Comparative Study of Homology Based Structure Prediction and Structure Validation Tools on Some Proteins from the bhlh family.

Tanvee Pardeshi*, Akhil Jobby, Sameer Chaudhary
RASA Life Science Informatics, 301, 3rd Floor, Dhanashree Apartment, Opposite Chittaranjan Vatika, Model Colony, Shivaji Nagar, Pune - 411016, India

ABSTRACT

Comparative homology modeling has become an efficient and easy method for predicting the unknown three dimensional structure of a protein based on sequence alignment. The steps involved are template alignment, loop assignment, model building and model refinement. However, it was noticed that though the basic steps for modeling the protein were same, the results produced by different tools that are available varied; possibly due to the efficiency of the algorithm and other factors. Here, five homology modelling tools were used to compare the results for some proteins of the bhlh family. It was also noticed that structure validation tools had different results. To compare the results ProcheckRamchandran plot from PDBSum Generate and Ramchandran plot from SPDBV were used. The differing results were compared using simple statistical approach and the inference was obtained as patterns for the tools.

Copyright © 2015 International Journal for Computational Biology. http://www.ijcb.in. All rights reserved.

1. INTRODUCTION

Comparative or homology modeling performs as one of the best methods for the prediction of unknown protein structures. It helps to bridge the gap between the available sequence and structure information by providing reliable and accurate protein models. A protein template with known three dimensional structures and sufficiently high sequence identity to the target can be modeled with high accuracy models and also the proper alignment can be seen[1]. Homology models of proteins are of great interest for planning and analyzing biological experiments when no experimental three-dimensional structures are available. But, inspite of the progress in the field of NMR spectroscopy and X-ray-crystallography, 3-D structures of many therapeutically relevant proteins are still not available [2]. Many of these therapeutically relevant protein structures are required in the initial phases of drug designing and homology based modeling methods are used to analyze them. So, there is a need to predict the 3D structures of such proteins. In this study, we have considered five basic helix-loop-helix proteins to carry out the comparative homology modelling to predict their 3D structures along with comparing the software used for the same[6][8].

The basic helix-loop-helix (bHLH) proteins are a part of large superfamily of transcriptional regulators which are found in organisms from yeast to humans and function in critical developmental processes, including sex determination and the development of the nervous system and muscles.[10]Following are the proteins which we have considered in our study;

OLIG1: It is a closely related basic helix-loop-helix (bHLH) transcription factors that is expressed in myelinating oligodendrocytes and their progenitor cells in the developing central nervous system (CNS). Olig2 is necessary for specification of oligodendrocytes, but the biological functions of Olig1 during oligodendrocyte lineage development are poorly understood, but, are assumed to be repairing the brain [5].

Article Info

Received Jan 31st, 2015
Revised Apr 13th, 2015
Accepted Apr 24th, 2015

Keyword:
bhlh proteins
Homology Modeling
protein structure prediction
Structure Validation

Corresponding Author:
Tanvee Pardeshi,
Jr. Application Scientist
RASA Life Science Informatics,
Pune, India.
Email: tanvee@rasalsi.com

How to Cite:
Pardeshi et. al. Comparative Study of Homology Based Structure Prediction and Structure Validation Tools on Some Proteins from the bhlh family. JCB. 2015; Volume 4 (Issue 2): Page 21-30.

Journal homepage: http://www.ijcb.in
NPAS1: This protein is a member of the basic helix-loop-helix (bHLH)-PAS family of transcription factors. Studies of a related mouse gene suggest that it functions in the neurons but the exact function is unclear. However, it is known that it may play protective or modulatory roles during late embryogenesis and postnatal development. [5]

MYOD1: It is involved in muscle differentiation (myogenic factor) and induces fibroblasts to differentiate into myoblasts. It activates muscle-specific promoters and interacts with, or, is inhibited by twist protein. This interaction probably involves the basic domains of both proteins. [5]

NPAS3: This protein is also a member of the basic helix-loop-helix and PAS domain-containing family of transcription factors. The protein is localized to the nucleus and may regulate genes involved in neurogenesis. Chromosomal abnormalities that affect the coding potential of this gene are associated with schizophrenia and mental retardation. [5]

