mucilaginous extracellular matrix that helps the fungal conidia to bind to the substrate [6]. This makes class I hydrophobins, which includes HYD1, to be essential for the *G. moniliformis* to complete its biological life cycle. However, the absolute understanding of the cause of the elastic film formation is yet to be understood [7].

The self-assembly property of HYD1 causes it to be insoluble in aqueous solvents and thermally resistant to heat treatments below 90°C, only to be dissociated by harsh treatment with strong acids such as concentrated TFA [8]. The



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

unique properties of HYD1 has caused *G. moniliformis* to be a destructive pathogen to more than 11,000 plant species including maize and rice [9]. This has resulted in the huge grain loss worldwide [10] accelerating the need to design potential antifungal agents which can inhibit HYD1 using *in silico* based methods. This however relies on knowledge of three dimensional structure of the biological target obtained through methods such as x-ray crystallography or NMR spectroscopy [11]. Nevertheless, the crystal structure of HYD1 is currently unavailable in the Protein Data Bank (PDB). Hence, there is need to create a homology model of HYD1 based on the experimental structure of related protein/s (template/s).

Currently available modeling programs, and especially the automated servers, typically consider one template to comparatively build the desired model for a target sequence [12]. Meanwhile results at CASPS meetings and other reports indicate that the use of multiple templates improves the quality of the comparative models [13, 14, 15]. Moreover, alignment methods commonly used are not consistently producing better alignments throughout the regions of compared sequences. This has prompted the use of the M4T server to produce accurate alignments and minimize the errors associated with template recognition and alignment [12]. The low homology of class I hydrophobins further confirms the need to use M4T server in comparative modeling of HYD1 since the M4T algorithm is known to improve alignment accuracy, especially at low sequence identity levels [12].

II. RELATED WORK

M4T server has been cited and used to construct reliable protein models in more than eight different articles. The modeled proteins range from transmembrane domains of serotonin receptors, to kinase enzyme, to ribonucleoproteins, as well as potassium channels [12]. Based on single templates, comparative modeling of proteins were done using SWISS MODEL and Prime tools (which are automated servers) to build the 3D structure of the human Nek6 protein [16], in the construction of DHP receptor [17] and also in the development of novel virostatic agents against bluetongue virus [18].

III. MATERIALS AND METHODS

A FASTA format of the HYD1 sequence (126 amino acids long) was taken from the UNIPROT database (ID: Q6FY29) [19]. The fully automated comparative protein structure modeling server Multiple Mapping Method with Multiple Templates (M4T) was used to construct the 3D HYD1 model. The novelty of M4T resides in two of its major modules, Multiple Templates (MT) and Multiple Mapping Method (MMM). Here, the target sequence is used as query to search for homologous protein structure(s) that could serve as template(s) by running three iterations of PSI-BLAST against PDB with an *E*-value cut-off of 0.0001 [12].

Fig. 1 Home page of M4T server.

M4T Server ver. 3.0
Comparative Modelling using a combination of multiple templates and iterative optimization of alternative alignments
Bionformatics (2007) 23, 2558-65
Nucleic Acids Res (2007) 35, W363-68
J. Struct. Funct. Genomics (2009) 10, 95-9

Sequence Upload and Job Description

Enter Target Sequence

Paste your sequence here (raw sequence ONLY)

```
MGYMTIVAFLAATVAAGQIRAYPSIDQITVAQANNACGNBMQVTCCKRVNTFAG  
NAVNSGAGILNNLSLFDQCCKLQVNVLAIANGLLNKCCQANAAACCNSGGSATGGL  
VNVALPCTALSSLI
```

OR upload a text file: No file chosen

Job Submission

Send Results to: [your email address]

Enter Job Description: [optional]

Server Policy

I am a non-profit/academic user and this server will be used solely for educational purposes or for basic research intended to advance scientific knowledge.

After searching the PDB an iterative clustering procedure identifies the most suitable templates to combine in the MT-

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

module. Templates are selected or discarded according to a hierarchical selection procedure that accounts for sequence identity between templates and target sequence, sequence identity among templates, crystal resolution of the templates and contribution of templates to the target sequence. This results in clusters of templates in which all templates are aligned to the corresponding target sequence using the iterative-MMM approach. To construct profiles, the sequences of the target and template(s) are searched against the non-redundant database (NR).

