ISSN No:-2456-2165

# Structural Homology Modeling of C-Terminal Domain of the Dystrophin Protein: An in-Silico Approach

Akanksha Mishra[1*]; Pramod Sairkar[1]; Nipun Silawat[1]; Mohd. Maruf Khan[2] and Anil Kothari[1]
[1] Centre of Excellence in Biotechnology, Madhya Pradesh Council of Science and Technology, Vigyan Bhawan, Nehru Nagar, Bhopal-462003 (M.P.)
[2] H.S.Z.H. Government Autonomous Unani Medical College Bhopal-462003 (M.P.)

Corresponding Author: Akanksha Mishra[1*] (ORCID: 0009-0006-0345-0041)

**Abstract:- Dystrophin is one of the most significant and well-researched cytoskeletal proteins that is prominently expressed in skeletal and cardiac muscles. It is a large 400-kD protein, which is encoded by the largest gene in the human body- DMD gene. A significant decrease in dystrophin levels in muscles results in a gradual and severe skeletal muscular weakening. Lack of dystrophin results in muscular dystrophies such as DMD (Duchenne muscular dystrophy) and BMD (Becker muscular dystrophy. Understanding the dystrophin protein's structure is crucial for developing a cure for the disease. Comprehensive knowledge of protein conformation can offer essential insights for protein engineering and medication development. Currently complete structural information about dystrophin protein is not available, only the structure of N-terminal domain and spectrin repeats of central rod domain has been prepared. Our study aims to determine the structure of the C-terminal domain using the structural modelling software (Robetta and Phyre2) and perform the validation of the most accurate structure using the SAVESv6.0 software. The result was concluded on the following basis – 1. PROCHECK (Ramachandran plot) analysis - Robetta predicted model has 90.8% residues in the most favored region and only 0.7% residues in the disallowed region which makes it a good quality model, compared to the phyre2 predicted model where only 69% residues were in the most favored region with 7.0% residues in the disallowed region. 2. ERRAT analysis- Robetta predicted structure is the most accurate with 88.316% as the overall model quality which is more than the phyre2 model quality (64.127%). The study validates Robetta predicted C- terminal domain model as the most stable C-terminal structure of Dystrophin protein.**

***Keywords:-*** *DMD Gene; Dystrophin Protein; Phyre2; Ramachandran Plot; Robetta.*

## I. INTRODUCTION

Dystrophin, a rod-shaped cytoplasmic protein is the vital part of the protein complex DGC, located primarily in the skeletal and cardiac muscles. Neurons in the brain contain trace quantities of dystrophin. The protein is made up of 3684 amino acids with a calculated molecular weight of 427kDa. It is predominantly hydrophilic throughout its entire length with 31% of the amino acids being charged (i.e. Arg, Asp, Glu, His and Lys). According to a "Chou and Fasman" prediction of secondary structure, over much of the sequence there is a very high potential for an alpha-helical development. Dystrophin consists of four domains:

➢ *Actin-Binding Amino Terminal Domain (14-240aa)-*

Two calponin homology domains (CH1 and CH2) are present in the actin-binding amino terminal domain. By directly binding to F actin, this typical CH-actin binding domain connects dystrophin to the subsarcolemmal actin network. In addition, dystrophin and the dystrophin complex function as broader cytoskeletal integrators, which are essential for the stability of muscle membranes.

- *Central Rod Domain (253-3040 Aa)-*

The 24 spectrin repeats are present in the central rod domain of dystrophin. Additionally, the rod domain is home to a second actin-binding motif (ABD2), which spans a distinct set of spectrin repeats that are rich in basic amino acids. This suggests that the interaction with acidic actin filaments is mediated by an electrostatic potential [1]. ABD2 and ABD1 work together to create a robust lateral interaction with actin filaments, where ABD2 lies close near the middle of the rod [2]. Hinges (Four short proline-rich spacers) interrupts the 24 spectrin repeats. Hinge 4, which is situated near the end of the rod domain, contains the WW domain, that is involved in protein-protein interactions. The carboxy-terminus of β-dystroglycan is bound by the WW domain and two nearby EF-hands, which anchors the dystrophin at the sarcolemma.

- *Cysteine-Rich Domain (3080-3360 Aa)-*

The two EF-hands in dystrophin reside in the cysteine-rich domain, which lies between the C-terminus and the central rod. The zinc finger (ZZ) domain contains conserved cysteine residues which folds to produce domain structure in the presence of divalent metal cations like $Zn^{2+}$. It is another domain found in the cysteine rich domain [3]. The ZZ domain of dystrophin binds calcium-dependently to calmodulin.

- *Carboxy Terminal Domain – (3361-3685)*

Two polypeptides in the carboxy-terminal (CT) domain fold into α-helical coiled coils resembling the rod domain's spectrin repeats. Common protein motifs involved in protein-protein interaction include coiled coils.

ISSN No:-2456-2165

Table 1 Dystrophin Domain Description

| Domain Description | Amino acid | Exons |
|---|---|---|
| 1. Actin binding domain | 14-240 | 2-8 |
| 2. Central rod domain | 253-3040 | 9-61 |
| 3. Cysteine-rich domain | 3080-3360 | 62-69 |
| 4. Carboxy-terminal domain | 3361-3685 | 70-79 |

When muscles contract and relax, the protein complex and dystrophin protein work together to fortify and shield the muscle fibers from damage. Skeletal and cardiac muscle cells experience extreme strain during contraction; dystrophin is essential for maintaining the stability of the sarcolemma, the muscle membrane, during this process. Muscle-related disorders result from dystrophin deficiency, which compromises the integrity of the sarcolemmal membrane. Numerous illnesses range in severity from mild muscular atrophy to total muscle degeneration and death.

