"Structural and Evolutionary Profiling of Select Human Protein-Protein Interactions across Diverse Folds"

By

Ms. Deepashree M

ABSTRACT

Understanding the evolutionary conservation and functional significance of protein-protein interactions is crucial for explaining the complex biological mechanisms. This study explores the structural and functional relationships between two representative human proteins—Profilin-1 and Transthyretin and their respective interacting partners, Actin and Retinol Binding Protein 4 (RBP4). These proteins were selected from the SCOPe database to represent diverse structural classes (α+β and all-β) and well-characterized biological roles. The protein interactions were identified based on high-confidence interaction data obtained from STRING. The specific amino acid residues involved in these interactions were highlighted using the PDBe interface analysis tool and visually highlighted on the 3D structures with PyMOL. Profilin-1 was found to interact with Actin, contributing to the regulation of cytoskeletal structure and cell movement. Similarly, Transthyretin was shown to bind with Retinol Binding Protein 4 (RBP4), aiding in the transport of vitamin A throughout the body. Homologous sequences for each protein were obtained through BLAST searches across NR and PDB databases, followed by CD-HIT clustering to reduce redundancy. Multiple sequence alignments were performed using tools such as MAFFT, Kalign, MUSCLE, Clustal Omega, T-Coffee, Clustal (MEGA) and MUSCLE (MEGA), with Clustal (MEGA) consistently yielding the good alignment based on completeness score using ALISTAT evaluations. Structural alignment tools such as TM-align, DALI, and PROMALS3D were used to identify conserved three-dimensional motifs and domain structures. Phylogenetic trees generated using MEGA provided evolutionary relationships and potential co-evolution among the proteins and their interacting partners. To improve the clarity of these trees, annotation and node labeling were performed using iTOL, which allowed for a more detailed visualization of species-specific divergence. Building on this, functional analysis in Jalview enabled the mapping of conserved residues and disease-associated variants.

Overall, this study provides a deeper understanding of the evolutionary conservation, divergence, and potential functional significance of proteins by combining sequence clustering, multiple alignment, phylogenetic reconstruction, and functional residue analysis.

TABLE OF CONTENTS

Introduction

Proteins are essential macromolecules responsible for the majority of structural and functional activities in living organisms. They are composed of linear chains of amino acids linked by peptide bonds and fold into intricate three-dimensional shapes that determine their specific biological roles. These macromolecules perform a wide range of tasks, including acting as enzymes, hormones, antibodies, transporters, and structural scaffolds(1). Protein biosynthesis occurs through translation of messenger RNA on ribosomes, and their structure is organized hierarchically into primary, secondary, tertiary, and quaternary levels.

Protein-protein interactions (PPIs) are fundamental to nearly all biological processes. These interactions form the basis of multi-protein complexes and cellular pathways, enabling signal transduction, immune responses, gene regulation, and metabolic control (2). PPIs can be either transient or stable and vary in affinity and specificity. They often occur through structurally complementary interfaces and can be modulated by external stimuli and post-translational modifications.

Proteins interact via specific interfaces involving complementary shapes, electrostatic properties, and hydrogen bonding potentials. These interactions are primarily stabilized by non-covalent forces, including hydrophobic effects, van der Waals contacts, ionic interactions, and hydrogen bonds. Binding often involves conformational changes, supporting mechanisms such as induced fit or conformational selection to enhance specificity. The interface area and residue composition play crucial roles in determining interaction strength and selectivity. Domain-motif interactions, like SH3 domains binding to proline-rich motifs or PDZ domains binding to C-terminal peptide sequences, are classical examples of modular recognition in protein networks. Post-translational modifications such as phosphorylation and ubiquitination can serve as molecular switches that modulate the formation or dissociation of protein complexes (3).

For example, Profilin-1 (PFN1) is a small, actin-binding protein involved in the regulation of cytoskeletal dynamics. It plays an essential role in actin polymerization by binding to monomeric actin (G-actin) and modulating its availability for filament elongation. In STRING database, several proteins interact with Profilin-1 with strong experimental evidence. These include actin alpha skeletal muscle (ACTA1), Vasodilator-stimulated phosphoprotein (VASP), Diaphanous homolog 1 (DIAPH1), and Ena/VASP-like protein

(EVL), among others. For instance, PFN1 interacts directly with ACTA1 to facilitate the exchange of ADP for ATP on G-actin, thereby priming it for polymerization. VASP, a key player in cell motility and adhesion, interacts with PFN1 via its proline-rich motifs, enhancing actin filament elongation at the leading edge of motile cells. DIAPH1, a formin family protein, binds to both PFN1 and actin, acting as a scaffold to mediate actin nucleation and elongation.

Studying such protein-protein interactions provides valuable insights into the evolutionary dynamics of the proteome. By analyzing how these interactions have been conserved or altered over time, researchers can trace the functional significance of specific protein interfaces and networks.

To explore these interactions and how they have evolved, we selected two pairs of proteins for our study. We aim to analyze how these proteins interact and see how their structures and sequences have been conserved or changed over time.

1. Profilin-1 and Beta-actin
2. Transthyretin and Retinol binding protein-4

Profilin-1 (PFN-1)

Profilin-1 (PFN1) is a small, evolutionarily conserved actin-binding protein encoded by the *PFN1* gene in *Homo sapiens*. It plays a crucial role in the regulation of the actin cytoskeleton, a process fundamental to cell shape, migration, division, and intracellular transport(4). PFN1 primarily binds to monomeric globular actin (G-actin), enhancing the exchange of ADP for ATP and facilitating the incorporation of actin monomers into growing actin filaments (F-actin) (5).

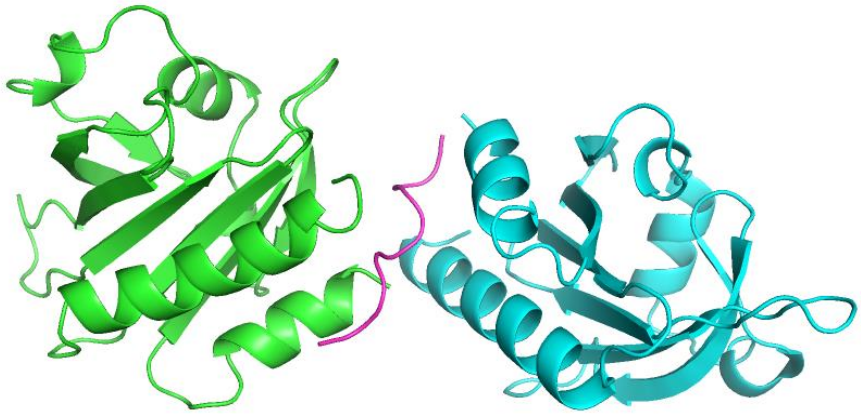


Figure 1: Crystal structure of Profilin-1 (PDB ID: 1AWI_A).

As described in Figure 1, Profilin-1 is a compact protein of approximately 140 amino acids (~15 kDa) composed predominantly of β-sheets and a few α-helices, forming a globular structure with distinct binding surfaces. It features a hydrophobic cleft that mediates actin binding, a separate surface for recognizing poly-L-proline motifs, and a lipid-binding region for phosphoinositides. These distinct functional surfaces enable PFN1 to participate in multiple cellular interactions simultaneously (6).

Beta-actin

Beta-actin (ACTB) is a highly conserved and abundant cytoskeletal protein encoded by the *ACTB* gene in *Homo sapiens*. It is one of the six actin isoforms expressed in mammalian cells and plays a fundamental role in maintaining cell shape, enabling motility, intracellular trafficking, and cell division. Beta-actin exists primarily in two forms within the cell: a monomeric globular form (G-actin) and a filamentous polymerized form (F-actin)(7), which assembles into dynamic microfilaments critical for the cytoskeletal network (8).

Figure 2: Crystal structure of Actin (PDB ID: 6NBW_A).

Figure 2 features a predominantly α/β fold, with each subdomain contributing to ATP binding, polymerization interfaces, and interaction with actin-binding proteins. The ATP or ADP nucleotide bound within the cleft regulates actin polymerization dynamics by influencing filament stability and growth. The protein undergoes conformational changes during the ATP hydrolysis cycle, which are essential for filament treadmilling and remodeling.

The actin monomer is approximately 42 kDa in size and comprises four subdomains arranged around a central ATP-binding cleft. This structure provides detailed insights into the monomeric form of actin and its interaction sites, facilitating the study of actin dynamics and its interaction (9).

Transthyretin

Transthyretin (TTR) is a homotetrameric protein encoded by the *TTR* gene in *Homo sapiens* that plays a key role in the transport of thyroid hormones and retinol-binding protein-vitamin A complex in the bloodstream and cerebrospinal fluid (10).

Each monomer of transthyretin is approximately 14 kDa in size, and the functional protein consists of four identical subunits arranged symmetrically to form a tetramer. Structurally, each monomer is composed predominantly of β-sheets organized into a β-sandwich fold, consisting of two four-stranded β-sheets that create a stable platform. The tetramer assembly creates two identical binding sites located at the dimer-dimer interface, which specifically bind thyroid hormones such as thyroxine (T4). This structural arrangement allows TTR to effectively transport thyroid hormones through the circulatory system (11).

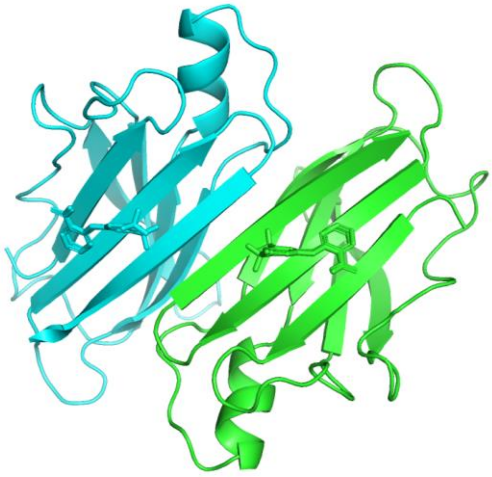


Figure 3: Crystal structure of Transthyretin (PDB ID: 1BM7_A).