BlastProfiler is run to build representative sequence profiles for both the target and template sequences. BlastProfiler parses all iterations of PSI-BLAST outputs, locates and stores those pairwise alignments between the query and database sequences that meet the filtering criteria (as reported in the pairwise Blast alignment) [12]. Three alternative profile-to-profile-based sequence alignments were generated from a combined implementation of CLUSTAL-W and MUSCLE algorithms, and used as input to MMM (Fig. 3). The MMM then iteratively compares and ranks alternatively aligned regions according to their fit in the structural environment of the template structure.

Models are then built with Modeller using the default values for HYD1 model.top routine. Selected template(s) and optimized alignment(s) from the MT and MMM modules described above are provided as inputs (Fig. 2). Two measures are calculated to assess model quality- DOPE (Discrete Optimized Protein Energy) and ProSA2003. The DOPE and PROSA2003 scores can be found in the header of the calculated coordinate file of the model while a separate html link leads to the PROSA2003 energy profile plot [12].

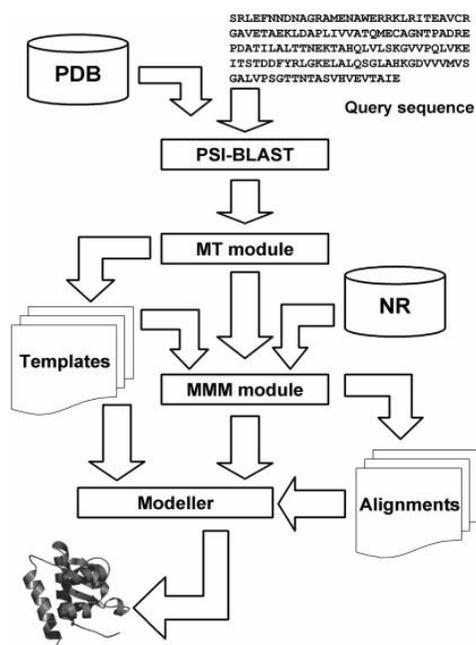


Fig. 2 General overview of the M4T algorithm (Source: Fiser [12]).

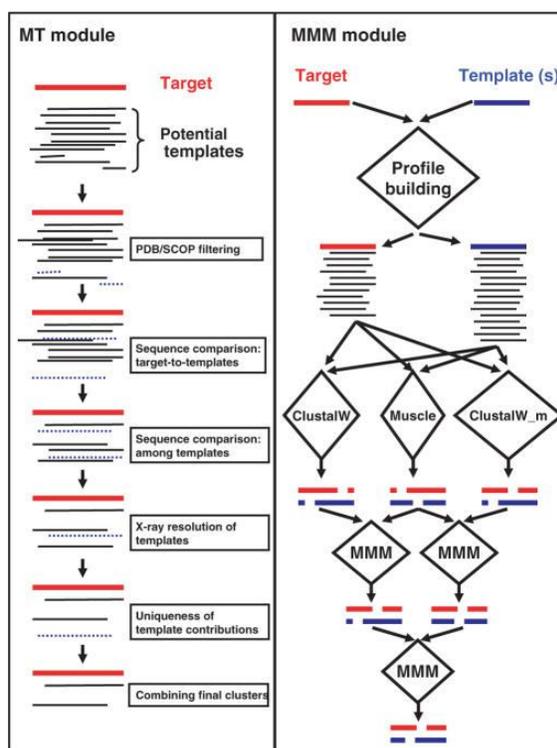


Fig. 3 Details of the MT and MMM modules of M4T. (Source: Fiser [12])

In Fig. 2 a PSI-BLAST search is first performed with a query sequence, then template(s) are selected in the MT-module; subsequently, MMM-module performs sequence-to-structure alignment(s), and finally Modeller builds the protein.

In Fig. 3; in the MT module the template candidates go through an iterative clustering and filtering process to select the least number of templates with a unique contribution to the target. The MMM module is an iterative implementation of the original Multiple Mapping Method using sequence profiles [12].