NPAS4: It is a member of the basic helix-loop-helix-PER (bHLH-PAS) class of transcriptional regulators, which are involved in a wide range of physiologic and developmental events. [5]

Experimental techniques such as NMR and X-ray Crystallography are employed to predict the 3D structure of proteins. These techniques are very tedious and prolonged, not always succeeding in determining the structure of proteins. Therefore the homology modeling approach is considered to be a good choice of method for predicting the structure of protein for further functional analysis study. Another advantage of homology modeling approach is that it is very quick method for predicting the structure of protein. Homology modeling of a protein is a procedure in which constructing an atomic resolution model of a protein from its amino acid sequences ("target protein") using an experimental 3D structure of a related homologous protein ("template") is carried out. Homology modeling is based on the concept that the protein sequences among homologues and identity above 30% in sequence is likely to give the similar structure. Homology modeling consists of 5 steps [13]

1. Fold assignment in which the similarity between the target sequence of interest and at least of one known protein structure (the template) is identified.
2. Target sequence and template sequence alignment.
3. The modeling done is based on the selected template.
4. Optimization of the model is performed.
5. Validation of the model

As the 3D structure of protein of bhlh family have yet not been build and in order to study the functional properties of protein more confidently, the objective of this work is to build the model of this protein using comparative modeling approach. It will provide insight into its structure and further some functional aspect. In this study we have analyzed and reviewed current approaches to carry out comparative homology modeling. As a case study, five tools for homology modelling and structure prediction were used to demonstrate the methodology and their structure prediction, but concentrating how the results from the tools differ.

2. SOFTWARE USED

For a given protein sequence, the comparative modelling procedure requires identification of homologous sequences with known structures, alignment of the query sequence to the selected template structure, 3D model construction, and refinement of the predicted model. The actual modeling process is of course much more complex, and the methods employed by various prediction servers to identify suitable templates and structures may widely differ. Following are the tools which we have utilised in our study. [2][7]

Phyre2

Protein Homology/analogy Recognition engine 2 (PHYRE2) is a free online homology modelling server. Phyre2 uses the alignment of hidden Markov models via HHsearch to significantly improve accuracy of alignment and detection rate. Phyre2 also incorporates a new ab-initio folding simulation called 'Poing' to model those regions of proteins in question which have no detectable homology to known structures. Poing is also used to combine multiple templates. Distance constraints from individual models are treated as linear elastic springs. Poing then synthesizes entire protein in the presence of these springs and at the same time models unconstrained regions using its physics simulation. [9]

MODELLER

It is a freely available offline tool used to model proteins. MODELLER builds a 3D structure that satisfies certain spatial restraints, including Ca-Ca bond length, main-chain and side-chain dihedral angles, and Van der Waals interactions. These restraints are expressed as probability density functions representing the probability of occurrence of a certain conformation and are calculated from the structures of homologous template sequences. Compilation of the probability density functions from individual template structures into a
single representative probability density function, followed by optimization of this function, then describes the overall spatial restraints for the query protein. As a result, the predicted model represents the most probable structure of the protein in question, based on structural information derived from known homologues. [12]

ESyPred 3D

ESyPred3D is an online tool which can be used to as an automated homology modelling program. This tool uses the advantage of increased alignment performances of a new alignment strategy. ESyPred3D uses the alignment obtained by weighing, combining and screening the results of various alignment programs. The final 3D structure of protein is modelled using the modelling program Modeller. [4]

Geno3D

Geno3D is a free and automatic web server for protein molecular modelling. For a query protein sequence, the homology modelling is performed by the server in six successive steps which are (i) identifying homologous proteins with known 3D structures using PSI-BLAST; (ii) providing the user all potential templates for target selection; (iii) performing the alignment of both query and subject sequences; (iv) extracting geometrical restraints (dihedral angles and distances) for corresponding atoms between the query and the template; (v) performing the 3D construction of the protein using a distance geometry approach and (vi) sending the results by e-mail to the user.[3]