Figure 3 is crystal structure of human transthyretin provides valuable insight into the quaternary arrangement of the protein and its ligand-binding properties. Transthyretin's stability and folding are critical because mutations or misfolding can lead to the formation of amyloid fibrils, which are implicated in diseases such as familial amyloid polyneuropathy (FAP) and senile systemic amyloidosis.

Retinol Binding Protein 4

Retinol Binding Protein 4 (RBP4) is a secreted plasma protein encoded by the *RBP4* gene in *Homo sapiens*, primarily responsible for the transport of retinol (vitamin A alcohol) from the liver to peripheral tissues (12). As the principal carrier of retinol in the bloodstream, RBP4 plays an essential role in maintaining vitamin A homeostasis, which is critical for vision, immune function, cellular differentiation, and embryonic development (11,13).

Structurally, RBP4 is a relatively small protein, approximately 21 kDa in size, characterized by a typical lipocalin fold—a β-barrel formed by eight antiparallel β-strands creating a hydrophobic pocket that tightly binds retinol. The binding pocket ensures the solubility and protection of the hydrophobic retinol molecule during transport through the aqueous environment of plasma.

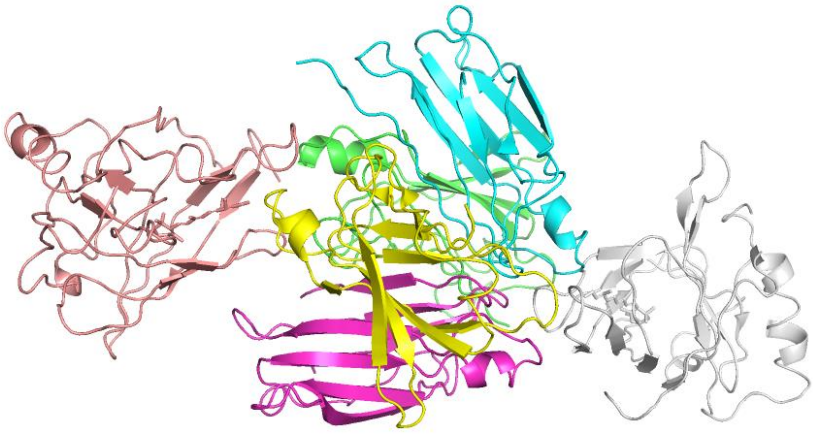


Figure 4: Crystal structure of Retinol binding protein 4 (PDB ID: 1QAB_E).

Figure 4 crystal structure of human RBP4, reveals the detailed interactions between RBP4 and retinol, including key amino acid residues involved in ligand binding and stabilization. The well-defined structure and critical physiological roles of RBP4 make it a valuable protein for understanding ligand-protein interactions, transport mechanisms, and evolutionary conservation in lipid-binding protein families.

Review of literature

A thorough understanding of protein structure, function, evolutionary history, and interactions is essential for decoding the molecular mechanisms that govern biological systems. Bioinformatics has emerged as a powerful approach, offering integrated tools to analyze structural data, protein networks, sequence alignments, and evolutionary patterns.

Structural domain classification serves as a fundamental strategy for comparative protein analysis. The SCOPe classification system enables researchers to detect conserved structural motifs that are maintained across functionally diverse protein families. Studies utilizing SCOPe have demonstrated that structural folds often correlate with conserved functional roles (14). Use of SCOPe in protein selection ensures that the chosen proteins are structurally characterized and biologically meaningful (15).

The STRING database combines protein-protein interaction data from multiple sources, including experimentally validated interactions, curated pathway databases and computational predictions, thereby generating an extensive network annotated with confidence scores (16). This resource enables the identification of interacting partners and functional modules, thereby offering enhanced insights into protein function and regulatory mechanisms.

Studies utilizing detailed interface analysis have demonstrated that identifying critical amino acids at the binding sites can reveal insights into allosteric regulation, complex stability, and specificity of interaction networks tools such as PyMOL enable researchers to explore the three-dimensional conformation of proteins, facilitating the analysis of key structural features including binding pockets, interface topology, and residue environments.

The identification of homologous sequences forms the evolutionary and comparative protein analysis. Tools such as the Basic Local Alignment Search Tool (BLAST) have become essential for detecting sequence similarity across diverse organisms due to their efficiency and reliability in discovering homologous relationships (17,18). By filtering out closely related sequences, CD-HIT helps ensure that conserved residues in the final alignment are indicative of true evolutionary conservation(19). Such curated alignments yield more accurate insights into structural and functional conservation, enabling downstream analyses such as phylogenetic reconstruction, motif identification, and functional annotation with greater confidence (20).

A range of multiple sequence alignment tools including MAFFT, MUSCLE, Kalign, T-Coffee, Clustal Omega, and MEGA-integrated versions of Clustal and MUSCLE have been developed, each employing distinct algorithms to balance computational efficiency and alignment accuracy. Studies have demonstrated that the selection of an alignment tool can substantially impact the quality of downstream analysis, such as phylogenetic tree inference and the detection of conserved residues. Tools like *AliStat* are commonly used to assess alignment quality. Literature suggests that Clustal, particularly when used through MEGA, often performs well in terms of producing complete and biologically meaningful alignments, making it a reliable option for evolutionary and functional analysis

Structure-based alignment approaches are particularly effective for discovering conserved features among related proteins, especially in cases where sequence similarity is limited. Tools such as TM-align, PROMALS3D, and DALI are instrumental in identifying structural motifs that may not be detectable through sequence-based methods alone. Existing literature shows the advantages of structural alignment tools in revealing biologically relevant similarities that sequence comparisons often overlook. These findings collectively emphasize the significance of integrating structural information into protein analysis to achieve a more comprehensive and accurate understanding of protein evolution and functional relationships.

The integration of phylogenetic analysis with functional residue annotation constitutes a powerful approach for explaining the evolutionary trajectories of proteins and understanding the functional consequences of sequence variation, particularly in the context of disease. Previous research has established that phylogenetic frameworks are instrumental in identifying highly conserved regions within protein sequences, which are representatives of essential functional roles. Assessing the extent to which functionally important or disease-associated residues coincide with these conserved regions enhances the interpretive value of the analysis, thereby offering a meaningful methodological basis for the investigation of protein function and evolutionary conservation.

AIM

To explore the structural, functional, and evolutionary properties of human proteins Profilin-1, Beta-actin, Transthyretin and Retinol Binding Protein 4 belonging to distinct structural folds, through comprehensive analysis of their protein-protein interactions, conserved residues and phylogenetic relationships using integrated bioinformatics approaches.

Objectives

- To select and classify structurally diverse proteins and their interacting partners using SCOPe and STRING.
- To retrieve and annotate 3D structures to identify interaction sites and functional residues.
- To obtain homologous sequences using BLAST and reduce redundancy through CD-HIT clustering.
- To perform multiple sequence and structure alignments to detect conserved regions.
- To construct and visualize phylogenetic trees to explore evolutionary relationships.
- To map conserved and disease-associated residues for functional interpretation.

Materials

In this study, a range of bioinformatics tools and databases were utilized for the retrieval of protein sequences, functional annotation, structural modeling, and interaction analysis, along with visualization and comparative analysis to explore the evolutionary relationships between selected protein pairs.

1. SCOPe

The Structural Classification of Proteins—extended (SCOPe) is a database that classifies protein structural domains based on evolutionary and structural relationships. It organizes proteins into a hierarchy of classes, folds, superfamilies, and families. In this study, SCOPe was used to identify the structural class and domain architecture of the selected proteins, aiding in evolutionary comparison. SCOPe integrates manual curation with automated updates from the Protein Data Bank (PDB), making it a reliable resource (21).

2. STRING

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a comprehensive database and web resource that compiles known and predicted protein-protein interactions from multiple sources, including experimental data, computational prediction methods, and public text collections. It provides confidence scores for interactions based on the evidence supporting them. In this study, STRING was used to identify and analyze the interaction partners of selected proteins, focusing on those with experimental validation to understand functional associations and biological pathways (16).

3. PDB

The Protein Data Bank (PDB) is a global repository that archives three-dimensional structural data of biological macromolecules, including proteins, nucleic acids, and complexes, determined by experimental methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. The RCSB PDB provides tools for structure visualization, analysis, and download, supporting research in structural biology, drug design, and bioinformatics. In this study, protein structures relevant to the selected proteins were retrieved from the PDB to analyze their three-dimensional conformation and interaction interfaces, facilitating structural comparisons and evolutionary insights (22).

4. PDBe

The Protein Data Bank in Europe (PDBe) is a part of the Worldwide Protein Data Bank (wwPDB) consortium and provides access to macromolecular structure data archived in the PDB. PDBe offers advanced visualization tools, annotations, and structural quality assessment features, facilitating interpretation of structural information in a biological context. In this study, PDBe was used to explore detailed structural annotations, visualize protein domains and binding sites, and download coordinate files for further comparative and functional analysis(22,23).

5. PDBsum

PDBsum is a web-based resource that provides detailed schematic summaries and graphical representations of the structural and functional features of proteins deposited in the Protein Data Bank (PDB). It includes information on secondary structures, ligand interactions, domain architecture, and protein-protein interfaces. In this study, PDBsum was used to analyze and visualize structural features of the selected proteins, including interaction sites, active sites, and secondary structure elements, aiding in the interpretation of their functional roles and interaction patterns(24).

6. PyMOL

PyMOL is an open-source molecular visualization tool widely used for analyzing and rendering high-quality 3D structures of biomolecules (25). It supports visualization of proteins, nucleic acids, ligands, and macromolecular complexes, allowing users to inspect structural features such as binding pockets, secondary structures, and interaction interfaces. In this study, PyMOL was employed to visualize the tertiary structures of selected proteins and to highlight interaction sites and conserved residues, contributing to comparative and evolutionary analysis of protein-protein interactions.

7. BLAST

The Basic Local Alignment Search Tool (BLAST) is a widely used algorithm for comparing primary biological sequences, such as protein or nucleotide sequences, to sequence databases

to identify regions of local similarity. It helps infer functional and evolutionary relationships between sequences and identify homologs. In this study, BLAST was used to retrieve homologous protein sequences from the NR and PDB database, enabling comparison of sequence conservation and selection of relevant structural entries for evolutionary and structural analysis (18).