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

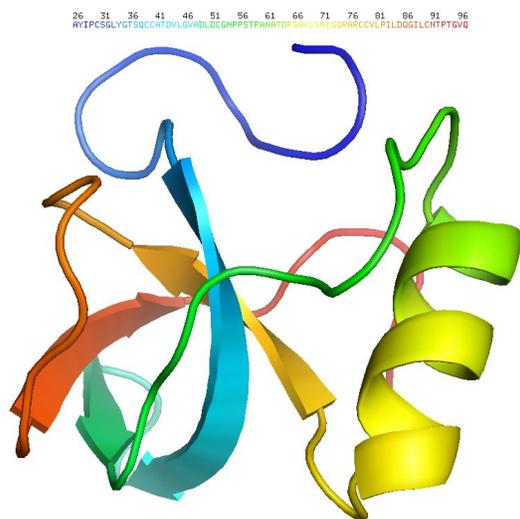


Fig. 5 M4T_HYD1 model

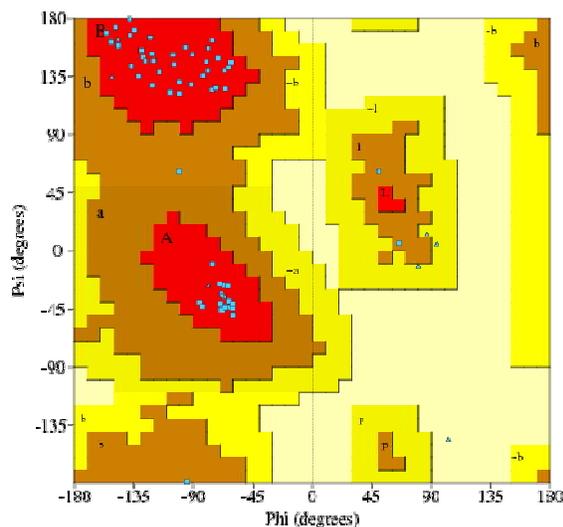


Fig. 6 Ramachandran plot for the M4T_HYD1 model

According to PROCHECK statistics, the distribution of the phi and psi angles for the amino acid residues was represented by Ramachandran plot (Fig. 6). The generated model was thus found to be highly plausible with 93.0% of the amino acid residues in the most favoured regions [A, B, L] (red colored areas). Such a satisfactory quality is based on the analysis of 118 structures of resolution of at least 2.0 Angstroms and R factor no greater than 20.0 where a good quality model is expected to have over 90% in the most favoured regions [20]. On the other hand, G-factors provide a measure of how unusual a property is. Values below -0.5* are considered to be unusual while values below -1.0** are termed highly unusual. A low or a negative G-factor indicates a low probability conformation [20]. However, G-factor assessment of the model reveals that all properties (dihedral angles and main-chain covalent forces) were in the ordinary range.

PROCHECK statistics

1. Ramachandran Plot statistics

	No. of residues	%-tage
Most favoured regions [A,B,L]	53	93.0%
Additional allowed regions [a,b,l,p]	4	7.0%
Generously allowed regions [~a,~b,~l,~p]	0	0.0%
Disallowed regions [XX]	0	0.0%

Non-glycine and non-proline residues	57	100.0%

End-residues (excl. Gly and Pro)	2	

Glycine residues	7	
Proline residues	6	

Total number of residues	72	

2. G-Factors

Parameter	Score	Average Score

Dihedral angles:-		
Phi-psi distribution	-0.18	
Chi1-chi2 distribution	-0.28	
Chi1 only	0.46	
Chi3 & chi4	0.67	
Omega	-0.08	
		-0.05
		=====
Main-chain covalent forces:-		
Main-chain bond lengths	-0.17	
Main-chain bond angles	-0.27	
		-0.23
		=====
OVERALL AVERAGE		-0.12
		=====

Fig. 7 PROCHECK: Ramachandran Plot and G-Factor results

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

ProSA and Verify3D are two methods that are sensitive in distinguishing between overall correct fold and those with an incorrect fold. ProSA-Protein Structure Analysis program statistically analyses all available protein structures for potential errors. The energy of the structure is evaluated using a distance-based pair potential and a potential that captures the solvent exposure of protein residues. From this a Z-score and a plot of residue energies are therefore derived and displayed. The Z-score indicates overall model quality and measures the deviation of the total energy of the structure with respect to an energy distribution derived from random conformations [17]. The more negative the Z-score is, the higher is the quality of the protein structure. Quality assessment of the model via ProSA thus revealed that the M4T_HYD1 model matched NMR region of the plot with Z-score (-3.43) which is reliable to the Z-scores of the templates; 2fz6C (-4.31), and 2pl6A (-4.10) [Figures 8,9 10]. This signifies the quality of the M4T_HYD1 model.