DMD, or Duchenne muscular dystrophy, is one of the most prevalent muscle illnesses [4]. Dystrophin loss is the cause of DMD, a fatal illness that affects the skeletal and heart muscles [5,6]. The less severe kind of DMD is called Becker muscular dystrophy (BMD). The symptoms of Duchenne Muscular Dystrophy (DMD) includes skeletal muscular atrophy and weakening. It is an X-linked, recessive, monogenic disorder. This progressive, life-shortening disease represents one of the most common genetic disorders affecting children: about l/3rd is the estimated world-wide incidence i.e. 300 male births annually [7]. The DMD gene mutation that causes this condition prevents any virtually or substantially functional dystrophin protein from being produced.

DMD primarily affects men and develops in early childhood. One birth per 3,500 live male births is the estimated birth prevalence. Usually, it starts when a child is three to five years old. Although DMD frequently runs in families, it is also possible for a male to develop the illness unexpectedly without any family history of the condition. There are a few potential explanations for this. The first is a genetic mutation, wherein the family's female members may have had DMD for generations without their knowledge or males of the previous generation were unaware of it, even if they had been affected. A new mutation that occurred in the mother's egg cell and caused the child to have DMD could be the second explanation. The X chromosome is inherited by the daughter, thus although a man cannot pass on the disease to his son, it will undoubtedly pass to his daughter.

Currently for DMD or BMD, there isn't an effective cure at the moment. However, a number of treatments, such as exon skipping molecules, protein replacement therapy, stop codon read through chemicals, and gene therapy techniques, are currently being investigated in clinical trials or in animal models of disease [8,9]. Large animal models of DMD have shown remarkable success with gene therapy; however, a clinical trial with DMD patients was unable to proceed because of immunological reactions [8,10].

To cure the disease, it is very important to have the solved structure of the protein as it gives us a deeper comprehension of the functioning of proteins, enabling us to formulate hypotheses regarding their modification and control. Currently only N-terminal actin binding domain 3D structure has been predicted out of the four domains, and some of the spectrin repeats has been resolved through which the central rod domain can be studied, but it is important to have knowledge of the complete structure. So our study aims *in-silico* structural modelling of C terminus domain of the dystrophin protein. The study predicted the structure in two different software ROBETTA and PHYRE2 and performed validation of the most accurate one in the SAVESv6.0 software.

## II. MATERIALS AND METHOD

- Target Protein Retrieval Using Uniprot Database

The sequence of target protein dystrophin is retrieved from the Uniprot database in the FASTA format.

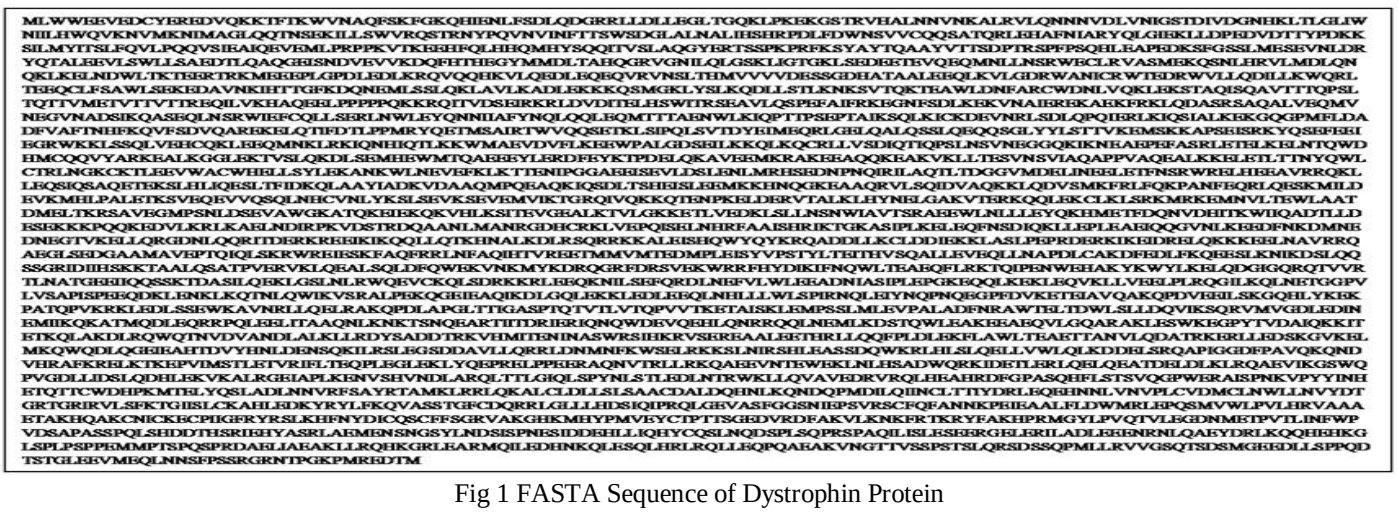

Fig 1 FASTA Sequence of Dystrophin Protein

➢ *In-Silico Structure Prediction Using Modelling Software* .

Structure prediction was performed using three different modelling software- Swiss-Model, Robetta and Phyre2 software.

- *Modelling Using Swiss Model* –

The C-terminal domain sequence was placed in Swiss-Model for modelling. Swiss-Model is an automated protein structure homology modelling server. It searches for library of experimental structures with suitable templates for the target protein. On the basis of a sequence alignment between the target protein and the template structure, a three-dimensional model for the target protein is generated. But it was seen that Dystrophin C terminal domain does not have any relative homologs present therefore Swiss model was unable to generate the 3D structure of the protein. Then we approached the *ab-initio* method where the distinct homologs were searched.