8. CD-HIT

CD-HIT (Cluster Database at High Identity with Tolerance) is a widely used clustering algorithm designed to efficiently group protein or nucleotide sequences based on user-defined sequence identity thresholds. It is commonly used to reduce redundancy and select representative sequences from large datasets. The algorithm functions by first sorting input sequences by length, with the longest sequence chosen as the initial cluster representative. It then employs a short-word-based heuristic to efficiently detect similarity between sequences, allowing rapid comparisons. Sequences that share identity above a defined threshold with an existing cluster representative are grouped into the same cluster, while dissimilar ones form new clusters. In this study, CD-HIT was employed to cluster homologous protein sequences retrieved via BLAST, ensuring a non-redundant set of representative sequences for structural and evolutionary analysis. The tool was executed in a Linux-based Ubuntu environment to enable efficient command-line processing and management of sequence data (20).

9. Multiple Sequence Alignment

9.1. Kalign

Kalign is a fast and accurate multiple sequence alignment (MSA) tool that utilizes the Wu-Manber string-matching algorithm, offering high performance and scalability for aligning large sets of sequences. It is especially useful in evolutionary and structural studies to identify conserved motifs and functionally important residues across homologous proteins. Due to its speed and reliability, Kalign is particularly effective for analyzing large datasets. In this study, Kalign was used to align non-redundant protein sequences to reveal conserved regions and patterns critical to structural and functional interpretation. All alignments were performed in a Linux-based Ubuntu environment to ensure efficient and reproducible command-line execution (26).

9.2. MAFFT

MAFFT (Multiple Alignment using Fast Fourier Transform) is a high-speed, accurate tool for multiple sequence alignment, widely used in comparative genomics and evolutionary studies. It supports various alignment strategies, including options for large datasets and highly divergent sequences. MAFFT uses fast Fourier transform techniques to improve speed without compromising accuracy. In this study, MAFFT was employed to align protein sequences clustered using CD-HIT, helping identify conserved regions and functional motifs. All alignments were performed using the command-line version of MAFFT in a Linux-based Ubuntu environment for efficiency and reproducibility (27).

9.3. MUSCLE

MUSCLE (Multiple Sequence Comparison by Log-Expectation) is a widely used tool for multiple sequence alignment, known for its accuracy and efficiency. It employs progressive alignment techniques along with refinement steps to improve alignment quality, making it suitable for analyzing conserved regions, phylogenetics, and structural predictions. In this study, MUSCLE was used to align homologous protein sequences to identify conserved residues and analyze sequence variability, which supported subsequent structural and evolutionary comparisons (28).

9.4. Clustal omega

Clustal Omega is a widely used multiple sequence alignment tool that combines speed, scalability, and accuracy. It uses a guide-tree and hidden Markov model (HMM) profile-profile techniques to align large numbers of sequences efficiently while maintaining alignment quality. Clustal Omega is commonly applied in evolutionary studies and functional annotation to identify conserved sequence regions across diverse proteins. In this study, Clustal Omega was utilized to perform multiple sequence alignments of protein sequences to reveal evolutionary conservation and functional motifs relevant to structural analysis (29).

9.5. T - Coffee

T-Coffee (Tree-based Consistency Objective Function For Alignment Evaluation) is a multiple sequence alignment program that improves alignment accuracy by combining results from several alignment methods into a consensus alignment. It uses a consistency-based

scoring system to refine alignments, making it particularly useful for aligning distantly related sequences or datasets with variable sequence similarity. In this study, T-Coffee was employed to generate accurate multiple sequence alignments of protein sequences to identify conserved and functionally important regions that inform evolutionary and structural analyses (30).

10. Multiple Structure Alignment

10.1. TM -align

TM-align is a computational tool used for the structural alignment of protein 3D structures. It compares two protein structures by optimizing the TM-score, a normalized measure of structural similarity that is independent of protein size. TM-align is widely used for identifying structural homologs and analyzing evolutionary relationships based on fold conservation. In this study, TM-align was employed to perform pairwise structural alignments of selected protein models to evaluate their conformational similarities and support evolutionary interpretations. All structural alignments were carried out using TM-align in a Linux-based Ubuntu environment to ensure efficient and reproducible command-line processing (31).

10.2. DALI

DALI (Distance-matrix ALIgnment) is a widely used algorithm for comparing protein three-dimensional structures by calculating intramolecular distance matrices and identifying structural similarities. It enables the detection of distant homologs and fold similarities, providing insights into protein evolution and function. In this study, DALI was used to perform structural comparisons between selected protein models to assess their conformational relatedness and evolutionary conservation (32).

10.3. PROMALS3D

PROMALS3D is an advanced multiple sequence and structure alignment tool that integrates sequence information with three-dimensional structural data to produce highly accurate alignments. By combining evolutionary information and structural constraints, PROMALS3D improves the alignment of distantly related proteins, which is critical for functional and

evolutionary studies. In this study, PROMALS3D was utilized to generate refined multiple sequence alignments incorporating structural data, thereby enhancing the identification of conserved residues and motifs relevant to protein function and evolution (33).

11. AliStat

Alistat is a tool designed for detailed analysis of multiple sequence alignments. It provides various statistics, including sequence conservation, composition, and completeness score across alignment columns, facilitating the identification of conserved and functionally important regions. In this study, Alistat was used to analyze multiple sequence alignments of protein families from various tools enabling the assessment of evolutionary conservation and guiding further structural and functional interpretations (34).

12. MEGA

MEGA (Molecular Evolutionary Genetics Analysis) is a comprehensive software suite widely used for conducting sequence alignment, phylogenetic tree construction, and evolutionary analysis. It offers various algorithms for inferring evolutionary relationships, estimating divergence times, and analyzing molecular sequence data. In this study, MEGA was utilized to construct phylogenetic trees based on aligned protein sequences, enabling the exploration of evolutionary relationships among homologous proteins (35).

13. iTOL

iTOL (Interactive Tree Of Life) is a web-based tool designed for the visualization, annotation, and management of phylogenetic trees. It allows users to upload tree files in various formats and customize them with metadata, colors, and labels, facilitating intuitive interpretation of evolutionary relationships. In this study, iTOL was used to visualize and annotate phylogenetic trees generated from protein sequence alignments, enabling clear presentation and comparative analysis of evolutionary data (36).

14. Jalview

Jalview is an interactive software tool for multiple sequence alignment visualization, editing, and analysis. It supports a range of features including alignment editing, annotation, secondary structure prediction, and integration with external databases. Jalview facilitates the identification of conserved regions, sequence motifs, and functional sites within protein or nucleotide alignments. In this study, Jalview was used to visualize and refine multiple sequence alignments, enhancing the interpretation of sequence conservation and functional relationships (37)**.**

Methods

1. Retrieval and Selection of Proteins.

The study began with the retrieval of proteins from the **SCOPe (Structural Classification of Proteins-extended)** database, focusing on different structural classes including **alpha and beta (α/β), alpha plus beta (α+β),** and **all-beta proteins**.

Among the retrieved entries, **Profilin-1 and Transthyretin** were selected based on their distinct biological functions and well-characterized structural data, making them suitable candidates for evolutionary comparison.

2. Identification of Interacting Partners.

To further understand the biological roles of these target proteins, their interacting partners were identified through the **STRING database**. To ensure high confidence, the search was restricted to interactions supported by experimental evidence, minimizing potential false positives.

This approach revealed **Beta-actin** as the interacting partner of Profilin-1.

Structural availability for both interacting partners was verified through the **Protein Data Bank (PDB)**, and each protein was also classified according to the SCOPe database. These identified interactions formed the basis for exploring possible co-evolutionary patterns and functional conservation within the protein interaction network.

3. Identification of Functionally Important Residues.

The key interacting residues at the protein-protein interface were identified using the PDBsum interface analysis tool and ligand binding sites and active sites were identified using PDBe.

The interacting residues were then visually represented using **PyMOL** software, enabling the highlighting of key residues on the three-dimensional structures.

4. Homologous Sequence Retrieval and Clustering.

Homologous sequences for the selected proteins and their partners were retrieved using **BLAST (Basic Local Alignment Search Tool)** from NCBI server searches against multiple databases, including the **NR (non-redundant)** database for broad sequence coverage, **UniProt** for high-quality annotated sequences, and **PDB** for structural homologs.

To reduce sequence redundancy and retain representative sequences, clustering was performed using **CD-HIT (Cluster Database at High Identity with Tolerance)**. The clustering process was executed in a **Linux (Ubuntu)** environment with the following command:

```
cd-hit -i input.fasta -o output.fasta -c 0.9 -n 5
```

where the arguments are as follows:

-i input.fasta → Input FASTA file containing protein sequences

-o output.fasta → Output file for representative sequences

-c 0.9 → Sequence identity threshold (90%)

-n 5 → Word length (default for 90% identity)

A range of sequence identity cut-offs (0.95 to 0.55) was tested to achieve an optimal balance between diversity and redundancy reduction. An identity threshold of **70% (0.7)** was chosen as the optimal cut-off.

5. Multiple Sequence Alignment (MSA)

To identify conserved motifs and regions, multiple sequence alignment (MSA) was performed using a range of established tools: **Kalign**, **MAFFT**, **MUSCLE**, **Clustal Omega**, **T-Coffee, Clustal (MEGA) and Muscle (MEGA)**.

Given the large number of sequences (often exceeding 500), command-line execution was used for **Kalign** and **MAFFT** to efficiently handle large datasets. For smaller datasets, web-based interfaces of **MUSCLE**, **Clustal Omega**, and **T-Coffee** were employed for alignment. **Clustal and Muscle** were executed using MEGA software.

The following command examples were used for command-line alignments:

```
kalign -i input.fasta -o output.aln
```

where the arguments are as follows:

-i input.fasta → Input FASTA file containing sequences to be aligned.

-o output.aln → Output file containing the aligned sequences. (It can be. aln, .fasta, .clu, etc.).

```
mafft input.fasta > output.aln
```

where the arguments are as follows:

input.fasta → Input FASTA file containing sequences to be aligned.

output.aln → Output file containing the aligned sequences.

The alignments produced by these tools were evaluated to assess sequence conservation, alignment consistency, and gap handling.