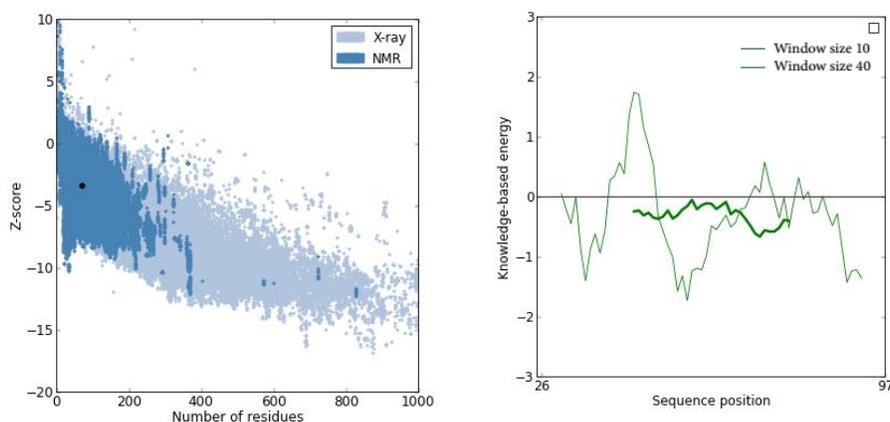


Fig. 8 ProSA overall and local quality plots for M4T_HYD1 with Z score of -3.43

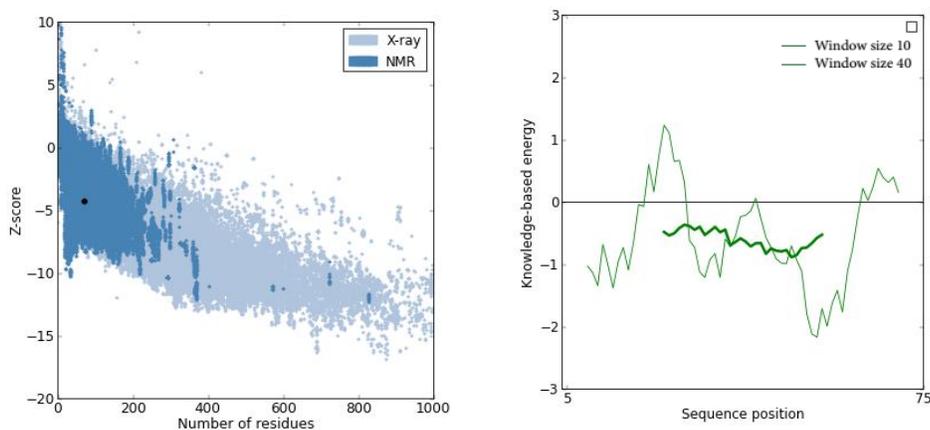


Fig. 9 ProSA overall and local quality plots for 2fz6C with Z score of -4.31

Verify3D analyzes the compatibility of an atomic model (3D) with its own amino acid sequence (1D) and hence tests the accuracy of the model. Each residue is assigned a structural class based on its location and environment. The environments are described by the area of the residue buried in the protein and inaccessible to solvent, the fraction of side chain area that is covered by polar atoms (O and N), and the local secondary structure [17]. Based on these parameters, each residue position is categorized into an environmental class.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

In this manner, a 3D protein structure is converted into a 1D string, like a sequence, which represents the environmental class of each residue in the folded protein structure. A collection of good structures is used as a reference to obtain a score for each of the 20 amino acids in this structural class. The scores of a sliding 21-residue window are added and plotted for individual residues. This method evaluates the fitness of a protein sequence in its current 3D environment. Residues with a score over 0.2 should thus be considered reliable and the sequences exhibiting lower scores are those of extracellular loops [17].