- *Modelling Using Robetta* –

In order to create a high-quality model that includes every residue in a provided sequence, it blends template-based (homology modeling) and de novo structure prediction techniques. The steps that were taken in Robetta were as follows. First, the query sequence was screened using BLAST and PSI-BLAST for regions that have been homologously characterized in an experiment. Then, putative domains were segmented into sequences based on matches to existing families and structures, multiple sequence information, and predicted secondary structure information. The portions of the query that are linked to any detected parents are saved and included to the template-based modeling methodology. After that, the remaining lengthy unassigned areas are divided into sizes that the Rosetta de novo procedure can model. Each potential domain then adheres to its designated protocol track following domain parsing. An automated version of the CASP-4 Rosetta protocol was used to generate a large number of alternate "decoy" conformations, filter the decoy ensemble to remove conformations that are not protein-like, and then cluster the remaining structures to identify broad low free energy minima in order to model the domains from scratch. Choosing the final models from the decoy clusters or from other low energy decoys is the last stage in the de novo domain modeling protocol. Using the RoseTTA fold approach, the 3D protein structure was successfully predicted.

- *Modelling Using Phyre2* –

Phyre2 is a collection of the tools available to predict and analyze the protein structure. It is one among the most widely used servers today. Phyre2 also incorporates a new *ab initio* folding simulation called Poing to model regions of your proteins with no detectable homology to known structures. When we submit the C-terminal domain sequence on the server following steps are followed. It begins with the-:

- ✓ Gathering the homologs sequences with PSI-BLAST.
- ✓ Building the Hidden Markov Model of the sequence.
- ✓ Fold library scanning and constructing simple models
- ✓ Loop modelling.
- ✓ Multiple templated modelling with Poing.
- ✓ Generating the result page.

The 3D protein structure is modelled using the intensive method, *ab initio* method which gives us the information about the confidence and percentage identity.

➢ *Evaluation and Model Validation of the Prepared Structure Using SAVES V6.0 Software*

The C terminal domain model which is prepared in the Robetta and Phyre2 software is downloaded in PDB format and now is evaluated in the SAVES v6.0 software using two tools PROCHECK and ERRAT.

- *Procheck-*

Input structure is uploaded then the server compares it with their PDB structure, since it has a PDB structure database. It analyses the overall model geometry and provides the stereochemical quality of a predicted model, generates the Ramachandran plot which gives information about the favored, allowed and disallowed regions. Thus on this basis we check for the structure that has better stability and validate it.

- *Eerat-*

Input structure is uploaded and the server now plots a graph of position of nine residues sliding window versus the error function. The result is then displayed in the form of overall quality score of the input structure. A overall quality score of more than 91% is considered good with resolution of 2-3 A0. Overall quality score below 85% cannot be consideration since it comes under bad score.

## III. RESULT AND DISCUSSION

➢ *Structure Modelling/ Prediction Results*

- *Predicted 3D Structure of C-Terminal Domain on Robetta-*

The 3d structure of the C-terminal domain on Robetta was build based on the RoseTTAfold method. The confidence of the predicted model is low (estimated TM score = 0.35). The estimated TM score is the based on the probability of the top predicted distance and the convergence of the top predicted models. TM score is between 0 and 1 and a score higher than 0.5 usually indicate a model with correctly predicted topology.