3D-Jigsaw

This server again is a freely available online server which is used to build 3D models of proteins based on homologues of the known protein structures using Hidden Markov Model (HMM). The returned alignments are used to build the models. All models are preselected using populus energy. Gaps and missing residues are closed and filled using populus repair. Finally all models are recombined using the basic populus approach. In the populus approach, genetic algorithm (GA) is used as conformational space search engine. GAs is powerful search and optimization techniques, and has been used previously in a number of protein modelling efforts ranging from ab initio folding to model-building by homology. This method uses multiple initial models from a variety of modelling approaches. After 10 rounds or conversion; all models are ranked using a fine and coarse energy function weighted according to the highest SEQID found. The top five models are returned. [11]

3. METHODOLOGY

3.1 Common Feature for Evaluation

The templates that were used to model the proteins mentioned, were kept common for each tool or server used for insilico modelling. Based on the alignment forming capacity of each server or tool, a template that was common to all the selected tools or servers was chosen for each of the five proteins. The PDB templates chosen were 3rty_a, 1mdy_a, 4f3l_b, 3rty_a, 1nkp_b for npas4, myod1, npas3, npas1, olig1 (each belonging to b-hl class) respectively and had ~18-22% identity with ~100% query coverage. Ramachandran plots from SPDB Viewer and PDBSum Generate (procheck) were used to evaluate the validation of protein structures.

3.2 Homology based Modelling and Energy Minimization

The npas4, myod1, npas3, npas1, olig1 protein sequences were retrieved from NCBI-Protein database and sequence alignment followed by modelling was performed on each server (Jigsaw3D v3-Interactive, Easypred3D, Geno3D, Modeller v9.11, Phyre2-Normal) and it was found that the models obtained varied in appearance (visualized in Rasmol and SPDB Viewer). The physical parameters such as number of residues, bonds, helices, and sheets etc. that were modelled also varied. Hence, the differences were noted down in a tabular format as shown in Table 1 and the energy minimization values were computed by Swiss PDB Viewer.

| Proteins | Modelling Tool | No of Residues | No of Atoms | No of Bonds | Helices | Sheets | Glysine | Proline | Energy Minimization Value |
|----------|----------------|----------------|-------------|-------------|---------|--------|---------|---------|---------------------------|
| NPAS4    | 3DJigaw        | 54             | 149         | 246         | 164     | 74     | 16      | 14      | -2373.55                  |
|          | EasyPred3D     | 155            | 172         | 179         | 175     | 54     | 53      | 98      | -2753.28                  |
|          | Geno3D         | 86             | 677         | 688         | 102     | 61     | 16      | 21      | -3401.53                  |
|          | Modeller9.11   | 234            | 1797        | 1839        | 406     | 74     | 77      | 119     | -9084.46                  |
|          | Phyre2         | 152            | 1173        | 1219        | 232     | 434    | 48      | 98      | -3785.14                  |
| MYOD1    | 3DJigaw        | 95             | 750         | 763         | 388     | 0      | 16      | 28      | -5154.21                  |
|          | EasyPred3D     | 0              | 0           | 0           | 0       | 0      | 0       | 0       | 0                         |
|          | Geno3D         | 65             | 489         | 492         | 342     | 0      | 4       | 14      | -4011.85                  |
|          | Modeller9.11   | 320            | 2420        | 2483        | 419     | 51     | 80      | 259     | -8222.58                  |

Table 1. Fluctuating values produced by different modelling tools
3.3 Structure Validation

The structures of the modelled proteins belonging to the basic helix-loop-helix (npas4, myod1, npas3, npas1, olig1) were checked for validity using Ramachandran plots from SPDB Viewer and PDBSum Generate. Here, SPDB Viewer has its Ramachandran plot whereas PDBSum Generate produces Procheck-Ramachandran Plot; hence, the results that were obtained differed. The varying results obtained were noted in a tabular format (Table 2 and Table3 respectively).