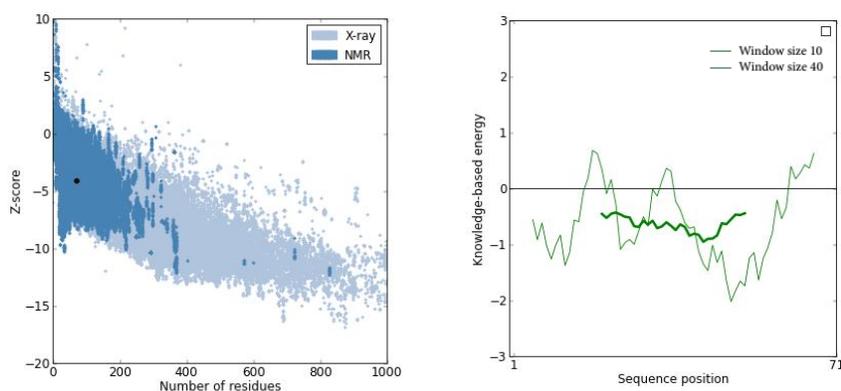


Fig. 10 ProSA overall and local (knowledge-based energy) quality plots for 2pl6A with Z score of -4.10

Evaluation of the model by Verify3D therefore revealed that 76.4 percentage of residues had an average score >0.2 (Fig. 11) which is a pass in the Verify3D assessment of protein structures [17]. Figure 11 shows the Verify3D profile for the M4T_HYD1 structure where 55 out of 72 residues are above the 0.2 mark.

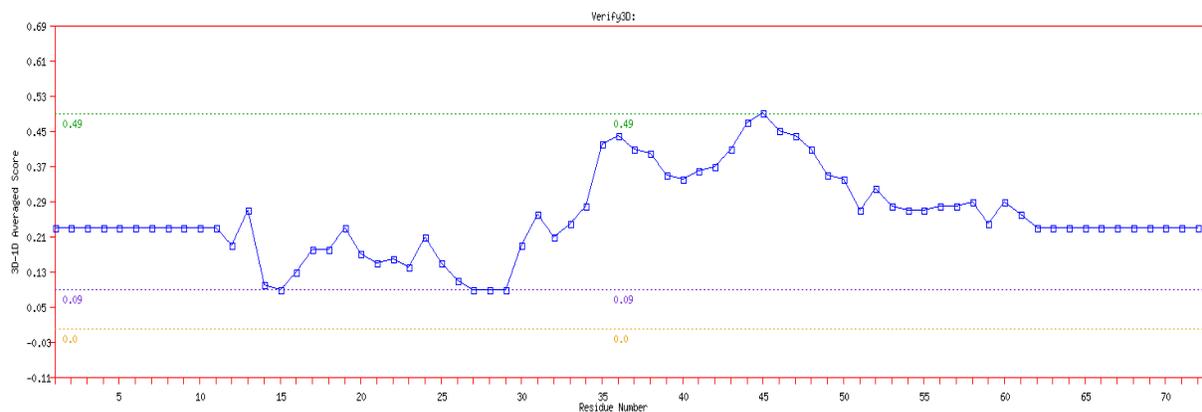


Fig. 11 Verify3D profile for model M4T_HYD1. Scores over 0.2 indicate a high quality model.

Since the overall residue score of the protein model is satisfactorily in the pass zone M4T_HYD1 is considered to have a reliable structural integrity.

When the 3dSS (3-dimensional Structural Superposition) server (which uses a STAMP algorithm) [21] was used to superimpose the model against its templates, the Root Mean Square Deviation (RMSD) between the predicted HYD1 model and templates was found to be 0.525 Å and 2.611 Å respectively (Table 1.0). Since the RMSD score reflects how close the model is to the templates, the low RMSD scores between the target and templates reflects the presence of strong homology [18].

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

Table 1.0 STAMP results obtained from the 3-Dss server

PDB ID	Chain ID	Superimposes	Sequence identity (%)	Stamp score (Max 10)	RMSD (A)	Color Scheme (Display)
M4T_HYD1.PDB	-	[Fixed Molecule]	100.00	10.000	-	Green
2FZ6	C	Yes	40.85	9.359	0.525	Orange
2P16	A	Yes	5.56	3.053	2.611	White

Table 2.0 Structural Alignment details between M4T_HYD1.PDB and 2FZ6

Fixed Molecule	Rotated Molecule	Stamp ^s Score	RMSD	Alignment Length
M4T_HYD1.PDB	2FZ6	9.359	0.525	71

Table 3.0 Rotation Matrix and Translation Vector used to calculate RMSD values and STAMP scores

Rotation Matrix			Translation Vector
0.93275	-0.24134	0.26783	3.21805
0.22587	0.97021	0.08763	-11.17414
-0.28100	-0.02125	0.95947	14.66459