6. **Alignment quality assessment**

Following multiple sequence alignment, the quality of each alignment was evaluated using **ALISTAT**, a tool designed to assess alignment statistics. This tool was used to analyze the number of sequences in the alignment and completeness score. Later gap percentage was found for each alignment to determine the overall alignment quality.

A script was used to run Alistat and calculate gap percentage across the alignments.

```
~/AliStat/alistat input_file.fasta > output_summary.txt
```

where the arguments are as follows:
input_file.fasta → Input FASTA file containing alignment to be evaluated.
output_summary.txt
→ Output summary file.

7. Multiple Structure Alignment

Multiple structure alignment was conducted to detect conserved structural features across homologous proteins. Structural alignment is particularly valuable for revealing conserved folds and motifs that are retained despite sequence divergence.

The tools employed for this task included **TM-align** (executed via command line), **PROMALS3D**, and **DALI** (both accessed through their web interfaces). TM-align was selected for its capability to handle large-scale alignments, while PROMALS3D and DALI provided structure-based alignment limited to 30 and 64 structures, respectively.

The following command was used for TM-align execution:

```
#!/bin/bash

mkdir -p results  # make sure results folder exists
cd inputs || { echo "inputs folder not found!"; exit 1; }

files=(*.pdb)

for ((i = 0; i < ${#files[@]}; i++)); do
    for ((j = i + 1; j < ${#files[@]}; j++)); do
        file1="${files[$i]}"
        file2="${files[$j]}"
        echo "Aligning $file1 vs $file2 ..."
        ../TMalign            "$file1"            "$file2"            >
"../results/${file1%.pdb}_${file2%.pdb}.txt"
    done
done

echo "All alignments complete. Check the 'results' folder."
```

8. Construction of Phylogeny tree

To explore the evolutionary relationships among the selected protein sequences, phylogenetic trees were constructed using MEGA (Molecular Evolutionary Genetics Analysis). Multiple sequence alignments generated from earlier steps (Clustal) were input into MEGA, where evolutionary distances were computed using the Poisson correction model. Tree construction was performed using the Neighbor-Joining algorithm with 100 bootstrap replicates to assess the robustness of the inferred relationships. The resulting phylogenetic trees offered valuable insights into the divergence patterns and evolutionary conservation within the studied protein families.

The command line used for MEGA is

```
~/megacc/megacc -a mega_phylogeny_parameters.mao -d input_file -o
output_file
```

9. Tree Visualization and Annotation

The phylogenetic trees generated in MEGA were exported in Newick (.nwk) format and visualized using iTOL (Interactive Tree of Life). To enhance interpretability and include dataset-specific annotations, Python scripts were written to generate labeling and metadata files compatible with iTOL. These files included custom labels, color schemes, and dataset-specific annotations sequence ID, protein names, organisms, which were uploaded alongside the tree file.

The script used to generate labels:

```
import os
meg_files_folder= '/home/meg_files'
meg_files = [f for f in os.listdir(meg_files_folder) if f.endswith('.meg')]
for meg_file in meg_files:
        with     open(f"{meg_file.split('.')[0]}_labels.txt",     'w')     as
metadata_file:
            meg_file = os.path.join(meg_files_folder, meg_file)
            metadata_file.write(""" LABELS
#use this template to change the leaf labels, or define/change the internal
node names (displayed in mouseover popups)
#lines starting with a hash are comments and ignored during parsing
#================================================================#
#                     MANDATORY SETTINGS                         #
#================================================================#
#select the separator which is used to delimit the data below (TAB,SPACE or
COMMA).This  separator  must  be  used  throughout  this  file  (except  in  the
SEPARATOR line, which uses space).
SEPARATOR TAB
#SEPARATOR SPACE
#SEPARATOR COMMA
#Internal tree nodes can be specified using IDs directly, or using the 'last
common ancestor' method described in iTOL help pages
#================================================================#
#       Actual data follows after the "DATA" keyword             #
#================================================================#
DATA
#NODE_ID,LABEL""")
            metadata_file.write("\n")
            with open(meg_file, 'r') as f:
                for ln in f.readlines():
                    if ln.startswith("#") and "_" in ln:
                        # print(f"Processing line: {ln.strip()}")
                        ln = ln.strip().replace('#', '')
                        accession_id = ln.strip().split('"')[0][:-1]
                        label_text = ln.strip().split('"')[1]
```

```
                        protein = label_text.split('[')[0][:-1]
            if 'hypothetical' in protein or 'uncharacterized' in protein or
'unnamed_protein_product' in protein:
                            protein = 'unannotated'
                        organism = label_text.split('[')[1].split(']')[0]
                        final_label                                                =
f"{accession_id}_{protein}_{organism}".replace("PREDICTED:","PREDICTED")

                        #  print(f"Accession  ID:  {accession_id},  Protein:
{protein},                        Organism:                        {organism}")
metadata_file.write(f"{accession_id}\t{final_label}\n")
```

The script used to generate color annotation:

```
from Bio import SeqIO
from collections import defaultdict
from matplotlib import cm
from matplotlib.colors import to_hex
import sys
# ----------
# INPUT
# ----------
genpept_file =  sys.argv[1] if len(sys.argv) > 1 else print("Usage: python
get_color_annotation.py <genpept_file>")

if not genpept_file.endswith('.gb'):
    raise ValueError("The input file must be a GenBank file with a .gb
extension.")

# Parsing the GenBank file to extract accession IDs and taxonomy

itol_annotation_file = genpept_file.replace('.gb', '_itol_annotation.txt')
accession_ids_list = []
taxonomy_list = []
for record in SeqIO.parse(genpept_file, "genbank"):
    accession_ids_list.append(record.id)
    # record.annotations['taxonomy'] is already a list of ranks:
    taxonomy_list.append(record.annotations.get('taxonomy', []))
assert len(accession_ids_list) == len(taxonomy_list)
taxonomy_dict = dict(zip(accession_ids_list, taxonomy_list))
print(len(taxonomy_dict.keys()))
#  Generating  the  iTOL  annotation  file  with  colors  based  on  2nd-level
taxonomy
# 1. Collect all unique 2nd-level taxa
second_levels = sorted({tax[1] for tax in taxonomy_dict.values() if len(tax)
> 1})
# 2. Generate distinct colors for each 2nd-level group using a colormap
cmap = cm.get_cmap('tab20', len(second_levels))
color_map = {lvl: to_hex(cmap(i)) for i, lvl in enumerate(second_levels)}
# 3. Write the iTOL annotation file
output_file = 'itol_annotation.txt'
with open(output_file, 'w') as out:
    out.write("TREE_COLORS\nSEPARATOR TAB\nDATA\n")
    for acc, tax in taxonomy_dict.items():
```

```
        if len(tax) > 1:
            lvl = tax[1]
            color = color_map[lvl]
            out.write(f"{acc}\trange\t{color}\t{lvl}\n")

print(f"Annotation file written to: {output_file}")
```

10. Alignment Visualization

Jalview was employed to visualize and analyze the multiple sequence alignments generated during the study. Alignment files from Clustal (MEGA) were imported into Jalview to inspect conserved residues and assess alignment quality. Jalview' s integrated features such as color schemes, secondary structure mapping, and annotation tools enabled a clearer analysis of conserved sites and sequence variation within the studied protein families.

To focus on biologically relevant homologs, sequence clusters were retrieved from CD-HIT at a 0.7 identity cut-off, Clusters containing ***Homo sapiens*** **counterparts** were specifically selected and retrieved using python script and then it was visualized using Jalview.

```
from Bio import SeqIO
# === Step 1: Extract IDs from Cluster no ===
cluster_file = "Protein_name.txt.clstr"
cluster_number = "Cluster no"
ids = set()
with open(cluster_file) as f:
    in_cluster = False
    for line in f:
        line = line.strip()
        if line.startswith(">Cluster"):
            in_cluster = (line == f">{cluster_number}")
        elif in_cluster and ">" in line:
            try:
                seq_id = line.split(">")[1].split("...")[0]
                ids.add(seq_id)
            except IndexError:
                continue

# === Step 2: Extract sequences from FASTA file ===
input_fasta = "protein name.txt"
output_fasta = "cluster_sequences.fasta"

count = 0
with open(output_fasta, "w") as out_handle:
    for record in SeqIO.parse(input_fasta, "fasta"):
        fasta_id = record.id.split()[0]
        if fasta_id in ids:
            SeqIO.write(record, out_handle, "fasta")
            count += 1
print(f" Extracted {count} sequences from Cluster no  to '{output_fasta}'")
```

11. **Identification of Mutant Variants**

To investigate functionally important variations, mutant variants of each protein were retrieved from the UniProt database. The search was performed using the reviewed UniProt entry corresponding to the ***Homo sapiens*** version of each protein. Within each entry, the "Natural Variants" and "Sequence annotation (Features)" sections were explored to identify missense mutations and known polymorphisms. For each variant, the following details were extracted: the **original residue**, the **mutated residue**, the **position of mutation**, the **documented phenotypic effect**, and any **associated diseases or conditions** as reported in UniProt or linked literature.

All collected data were systematically tabulated to provide a comparative overview. The resulting table included the following columns: (i) **Residue Position**, (ii) **Original Residue**, (iii) **Mutated Residue**, (iv) **Effect**, and (v) **Associated Disease**. Variants with no known disease correlation but listed as polymorphisms or with experimental evidence were also included for completeness.

Results

1.a. Retrieval of proteins from SCOPe database.

	PROTEIN	**INTERACTING PROTEIN**
PROTEIN NAME	**Profilin-1**	**Beta-actin**
CLASS	Alpha and beta **(α+β)**	Alpha and beta (**(α/β)**
FOLD	Profilin-like	Ribonuclease H - like motif
SUPERFAMILY	Profilin	Actin like ATPase domain
FAMILY	Profilin	Actin
SPECIES	*Homo sapiens*	*Homo sapiens*
PDB ID	1AWI_A	6NBW_A

Table 1. Structure and Classification Details of Profilin-1 and its Interacting Protein Beta-Actin.