Table 2. Values produced from SPDB Viewer Ramachandran Plot

| Proteins | Modelling Tool | Most Favoured Region (%)| Allowed Region | Disallowed Region |
|----------|----------------|-------------------------|----------------|------------------|
|          |                | SPDBViewer Ramachandran Plot |                | SPDBViewer Ramachandran Plot |
|          | 3DJigaw       | 38                      | 8             | 17.02            | 1                | 2.12             |
|          | EasyPred3D    | 118                     | 85            | 13               | 9.35             | 8                | 5.7              |
|          | geno3D        | 53                      | 68.8          | 19               | 24.6             | 5                | 6.4              |
|          | Modeller9.11  | 191                     | 88.8          | 19               | 8.8              | 5                | 2.3              |
|          | Phyre2        | 117                     | 84.7          | 17               | 12.31            | 4                | 2.9              |
|          | 3DJigaw       | 65                      | 74            | 18               | 20               | 4                | 4.5              |
|          | EasyPred3D*   | 0                       | 0             | 0                | 0                | 0                | 0                |
|          | geno3D        | 58                      | 93.5          | 4                | 6.4              | 0                | 0                |
|          | Modeller9.11  | 180                     | 92.3          | 11               | 5.6              | 4                | 2.05             |
|          | Phyre2        | 51                      | 78.4          | 13               | 20               | 1                | 1.5              |
|          | 3DJigaw*      | 0                       | 0             | 0                | 0                | 0                | 0                |
|          | EasyPred3D    | 288                     | 91.7          | 19               | 6.05             | 7                | 2.2              |
|          | geno3D        | 276                     | 72.8          | 63               | 16.6             | 40               | 10.5             |
|          | Modeller9.11  | 751                     | 80.6          | 113              | 12.13            | 67               | 7.19             |
|          | Phyre2        | 293                     | 95.7          | 10               | 3.2              | 3                | 0.9              |
|          | 3DJigaw       | 209                     | 83.9          | 24               | 9.6              | 16               | 6.4              |
|          | EasyPred3D    | 279                     | 89.71         | 26               | 8.36             | 6                | 1.9              |
|          | geno3D        | 184                     | 70.2          | 59               | 22.5             | 19               | 7.2              |
|          | Modeller9.11  | 494                     | 84.01         | 55               | 9.3              | 39               | 6.6              |
|          | Phyre2        | 204                     | 89.08         | 18               | 7.8              | 7                | 3.05             |
|          | 3DJigaw       | 73                      | 90.1          | 4                | 4.9              | 4                | 4.9              |
|          | EasyPred3D    | 66                      | 92.9          | 3                | 4.22             | 2                | 2.8              |
|          | geno3D*       | 0                       | 0             | 0                | 0                | 0                | 0                |
|          | Modeller9.11  | 245                     | 91.07         | 18               | 6.6              | 6                | 2.2              |
|          | Phyre2        | 64                      | 96.96         | 2                | 3.03             | 0                | 0                |

* No model was obtained as common template was not found.
4. RESULTS AND DISCUSSION

4.1 Data Interpretation of Table 1

Table 1 displays the fluctuating structural attributes from different tools used. Standard deviations (σ) of each protein from the different tools were obtained to understand variation from the average of Most Favoured Regions (MFR) from PDBSum Generate Procheck-Ramchandran Plot. Similarly, to analyse if there is pattern similarity, standard deviation (σ) of Number of Residues modelled for each protein by the tools were also obtained (Table 4).

| Proteins | Modelling Tool | Most Favoured Region (PDBSum Generate) | MFR (%) | No of Residues | σ (No of Residues) |
|----------|----------------|----------------------------------------|---------|----------------|--------------------|
| NPAS4    | 3DJigaw        | 39                                     | 52.18429| 54             | 69.72231           |
|          | EasyPred3D     | 105                                    | 155     |                |                    |
|          | Geno3D         | 49                                     | 86      |                |                    |
|          | Modeller9.11   | 169                                    | 234     |                |                    |
|          | Phyre2         | 104                                    | 152     |                |                    |
| MYOD1    | 3DJigaw        | 61                                     | 61.75961| 95             | 122.7436           |
|          | EasyPred3D*    | 0                                      | 0       |                | 2.8                |
|          | Geno3D         | 61                                     | 65      |                |                    |
|          | Modeller9.11   | 180                                    | 320     |                |                    |
|          | Phyre2         | 49                                     | 68      |                |                    |
|          | 3DJigaw*       | 0                                      | 0       |                | 338.3825           |
| NPAS3    | EasyPred3D     | 258                                    | 320     |                |                    |
|          | Geno3D         | 227                                    | 381     |                |                    |
|          | Modeller9.11   | 595                                    | 933     |                |                    |