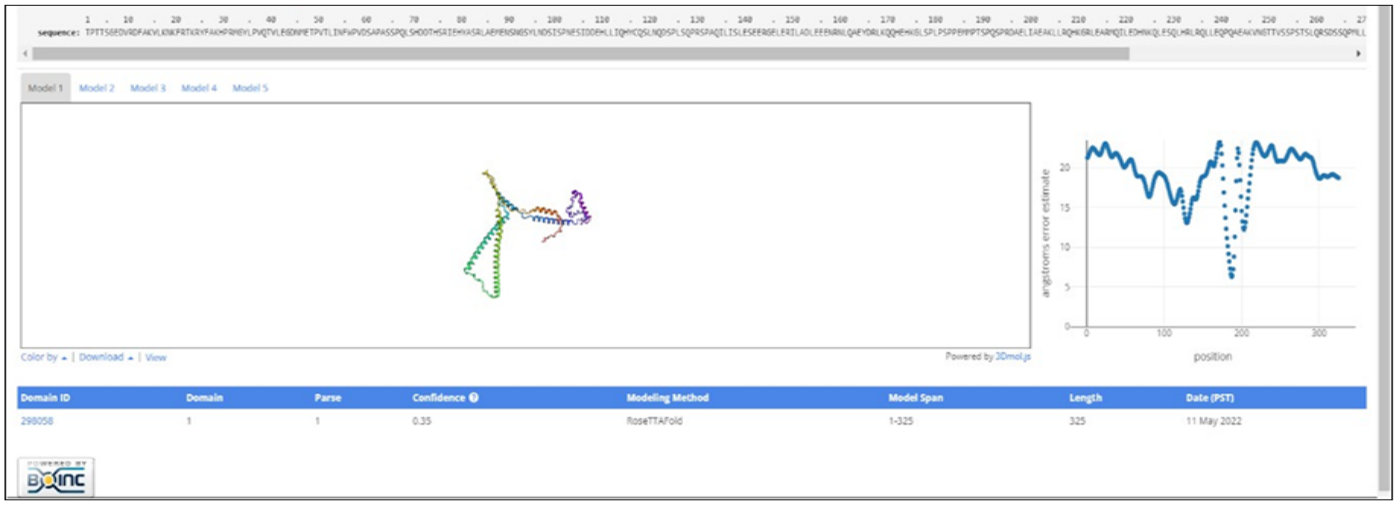


Fig 2 C Terminal Structure Modelling on Robetta

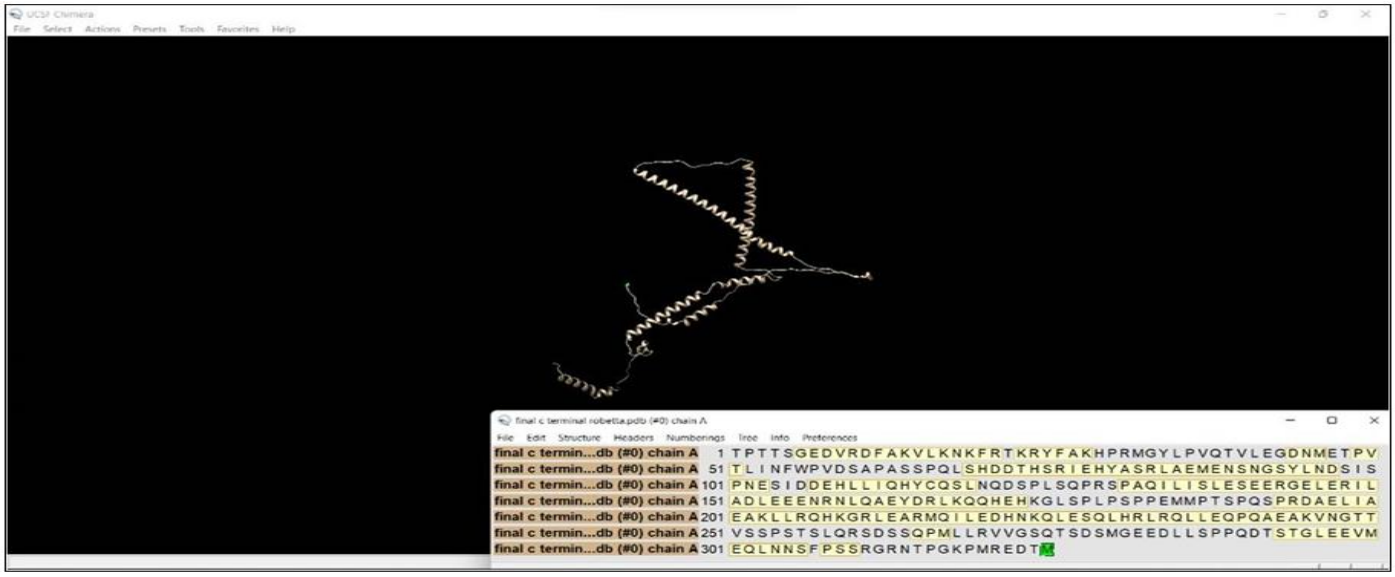


Fig 3 Visualization of the Predicted Structure on Chimera Software

➢ *Predicted 3D Structure of C-Terminal Domain on Phyre2 –*

The 3d structure of the C-terminal domain on Phyre2 was build based on the *ab-initio* method. The confidence of the predicted model is 94% concluding that 94% of residues are modelled at >90% confidence.