	PROTEIN	**INTERACTING PROTEIN**
PROTEIN NAME	**Transthyretin**	**Retinol binding protein 4**
CLASS	All beta **(β)**	All beta **(β)**
FOLD	Pre-albumin like	Lipocalins
SUPERFAMILY	Transthyretin	Lipocalins
FAMILY	Transthyretin	Retinol binding protein - like
SPECIES	*Homo sapiens*	*Homo sapiens*
PDB ID	1BM7_A	1QAB_E

Table 2. Structure and Classification Details of Transthyretin and its Interacting Protein Retinol binding protein 4.

Tables 1 and 2 summarize the structural classification of the human Profilin-1 (PDB ID: 1AWI_A) and Transthyretin (PDB ID: 1BM7_A) its experimentally validated interacting partner Beta-Actin (PDB ID: 6NBW_A) and Retinol binding protein 4 (PDB ID: 1QAB_E). The information includes SCOPe-based class, fold, superfamily, and family levels, along with species and PDB identifiers.

1.b. Selection of interacting proteins using string database.

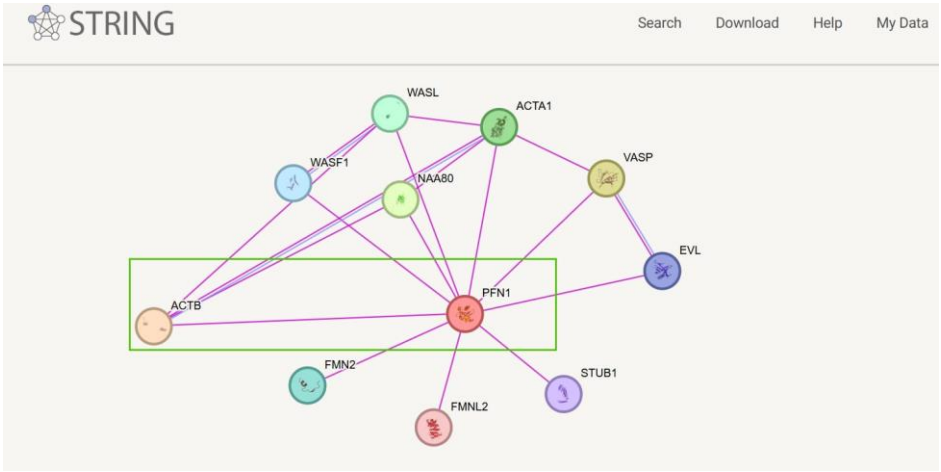
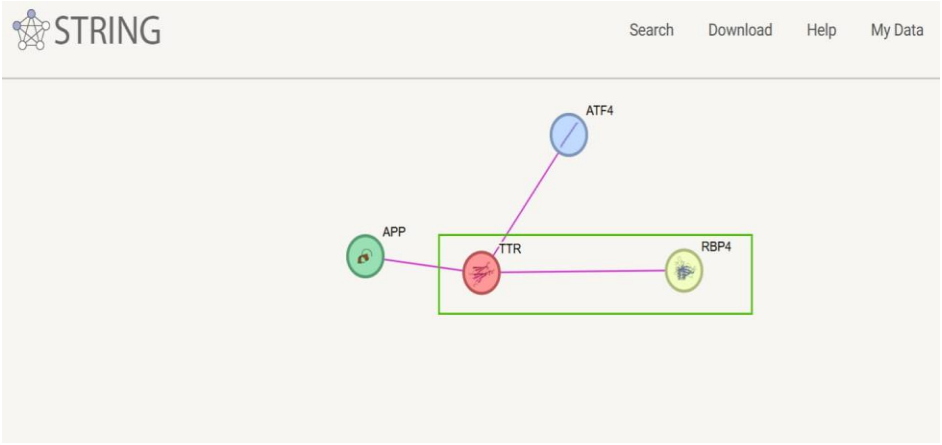


Figure 5. Protein–Protein Interaction (PPI) Network of Profilin-1 Constructed Using STRING Database.

The Figure 5 represents the predicted and experimentally validated protein–protein interaction network of Profilin-1 (PFN1) as obtained from the STRING database. Each node represents a protein, while edges represent evidence of interaction. The central node (PFN1) shows direct associations with multiple proteins including Beta-actin (ACTB), ACTA1, VASP, WASF1, EVL, STUB1, FMNL2, and others. The green box highlights the interaction



between PFN1 and ACTB, supported by experimental data.

Figure 6. Protein-Protein Interaction (PPI) Network of Transthyretin Constructed Using STRING Database.

The Figure 6 represents the predicted and experimentally validated protein-protein interaction network of Transthyretin as obtained from the STRING database. Each node represents a protein, while edges represent evidence of interaction. The central node (TTR) shows direct

associations with multiple proteins including Retinol binding protein 4 (RBP4), APP, ATF4. The green box highlights the interaction between TTR and RBP4, supported by experimental data. Such networks help identify functional partnerships, co-expression trends, and evolutionarily conserved interaction pathways.

2. Functionally important residues

To identify functionally important residues, information was retrieved from both the **PDBe** and **PDBsum** databases. PDBe was used to access detailed structural annotations of each protein, while PDBsum provided comprehensive residue-level insights, including **ligand-binding sites**, **active sites**, **interaction interfaces**, and **catalytic residues**. The 3D structure of the protein from PyMOL is shown and colored by chain. Functionally important residues are highlighted and represented as **spheres**. **Ligand binding sites** are enclosed within a **red box**, **interaction interfaces** are marked by a **yellow box**, and **active site residues** are highlighted within a **black box**. This visualization provides insight into the spatial organization of critical functional regions and their potential roles in protein activity and interactions. These residues are considered as residue of interest.

In the figure 7 crystal structure of Profilin-1 (PDB ID: 1AWI), the protein is complexed with **actin** and an **ATP molecule**, which plays a role in stabilizing the interaction. Profilin-1 is also known to interact with polyproline-rich motifs and phosphoinositide lipids such as PIP2, contributing to its role in actin cytoskeleton regulation and signal transduction.

For **Profilin-1 (Chain A)**, the following surface-exposed residues were identified as part of the

- **Interaction - interface:** Gly1, Asn3, His133, Arg135
- **Ligand binding sites:** Trp2, Asn3, Ile7, Lys24, Asp25, Ser26, Val29, Trp30, Ala31, Ala32, Ile72, Asp86, Arg87, Val99, Thr100, Lys106, His118, Gly119, Gly120, Asn123, Tyr138
- **Active site:** Trp3, Trp31, Tyr6, His133, Tyr139.

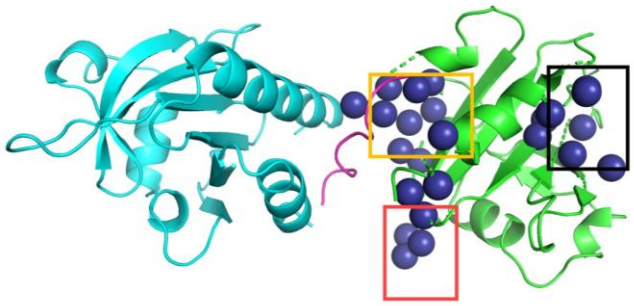


Figure 7: PyMOL visualization of the 1AWI_A PFN-1 is shown in green color. Residues of C^{a} atoms are shown as dark blue spheres. Residues involved in interaction interface, ligand binding sites, active sites are shown in yellow, red and black color boxes respectively.

In the figure 8, PDB ID: 6NBW, Actin proteins typically bind to **ATP or ADP** along with a **divalent metal ion** such as **Mg^{2+}** or **Ca^{2+}**, which are crucial for nucleotide stabilization and actin polymerization dynamics. In structural studies, these ligands are consistently observed in the nucleotide-binding pocket of G-actin monomers.

For **Actin (Chain A)**, the residues included

- **Interaction interface**: Asp1, Ala5, Phe20, Gly22, Asp24, Arg27, Leu93 - Val95, Glu99 - His100, Lys112, Tyr132, Tyr143- Ala144, Ser145, Gly146, Arg147, Leu 166, Tyr167, Glu 168, Gly169, His173, Pro172, Ile233 - Asp 286, Val287, Ile289, Ile341, Gly343, Ser344- Ile345, Ser348, Leu349, Ser350, Gln354- Met355, Glu360, Gln352
- **Ligand binding sites:** Gly12- Met15, Lys17, Pro31, Ile33, Gln58, Leu66, Tyr68, Glu71, Gly155- Val158, Gly181, Thr185, Asp 183, Ser 198- Phe199, Glu 204, Arg205- Glu206, Arg209, Lys212, Glu213, Gly300-Thr302, Met304, Tyr305, Lys335.

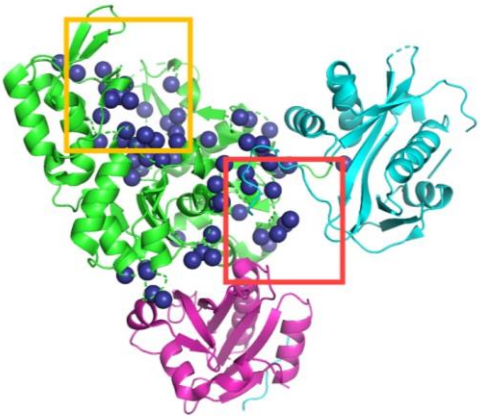


Figure 8: PyMOL visualization of the 6NBW_A Actin is shown in green color. Residues of C^{a} atoms are shown as dark blue spheres. Residues involved in interaction interface and ligand binding sites are shown in yellow and red color boxes respectively.

In the figure 9, crystal structure of human transthyretin (PDB ID: 1BM7), the protein is complexed with **thyroxine (T4)**, a thyroid hormone. Thyroxine binds at the dimer–dimer interface of the transthyretin tetramer, playing a vital role in hormone transport. This ligand interaction is also pharmacologically significant in stabilizing TTR to prevent amyloid fibril formation.

For **Transthyretin (Chain A)**, the highlighted interacting residues included

- **Ligand binding sites**: Lys15, Leu17, Thr106, Ala108, Leu110, Ser117, Thr119, Val121.
- **Interaction interface:** Leu17, Ala19 - Pro24, Ile68, Phe87- His90, Glu92 - Thr96, Tyr105, Ile107, Leu 110, Ser112 - Val124.