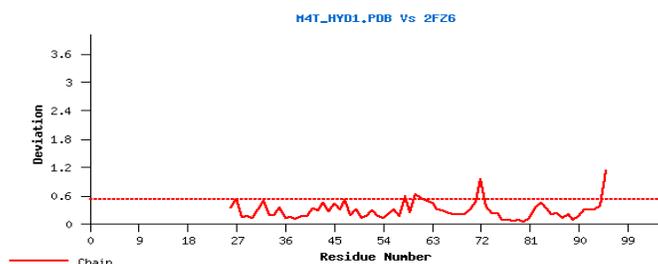


Fig. 12 Snapshot of graphical plot for superimposition between M4T_HYD1 and 2FZ6

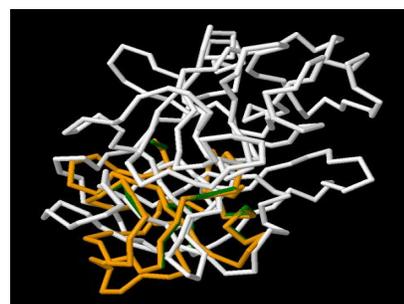


Fig. 13 Backbone structures for superimposition between M4T_HYD1 and its two templates (2fz6C, 2pl6A) as indicated by color scheme in Table 1.0.

The low deviation between the model protein and its template 2FZ6 shown in Fig. 12 (i.e., less than 1.2) proves the existence of a strong homology between the two protein structures. This is also reflected in Fig. 13 by the respective superimposed backbone structures (green and orange), thus validating the quality of the M4T_HYD1 protein structure.

V. CONCLUSION AND FUTURE WORK

The homology results showed that the M4T algorithm can produce a better 3D protein model as exemplified by the M4T_HYD1 model output whose quality was satisfactory based on Ramachandran plot analysis. Further assessment of the stereo-chemical quality of the M4T_HYD1 model using ProSA and Verify3D confirmed that the 3D structure was of high structural integrity. When compared to template structures, the ProSA and RMSD scores indicated that the constructed model was in the expected quality range and closely related to the templates, respectively. Hence, the predicted model proved to be well ratified in terms of geometry and energy contours. This shows that the model is reasonable and reliable for future molecular docking studies.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. A. Govardhan, Director, School of Information Technology (SIT), Jawaharlal Nehru Technological University Hyderabad (JNTUH) for his support and encouragement in this research.

REFERENCES

1. Wosten, H. A. B., "Hydrophobins, the fungal coat unraveled", *Biochimica et Biophysica Acta* Vol.1469, pp.79-86, 2000.
2. Wessels, J. G., "Hydrophobins: proteins that change the nature of the fungal surface", *Adv Microb Physiol* Vol.38, pp.1-45, 1997.
3. Stubner, M., Lutterschmid, G., Vogel, R. F., and Niessen L., "Heterologous expression of the hydrophobin FcHyd5p from *Fusarium culmorum* in *pastoris* and evaluation of its surface activity and contribution to gushing of carbonated beverages", *Int. J. Food Microbiol.* Vol.141, pp.110-115, 2010.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