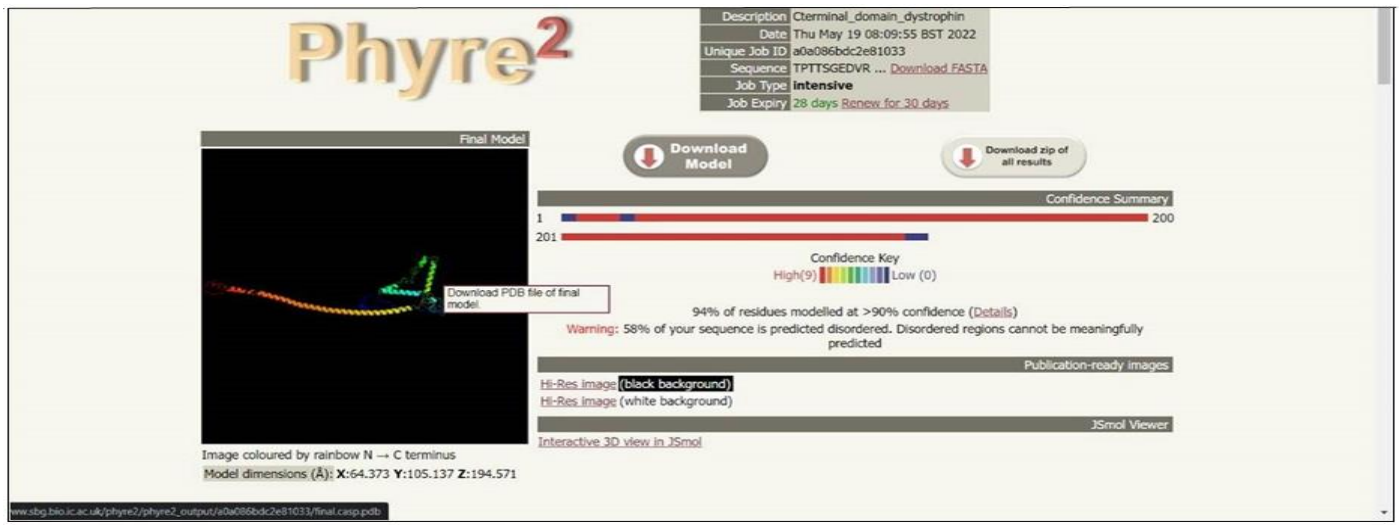


Fig 4 C Terminal Structure Modelling on Phyre2

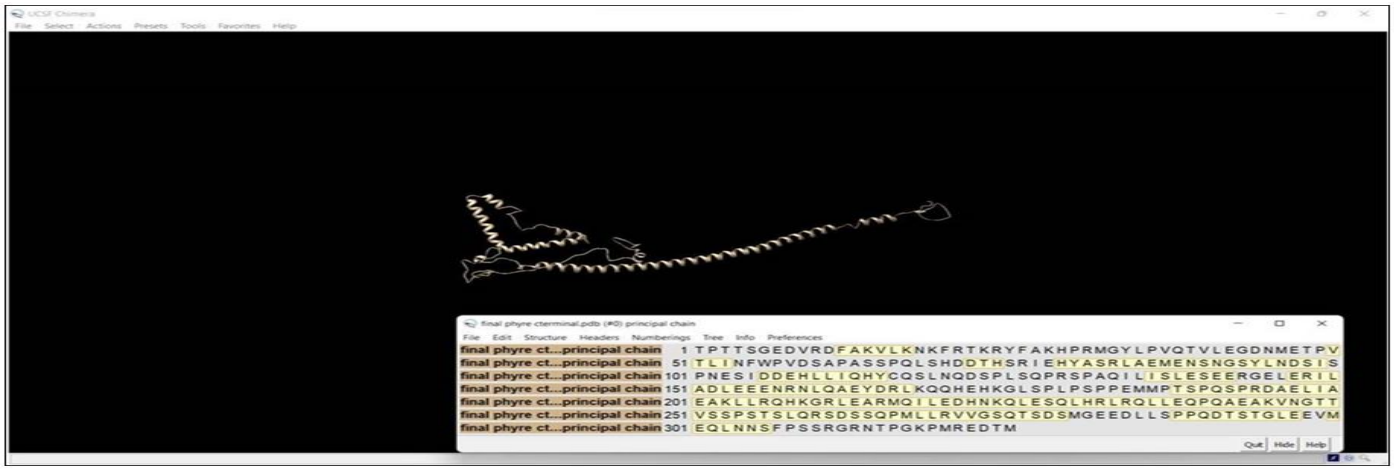


Fig 5 Visualization of the Phyre2 Predicted Structure on Chimera Software

➢ *Evaluations, Validation of the Predicted Structures on Savesv6.0*

After the model structure prediction, quality assessment and model validation is the most important step. This is done using the SAVES v6.0 software were both the structures are individually evaluated and the one with the best results is declared as the valid C- terminal domain structure. For validating the best model between the two predicted C terminal structure, 2 tools PROCHECK and ERRAT is taken into consideration.

➢ *Validation Using PROCHECK Tool*

Robetta and Phyre2 predicted sequence is individually uploaded as input file and the tool generates the Ramachandran plot. The Ramachandran plot analysis is the widely used method for model accuracy analysis and its validation. It gives result in info in form of - the most favored regions, additionally allowed regions, generously allowed regions, and disallowed regions, respectively. It is expected that no residue should reside in the disallowed or outlier region.