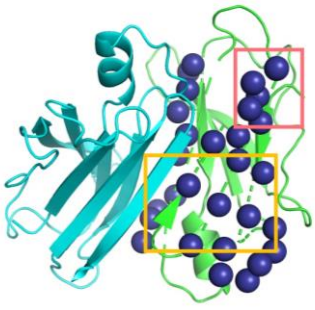


Figure 9: PyMOL visualization of the 1BM7_A TTR is shown in green color. Residues of C^{a} atoms are shown as dark blue spheres. Residues involved in interaction interface and ligand binding sites are shown in yellow and red color boxes respectively.

In the figure 10, Retinol Binding Protein 4 (RBP4) specifically binds **all-trans-retinol (ROL)** within a hydrophobic β-barrel cavity. This interaction facilitates the transport of vitamin A from hepatic stores to peripheral tissues, essential for vision, cellular growth, and differentiation. In structural studies such as PDB ID: 1QAB_E, retinol is observed tightly bound within the central cavity of RBP4.

For **Retinol Binding Protein 4 (Chain E)**,

- Ligand binding sites: Leu32 - Leu34, Phe42, Ala52, Ala54, Val58, Met70, Val71, Phe74, Met85, Tyr87, Leu94, Gln95, His 101, Tyr130, Phe132,
- Interaction interfaces: Leu32, Leu60 - Trp64, Lys86, Trp88, Val90, Ser92 - Lys96, Glu176, Leu179, Leu180.

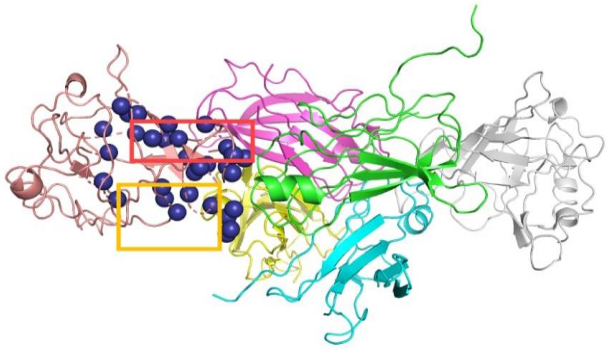


Figure 10: PyMOL visualization of the 1QAB-E RBP4 is shown in light pink colour. Residues of C^{a} atoms are shown as dark blue spheres. Residues involved in interaction interface and ligand binding sites are shown in yellow and red color boxes respectively.

3.1 BLAST result summary

A BLASTp (protein BLAST) search was performed for each selected protein, with the maximum number of target sequences set to 5000. The searches were conducted separately against the NR and PDB databases to identify homologous protein sequences.

Protein name	UNIPROT ID	NR	PDB
PFN-1	P07737	2513	16
ACTB	P60709	5000	169
TTR	P02766	4348	131
RBP4	P02753	3380	30

Table 3. Summary of BLASTp Results Across NR and PDB Databases for Selected Proteins.

Table 3 summarizes the number of homologous sequences retrieved for each protein from the NR and PDB databases. All proteins returned the maximum of 5000 sequences from the NR database, indicating a high level of conservation across species. The PDB database provided a distinct set of structurally characterized homologs. The retrieved sequences were subsequently clustered using CD-HIT to reduce redundancy.

3.2 CD-HIT result summary

Sequences obtained from the NR and PDB databases were clustered using CD-HIT to reduce redundancy and retain representative homologs for each protein. CD-HIT clustering was performed at various identity thresholds (0.9, 0.85, 0.8, 0.75, 0.7, and 0.65) to assess the redundancy among homologous sequences from the NR and PDB databases. We observed that at identity levels above 0.7, the number of clusters remained very high, reflecting substantial redundancy. Similarly, lowering the threshold below 0.7 caused a sharp drop in cluster numbers, risking the loss of important sequence diversity. Based on these results, a 0.7 identity cutoff was chosen as the optimal balance between reducing redundancy and preserving diversity.

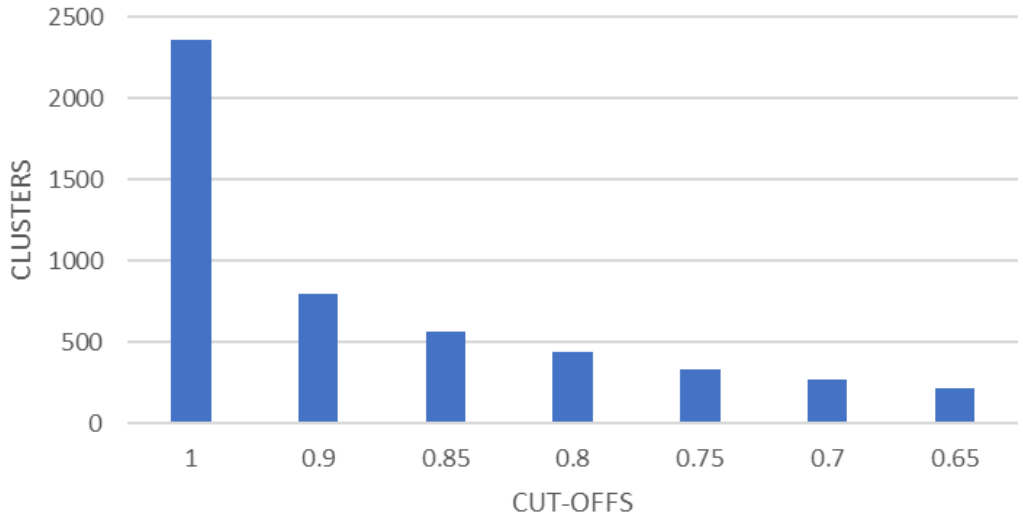


Figure 11. CD-HIT Clustering of Profilin-1 Homologs from the NR Database at Various Identity Thresholds. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

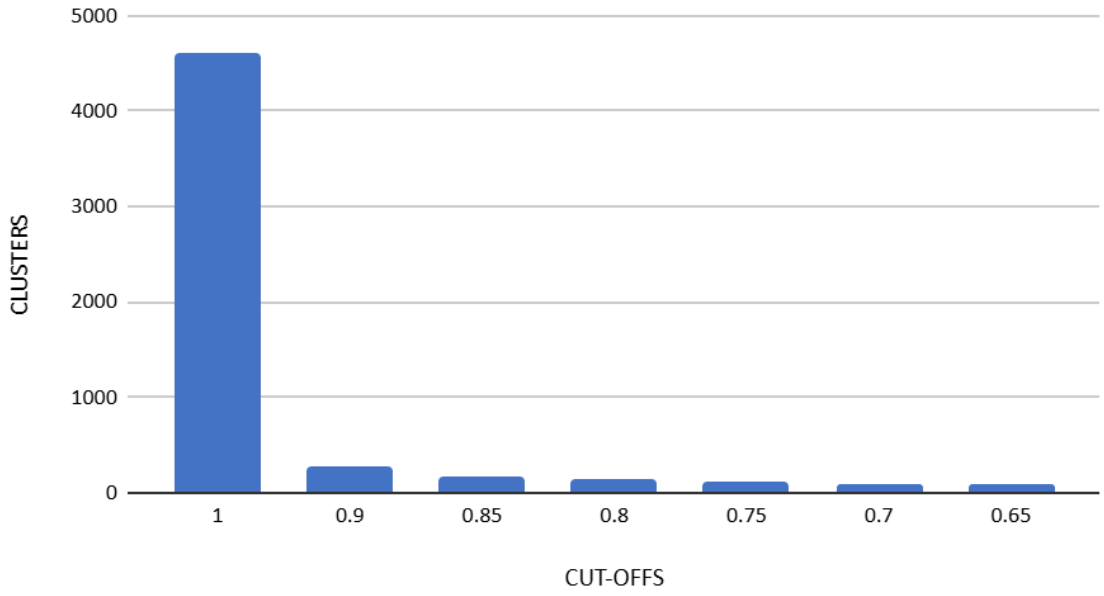


Figure 12. CD-HIT Clustering of ACTB Homologs from the NR Database at Various Identity Thresholds. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

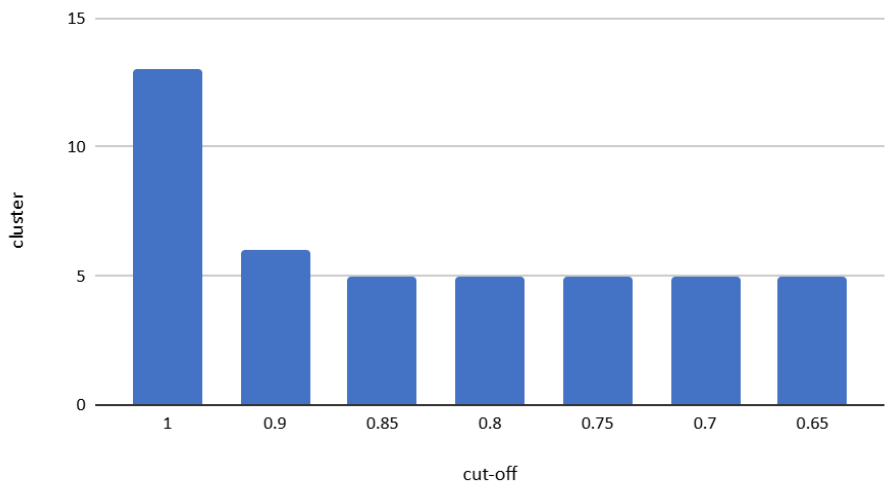


Figure 13: The plots illustrate the number of clusters across different identity cut-offs for profilin-1 protein from PDB database. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

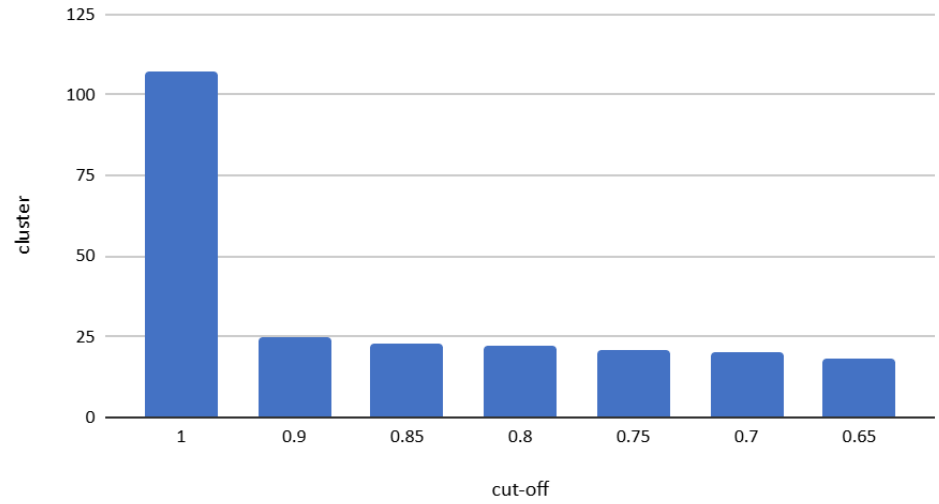


Figure 14: The plots illustrate the number of clusters across different identity cut-offs for ACTB protein from PDB database. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

Figure 11 to 14 plots illustrate the number of clusters across different identity cut-offs for profilin-1 and actin protein from NR and PDB database.