4. Talbot, N. J., Kershaw, M. J., Wakley, G. E., De Vries, O., and Wessels, J., "MPG1 Encodes a fungal hydrophobin involved in surface interactions during infection-related development of *Magnaporthe grisea*", *Plant Cell* Vol.8, pp.985-999, 1996.
5. Linder, M. B., "Hydrophobins: proteins that self assemble at interfaces", *Curr. Opin. Colloid Interface Sci.* Vol.14, pp.356-363, 2009.
6. Wang, X., Permentier, H. P., Rink, R., Kruijtzter, J. A., Liskamp, R. M., Wösten, H. A., Poolman, B., and Robillard, G. T., "Probing the self-assembly and the accompanying structural changes of hydrophobin SC3 on a hydrophobic surface by mass spectrometry", *Biophys J* Vol.87, pp.1919-28, 2004.
7. Bayry, J., Amanianda, V., Guijarro, I., Sunde, M., and Latge, J. P., "Hydrophobins—Unique Fungal Proteins", *PLoS Pathog* Vol.8 Issue 5, e1002700, 2012.
8. Askolin, S., Linder, M., Scholtmeijer, K., Tenkanen, M., Penttilä, M., de Vocht, M. L., Wösten, H. A., "Interaction and comparison of a class I hydrophobin from *Trichophyllum commune* and class II hydrophobins from *Trichoderma reesei*", *Biomacromolecules* Vol.7, pp.1295-301, 2006.
9. Bacon, C. W., Yates, I. E., Norred, W. P., and Leslie, J. F., "Production of fusaric acid by *Fusarium* species", *Applied and Environmental Microbiology* Vol.62, pp.41, 39-41, 43, 1996.
10. Jurgenson, J. E., Zeller, K. A., and Leslie, J.F., "Expanded genetic map of *Gibberella moniliformis* (*Fusarium verticillioides*)", *Applied and Environmental Microbiology* Vol.68, pp.1 972- 1 979, 2002.
11. Leach, A. R., and Harren, J., "Structure-based Drug Discovery". Berlin, Springer, 2007.
12. Fernandez-Fuentes, N., Madrid-Aliste, J. C., Rai, K. B., Fajardo, E. J., and Fiser, A., "M4T: A Comparative Protein Structure Modeling Server", *Nucleic Acids Res.* Vol.35, pp.363-368, 2007.
13. Venclovas, C., Zemla, A., Fidelis, K., and Moulton, J., "Assessment of progress over the CASP experiments", *Proteins.* Vol.53 Suppl. 6, pp.585-595, 2003.
14. Sanchez, R., and Sali, A., "Evaluation of comparative protein structure modeling by MODELLER-3", *Proteins. Suppl.* 1, pp.50-58, 1997.
15. Contreras-Moreira, B., Fitzjohn, P. W., and Bates, P. A., "In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling", *J. Mol. Biol.* Vol.328, pp593, 2003.
16. Srinivasan, P., Perumal, C. P., and Sudha, A., "Discovery of Novel Inhibitors for Nek6 Protein through Homology Model Assisted Structure Based Virtual Screening and Molecular Docking Approaches", *The Scientific World Journal* Vol.2014, pp.1-9, 2014.
17. Shaldam, A. M., Elhamamsy, H. M., Esmat, A. E., and El-Moselhy, F. T., "1,4-Dihydropyridine Calcium Channel Blockers: Homology Modeling of the Receptor and Assessment of Structure Activity Relationship", *Medicinal Chemistry* Vol.2014, pp.1-14, 2014.
18. Anupama, P., "In Silico Studies on Development of Novel Virostatic Agents against Bluetongue Virus", *Virology* Vol.2014, pp.1-8, 2014.
19. The UniProt Consortium., "Activities at the Universal Protein Resource (UniProt)", *Nucleic Acids Res.* Vol.42, pp.191-198, 2014, <http://www.uniprot.org/uniprot/Q6FY29>.
20. http://www.ebi.ac.uk/thornton/srv/databases/cgibin/pdbsum/GetPage.pl?pdbcode=o507&pdb_type=UPLOAD&code=111606&template=procheck_summary.html, 17 April 2014.
21. Sumathi, K., Ananthalakshmi, P., Roshan, M. N. A., and Sekar, K., "3-Dimensional Structural Superposition", *Nucleic Acids Res.* Vol.34, pp.128- 134, 2006.

BIOGRAPHY



Cephas Mawere is an M.Tech Student in the field of Bioinformatics currently underway at the SIT, JNTUH, India. He attained his B. Tech. Degree in Biotechnology from CUT, Zimbabwe in 2010. He is a Harare Institute of Technology staff development research fellow. His research interests are in the area of Microbiology, Computer- Aided Drug Design, High Performance Computing and Proteomics.



Suresh Kumar G is a Lecturer in Bioinformatics at School of Information Technology (SIT), Jawaharlal Nehru Technological University Hyderabad (JNTUH), Hyderabad, India. He has received Masters Degree (M.Tech) in Bioinformatics from University of Hyderabad, Hyderabad, India. His research interests include Computer- Aided Drug Design: Ligand based, Structure based Drug Design, Molecular Modeling, Docking, 3D-QSAR studies, Genomics and Proteomics, Computational Systems Biology.



Shumbeyi Muzondo is an M.Tech Biotechnology Student at JNTUH, India. She attained her B. Sc. Degree in Applied Biology and Biochemistry from NUST, Zimbabwe in 2007. She is a Harare Institute of Technology staff development research fellow. Her research interests are in the area of Gene Transformation, Microbiology, and Computer- Aided Drug Design, High Performance Computing as well as Proteomics.