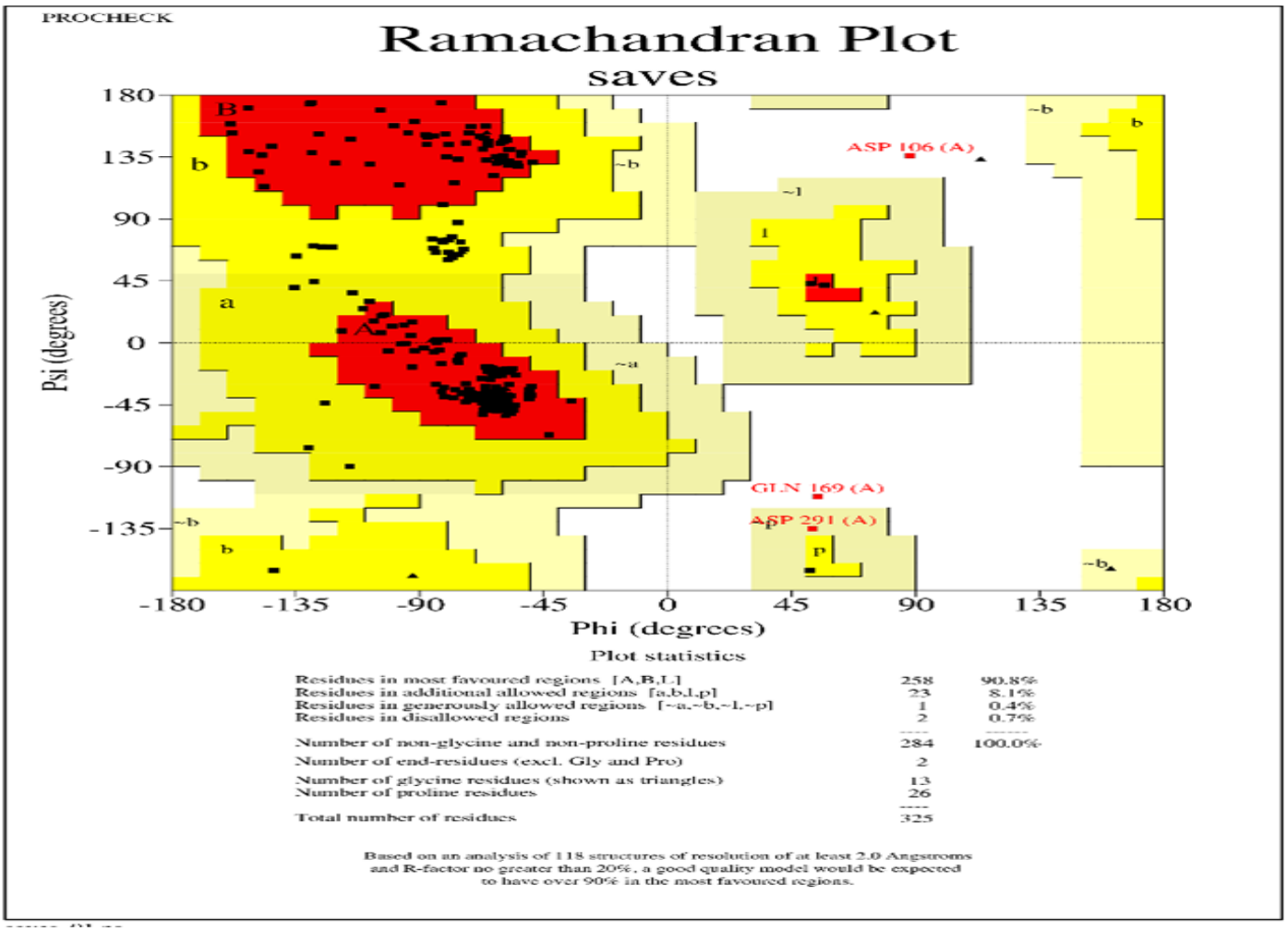


Fig 6 Robetta Predicted Model Validation

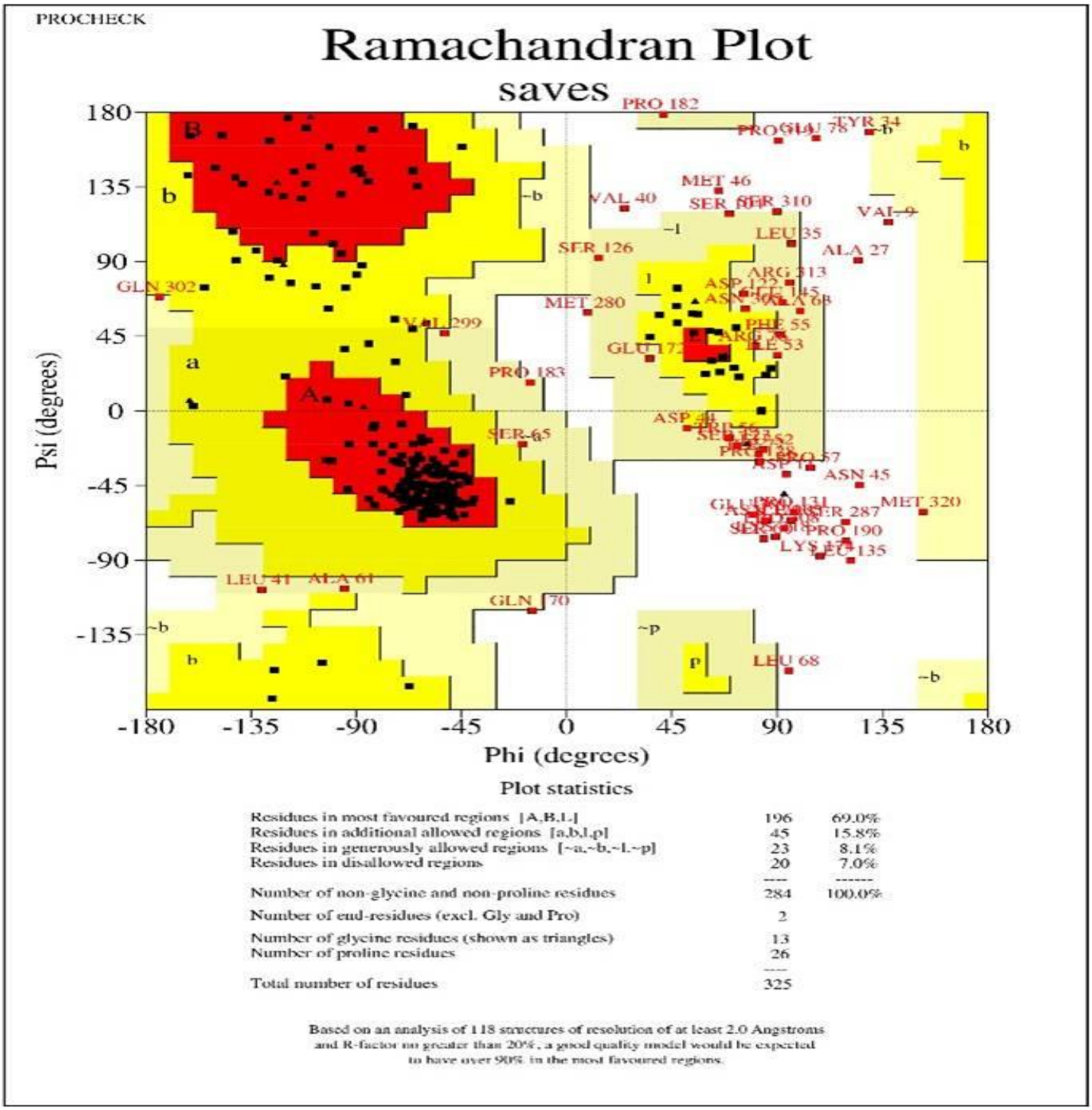


Fig 7 Phyre2 Predicted Model Validation

Table 2 Comparison of the Obtained Ramachandran Plot

| S.No. | Ramachandran plot statistics | Robetta predicted structure | | Phyre2 predicted structure | |
|---|---|---|---|---|---|
| | | Total residue | % | Total residue | % |
| 1. | Residues in most favoured region (red colour) | 258 | 90.8% | 196 | 69% |
| 2. | Residues in additional allowed region (yellow) | 23 | 8.1% | 45 | 15.8% |
| 3. | Residues in generously allowed region (cream) | 1 | 0.4% | 23 | 8.1% |
| 4. | Residues in disallowed region (white) | 2 | 0.7% | 20 | 7.0% |
| 5. | No. of non-glycine and non-proline residues | 284 | 100% | 284 | 100% |
| 6. | Number of end residues (excl. Gly and Pro) | 2 | | 2 | |
| 7. | Number of glycine residues (shown in triangle) | 13 | | 13 | |
| 8. | Number of proline residues | 26 | | 26 | |
| 9. | **Total number of residues** | **325** | | **325** | |

Based on the PROCHECK (Ramachandran plot) analysis of the two predicted C-terminal structure, it can concluded that Robetta predicted C terminal model is accurate and valid compared to Phyre2 predicted C terminal structure, since a high-quality model should contain more than 90% of residues in the most favored area. According to results, in the Robetta predicted model about 90.8% of the residues fall in the most favored region with just 0.7% residues in the disallowed region, whereas in the Phyre2 predicted model 69% residues fall in the most favored region with about 7.0% residues in the disallowed region.

➢ *Validation Using ERRAT Tool*

Robetta and Phyre2 predicted sequence is individually uploaded as input file and the tool generates the graph between the residue and the error value. The problematic portions of the structure are shown by red and yellow color regions in the ERRAT graph, whilst the normal parts are represented by white color. Plot analysis makes residues with error values greater than 95% and 99% easy to find. The result is displayed in form of overall quality score. A overall quality score of more than 95% is considered good with resolution of 2-3 $A^0$. Overall quality score below 85% cannot be consideration since it comes under bad score.