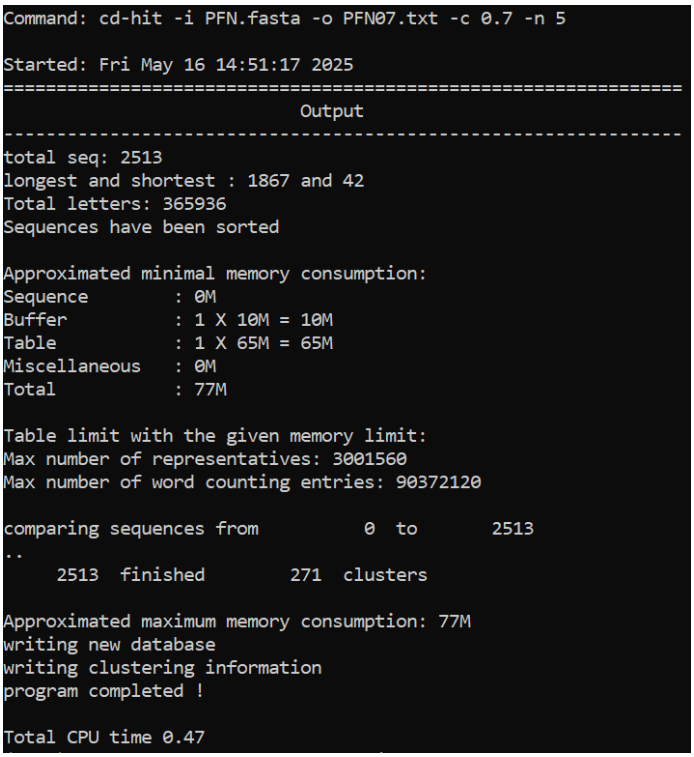


Figure 15. CD-HIT Clustering Output for Profilin-1 (NR Database) at 70% Identity Threshold

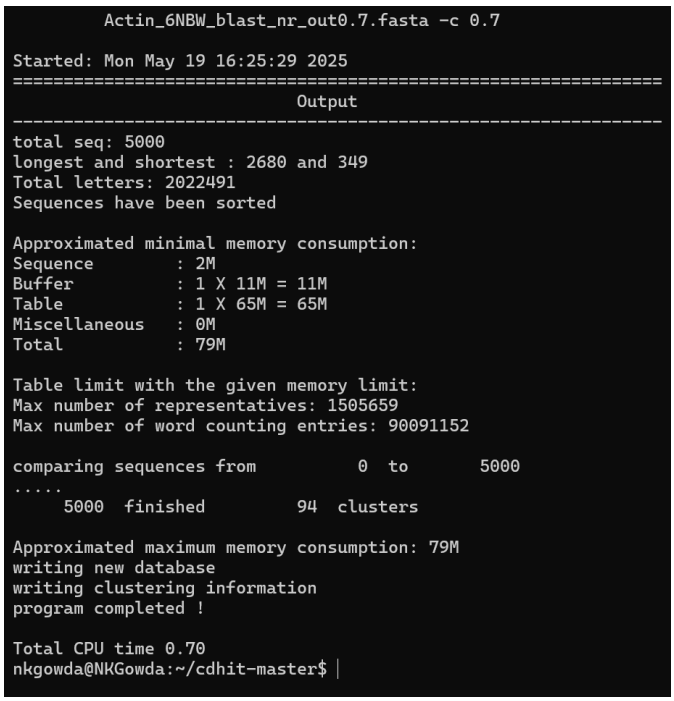


Figure 16. CD-HIT Clustering Output for Beta-Actin (NR Database) at 70% Identity Threshold.

Figure 15 and 16 shows the clustering output for profilin-1 and actin protein at 70% threshold.

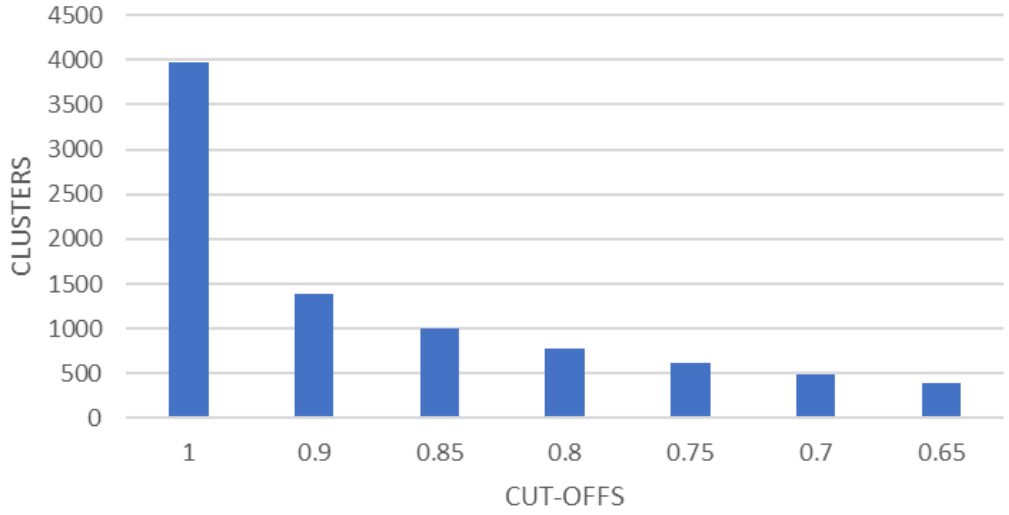


Figure 17. CD-HIT Clustering of Transthyretin Homologs from the NR Database at Various Identity Thresholds. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

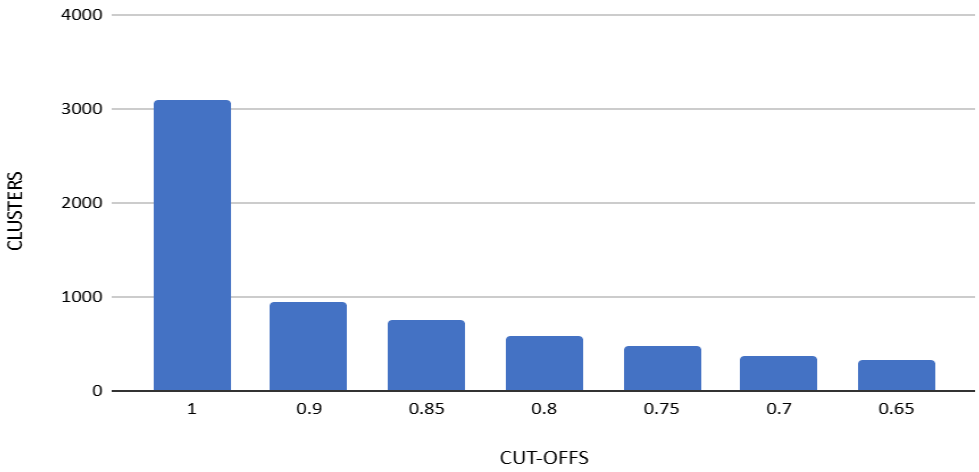


Figure 18. CD-HIT Clustering of Retinol binding protein 4 Homologs from the NR Database at Various Identity Thresholds. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

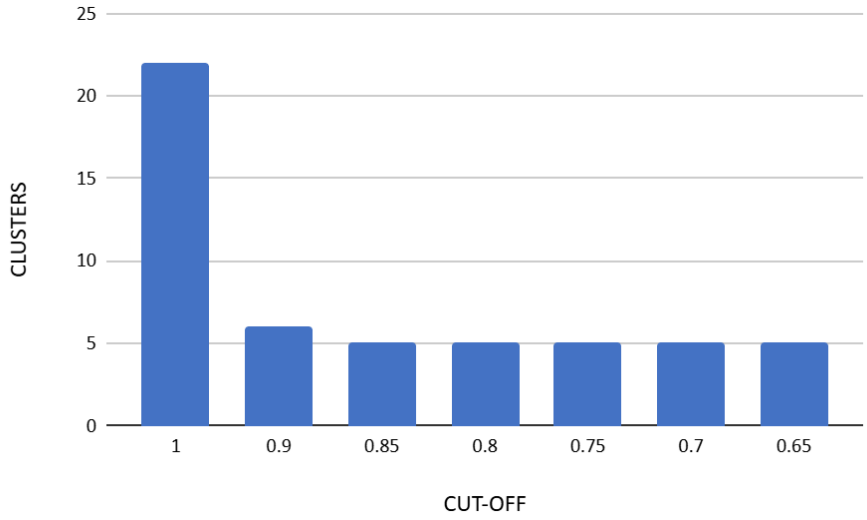


Figure 19: The plots illustrate the number of clusters across different identity cut-offs for transthyretin protein from PDB database. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