* No model was obtained as common template was not found.
The standard deviations obtained were plotted on a graph (Graph 1 and Graph 2) and a pattern was observed. This shows that there is constant variation in the models produced by the tools (Jigsaw3D v3-Interactive, EasyPred3D, Geno3D, Modeller v9.11, Phyre2-Normal)

Graph 1. Standard deviations of MFR of models from each tool for the proteins

Graph 2. Standard deviations of number of residues of models from each tool for the tool
The models obtained were checked for structure validation using Ramachandran Plots from SPDB Viewer and PDBSum Generate and it was observed that the results from Plots also varied. To understand the variations in the results for structure validation tools, standard deviation of the MFR produced by SPDB Viewer-Ramachandran Plot and PDBSum Generate-Ramachandran Plot were obtained and graphs were plotted (Table 5, Graph 3).

Table 5. Standard deviations (σ) of MFR from SPDBV and PDBSum Generate Plots

| Proteins | Modelling Tool | Most Favoured Region (SPDB Viewer-Ramchandran Plot) | σ | Most Favoured Region (PDBSum Generate-Ramchandran Plot) | σ |
|----------|----------------|----------------------------------------------------|----|-------------------------------------------------------|----|
| NPAS4    | 3DJigaw        | 38                                                 | 7.704024922 | 38                                                    | 79 |
|          | EasyPred3D     | 118                                                | 105 |                                                        | 9.262397098 |
|          | geno3D         | 53                                                 | 49  |                                                        | 9.262397098 |
|          | Modeller9.11   | 191                                                | 169 |                                                        | 104 |
|          | Phyre2         | 117                                                | 104 |                                                        | 104 |
|          | 3DJigaw        | 65                                                 | 61  |                                                        | 61  |
|          | EasyPred3D*    | 0                                                  | 0   |                                                        | 0   |
|          | geno3D         | 58                                                 | 54  |                                                        | 10.14379942 |
| MYOD1    | Modeller9.11   | 180                                                | 180 |                                                        | 180 |
|          | Phyre2         | 51                                                 | 49  |                                                        | 49  |
|          | 3DJigaw*       | 0                                                  | 0   |                                                        | 0   |
|          | EasyPred3D*    | 288                                                | 258 |                                                        | 258 |
|          | geno3D         | 276                                                | 227 |                                                        | 227 |
|          | Modeller9.11   | 751                                                | 595 |                                                        | 595 |
|          | Phyre2         | 293                                                | 258 |                                                        | 258 |
|          | 3DJigaw        | 209                                                | 174 |                                                        | 174 |
|          | EasyPred3D*    | 184                                                | 147 |                                                        | 147 |
|          | geno3D         | 279                                                | 227 |                                                        | 227 |
|          | Modeller9.11   | 494                                                | 360 |                                                        | 360 |
|          | Phyre2         | 204                                                | 162 |                                                        | 162 |
|          | 3DJigaw        | 73                                                 | 56  |                                                        | 56  |
|          | EasyPred3D*    | 66                                                 | 60  |                                                        | 60  |
|          | geno3D*        | 0                                                  | 0   |                                                        | 0   |
|          | Modeller9.11   | 245                                                | 189 |                                                        | 189 |
|          | Phyre2         | 64                                                 | 54  |                                                        | 54  |

*Note:* No model was obtained as common template was not found. For such cases the values are considered to be zero and were not calculated while computing Standard Deviations.

For better understanding, the standard deviations were plotted on a graph and the difference was observed.

Graph 3. Standard deviations of MFR from Ramachandran Plots of SPDB Viewer and PDBSum Generate.

Hence using these tools homology model of these five bhlh proteins was generated. Using the swiss model software the models of these five proteins generated from the five software were superimposed on each other and a single image was obtained from it. Fig 1 illustrates the same.
Fig 1: This figure shows the homology models generated for MYOD1, NPAS1, NPAS3, NPAS4 and OLIG1.