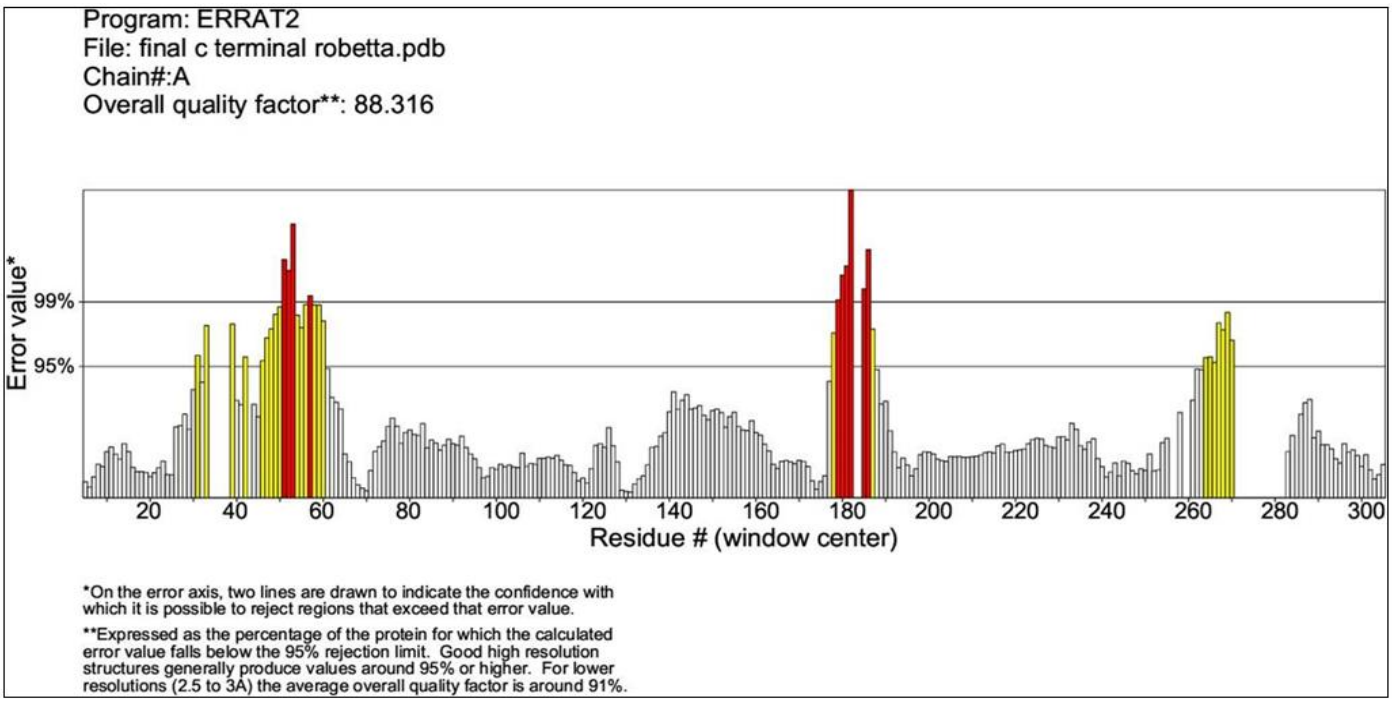


Fig 8 Robetta Predicted Model Validation Using EERAT Tool

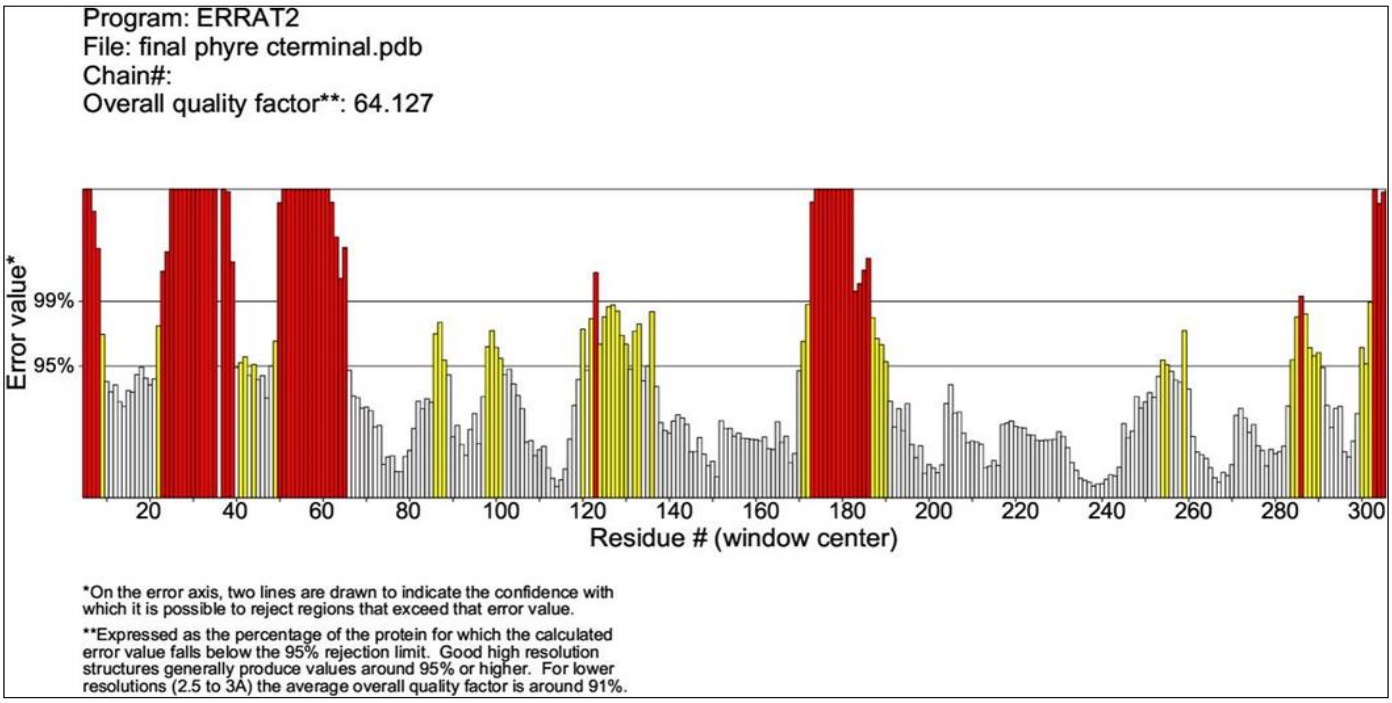


Fig 9 Phyre2 Predicted Model Validation Using EERAT Tool

Based on the ERRAT (residue v/s error value graph) analysis of the two predicted C-terminal structure, it can be concluded that Robetta predicted C terminal model is more accurate and valid compared to Phyre2 predicted C terminal structure, since a good quality model is expected to have overall quality value above 91%. According to results, Robetta predicted model has 88.316% overall quality value which is more then the phyre2 predicted model overall quality value i.e. 64.127%. The model quality value of Robetta predicted model is less than 91% but can be taken into consideration as the value is quite close whereas Phyre2 predicted model according to the value comes under the category of the bad quality model.

## IV. CONCLUSION

In this study the in-silico structure prediction methods were used to model the C-terminal domain of the dystrophin protein. According to the review of literature dystrophin protein is a part of protein complex DGC that works together to strengthen muscle fibers and protect them from injury as muscle contract and relax. Alteration of the protein causes severe diseases like DMD and BMD. To understand the reason for mutation or alteration, the complete structure of the dystrophin protein needs to be known. Structure of the N-terminal domain and spectrin repeats of the central rod domain is known, but there is no structure available for the C-terminal domain. So, for prediction of the C- terminal domain structure, in-silico method of structural modelling was used. Two structures of C-terminal domain were prepared on different software (ROBETTA and PHYRE2). This was done to get the most accurate structure. The predicted structures were then assessed for accuracy using the SAVESv6.0 software. Two tools of the SAVESv6.0 software were used - PROCHECK and ERRAT. Based on the results of the two tools, it was found that Robetta predicted C-terminal structure was the most accurate then the Phyre2 predicted structure. According to PROCHECK (Ramachandran plot analysis) results, Robetta predicted model has 90.8% residues in the most favored region with just 0.7% residues in the disallowed region which makes it a good quality model compared to the phyre2 predicted model which has only 69% residues in the most favored region with 7.0% residues in the disallowed region. The ERRAT results also shows that Robetta predicted structure is most accurate with overall model quality 88.316% compared to the phyre2 model quality which is only 64.127%. Based on the above results Robetta predicted C- terminal domain model was validated as the final C-terminal structure. Due to this preliminary encouraging result based on In-silico assessment, further studies are required for validation of our findings.