From the tables (Table 2 and Table 3) and Graph 3, it can be analysed that though the number of residues were the same for the Ramchandran Plots, both 'SPDB Viewer' and 'PDBSum Generate' produced varying results. It is evident from the graph that the values obtained for MFR (Most Favoured Region) had more residues from SPDB Viewer when compared to MFR from PDBSum Generate as the Standard deviation of MFR-PDBSum Generate was more. Also, Ramchandran plots from SPDBV had Most Favoured Region, Allowed Regions and Disallowed Regions whereas that from PDBSum Generate had Most Favoured Regions, Additionally Allowed Regions, Generously Allowed Region and Disallowed Regions. Here, 'Additionally Allowed' and 'Generously Allowed Regions' were combined as 'Allowed Region' for convenience.

5. CONCLUSION
High-resolution crystallographic structures becoming available for all human and pathogen proteins and complexes are a very unlikely scenario. The only practical manner of exploring ligand–protein interactions for most systems is to use comparative protein structure models. It is now clear that comparative models, based on as little as 30% sequence identity to known template structures, can be useful. It can be concluded that different homology modelling tools (Jigsaw3D v3-Interactive, EasyPred3D, Geno3D, Modeller v9.11, Phyre2-Normal) produced varying results although common templates were used and the difference was observed to be varying in terms of the number of residues modelled. It was also observed that Ramchandran plots from structure validation tools (SPDB Viewer and PDBSum Generate) also varied. It was inferred that Ramchandran plot from SPDB Viewer produced more residues in Most Favoured Regions than Procheck Ramchandran plot from PDBSum Generate. However, to establish a much clearer picture, more proteins need to be considered for the study and further validation has to be performed.

REFERENCES
[1] Akbar Nayeem, Doree Sitkoff, Stanley Krystek Jr., A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models. Protein Science, 2006; 15: 808-824.
[2] András Fiser, Richard Kinngian Do, Andrej S Ali (2000). Modeling of loops in protein structures. Protein Science, 9, 1753-1773.
[3] Christophe Combet, Martin Jambon, Gilbert Delage, Christophe Gourjun, Geno3D: automatic comparative molecular modelling of proteins, Protein Science, 2001; 18(1): 213-214
[4] Christophe Lambert, Nadia Leonard, Xavier De Bolle, Eric Depiereux, eSyPred 3D: Prediction of proteins 3D structures. Bioinformatics, 2002; 18(9): 1250-1256.
[5] Donna Maglott, Jim Ostell, Kim D. Pruitt, Tatiana Tatusova, Entrez Gene: gene-centered information at NCBI. Nucleic Acid Research, 2005; 33(1): 54-58.
[6] E. Michalsky, A. Goede and R. Preissner, Loops In Proteins (LIP) A comprehensive loop database for homology modelling. Protein engineering, 2003; 16(2): 979-985.
[7] Fischer, D. Servers for protein structure prediction, Curr Opin Struc Biol, 2006; 16: 178-82.
[8] Konstantin Arnold, Lorenza Bordol, Jurgen Kopp, Torsten Schwede, The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics, 2005; 22(2): 195-201.
[9] Lawrence A Kelley & Michael J E Sternber, Protein structure prediction on the Web: a case study using the Phyre server, Nature, 2009; 4(3): 364-371.
[10] Mark Eben Massari & Cornelis Murre, Helix-Loop-Helix Proteins: Regulators of Transcription in Eucaryotic Organisms. Molecular and Cellular Biology, 2000; 20(2), 429-440
[11] Paul A. Bates, Lawrence A. Kelley, Robert M. MacCallum, Michael J.E. Sternberg, Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Protein Science, 2002; 45(2), 39-46.
[12] Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., Karplus, M., Evaluation of comparative protein modeling by MODELLER. Proteins, 1995; 23: 318-26.
[13] Singh Ameesha & Chaube Radha, Bioinformatic Analysis, Structure Modeling and Active Site Prediction of Aquaporin Protein from Catfish Heteropneustes fossils. International Journal on Recent and Innovation Trends in Computing and Communication, 2014; 2(10): 3208-3215.