## ACKNOWLEDGMENT

All authors contributed to the study conception and design. Ms. Akanksha Mishra did conceptualization, performed review writing of the manuscript, material preparation, data collection and curation. Mr. Mohd. Maruf Khan helped in review writing. Dr. Anil Kothari, Dr. Nipun Silawat, and Dr. Pramod Sairkar made critical revisions, analyzed manuscript and supervised throughout the process. All authors read and approved the final manuscript. We acknowledge Madhya Pradesh Council of Science and Technology, Vigyan Bhawan, Nehru Nagar, Bhopal for providing infrastructure required to conduct the study.

➢ *Funding*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

➢ *Conflict of Interest*

All authors of this manuscript declare to have no conflict of interest.

## REFERENCES

[1]. Amann KJ, Renley BA, Ervasti JM(1998). A cluster of basic repeats in the dystrophin rod domain binds F-actin through an electrostatic interaction. J Biol Chem.; 273:28419–28423.
[2]. Rybakova IN, Amann KJ, Ervasti JM(1996). A new model for the interaction of dystrophin with F-actin. J Cell Biol.; 135:661–672.
[3]. Ponting CP, Blake DJ, Davies KE, Kendrick-Jones J, Winder SJ. ZZ and TAZ (1996): new putative zinc fingers in dystrophin and other proteins. Trends Biochem Sci.; 21:1113.
[4]. Verma, S., Anziska, Y. and Cracco, J. (2010) Review of Duchenne muscular dystrophy (DMD) for the pediatricians in the community. Clin. Pediatr. (Phila). 49, p1011-1017.
[5]. Hoffman, E.P., Brown, R.H., Jr. and Kunkel, L.M. (1987) Dystrophin: the protein product of the Duchenne muscular dystrophy locus. Cell Tissue Res., 51, pp 919-928.
[6]. Koenig, M., Hoffman, E.P., Bertelson, C.J., Monaco, A.P., Feener, C. and Kunkel, L.M. (1987) Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell, 50, pp509-517.
[7]. Williams, W., Thompson, M., and Morton, N. (1983). Complex segregation analysis and computer-assisted genetic risk assessment for Duchenne muscular dystrophy. American Journal of Medical Genetics 14, pp 315-333.
[8]. Muir, L.A. and Chamberlain, J.S. (2009) Emerging strategies for cell and gene therapy of the muscular dystrophies. Expert Reviews in Molecular Medicine, 11, e18.
[9]. Pichavant, C., Aartsma-Rus, A., Clemens, P.R., Davies, K.E., Dickson, G., Takeda, S., Wilton, S.D., Wolff, J.A., Wooddell, C.I., Xiao, X. et al. (2011) Current Status of Pharmaceutical and Genetic Therapeutic Approaches to Treat DMD. Mol Ther.
[10]. Mendell, J.R., Campbell, K., Rodino-Klapac, L., Sahenk, Z., Shilling, C., Lewis, S., Bowles, D., Gray, S., Li, C., Galloway, G. et al. (2010) Dystrophin immunity in Duchenne's muscular dystrophy. The New England Journal of Medicine 363, 1429- 1437.