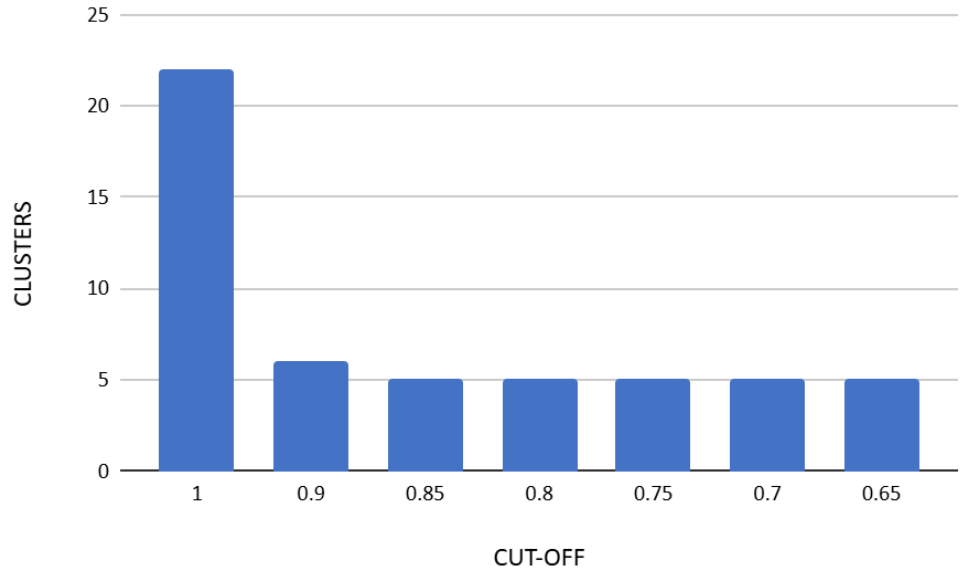


Figure 20. The plots illustrate the number of clusters across different identity cut-offs for retinol binding protein-4 protein from PDB database. The y-axis represents the number of clusters obtained, and the x-axis represents the identity cut-offs.

Figure 17 to 20 plots illustrate the number of clusters across different identity cut-offs for transthyretin and RBP4 protein from NR and PDB database.

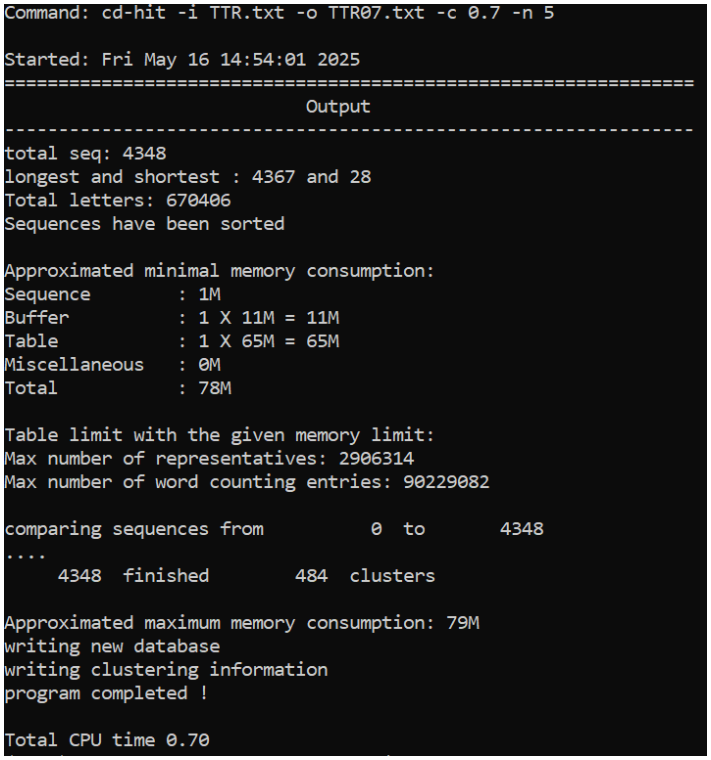


Figure 21. CD-HIT Clustering Output for Transthyretin (NR Database) at 70% Identity Threshold.

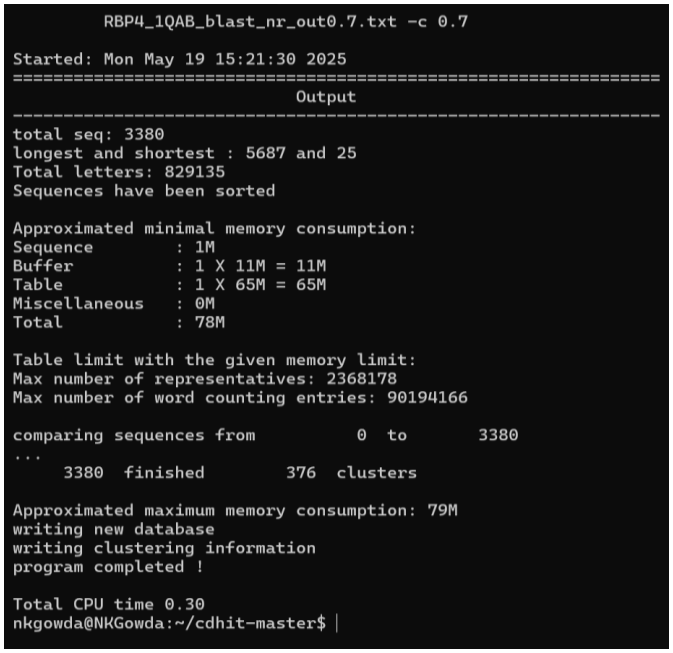


Figure 22. CD-HIT Clustering Output for RBP4 (NR Database) at 70% Identity Threshold.

Figure 21 and 22 shows the clustering output for TTR and RBP4 protein at 70% threshold.

4. Multiple sequence alignment

A comparative analysis was performed using multiple alignment tools—K-align, MAFFT, Clustal Omega, MUSCLE, T-Coffee, as well as Clustal and MUSCLE implemented in MEGA—on each protein sequence dataset clustered at a 0.7 identity threshold.

To objectively evaluate the alignment quality, the Alistat tool was employed to compute completeness scores. These scores reflect the degree of conserved alignment across the sequences, with higher values indicating better alignment accuracy and more consistently aligned regions.

Protein	Clustal	Muscle	Kalign	Mafft	T-coffee	ClustalW (MEGA)	Muscle (MEGA)
PFN-1	0.045	0.052	0.06	0.031	0.034	**0.050**	0.045
ACTB	0.049	0.074	0.063	0.02	0.04	**0.140**	0.124
TTR	0.032	0.039	0.036	0.023	0.04	**0.038**	0.037
RBP4	0.05	0.068	0.13	0.032	0.039	**0.152**	0.12

*Table 4: Completeness Scores of Multiple Sequence Alignments Evaluated Using Alistat. Scores close to **1** indicate a strong, reliable alignment, while scores near **0** point to poor alignment with many gaps or mismatches.*

Table 4 summarizes that among all the tools tested, Clustal (MEGA) consistently produced alignments with the highest completeness scores across the protein sets. The completeness score, calculated using Alistat, reflects the proportion of well-aligned, non-gap positions across sequences. A score close to **1** indicates a high-quality, complete alignment with minimal gaps, while a score near **0** suggests poor alignment and high variability or gaps. Based on this metric, Clustal (MEGA) was selected as the preferred alignment tool due to its consistently superior performance.

The **ClustalW algorithm** implemented in **MEGA** due to its flexibility in allowing fine-tuned control over key alignment parameters. Users can specify **evolutionary substitution models**, as well as adjust **gap opening and gap extension penalties**, which directly influence the alignment accuracy and sensitivity. This level of customization enables the generation of more biologically meaningful alignments, particularly when dealing with divergent sequences. The resulting alignments produced using ClustalW in MEGA demonstrated **higher completeness scores**, indicating a greater level of sequence conservation across homologous regions and a **reduced frequency of insertions and deletions (gaps)**.

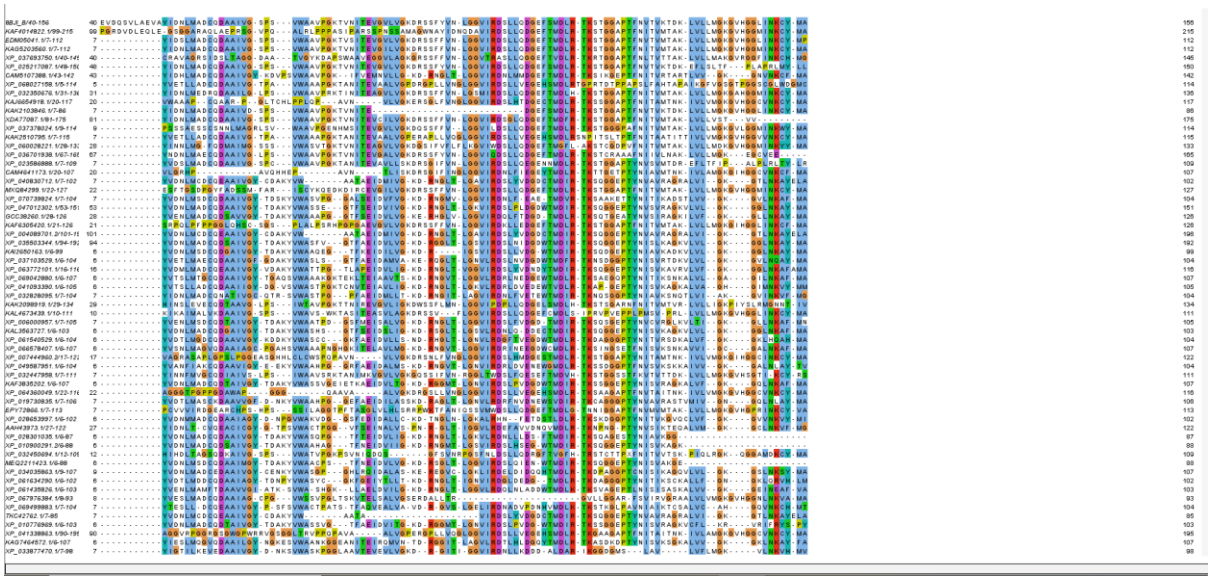


*Figure 23: Multiple Sequence Alignment of PFN-1 Using ClustalW (MEGA) Visualized in Jalview. **Blue** regions represent hydrophobic residues, **magenta** represents the negatively charged residue, **Green** represents the presence of polar residues, **Red** indicates positively charged residues, **Orange** represents glycines, **yellow** represents the proline residues.*

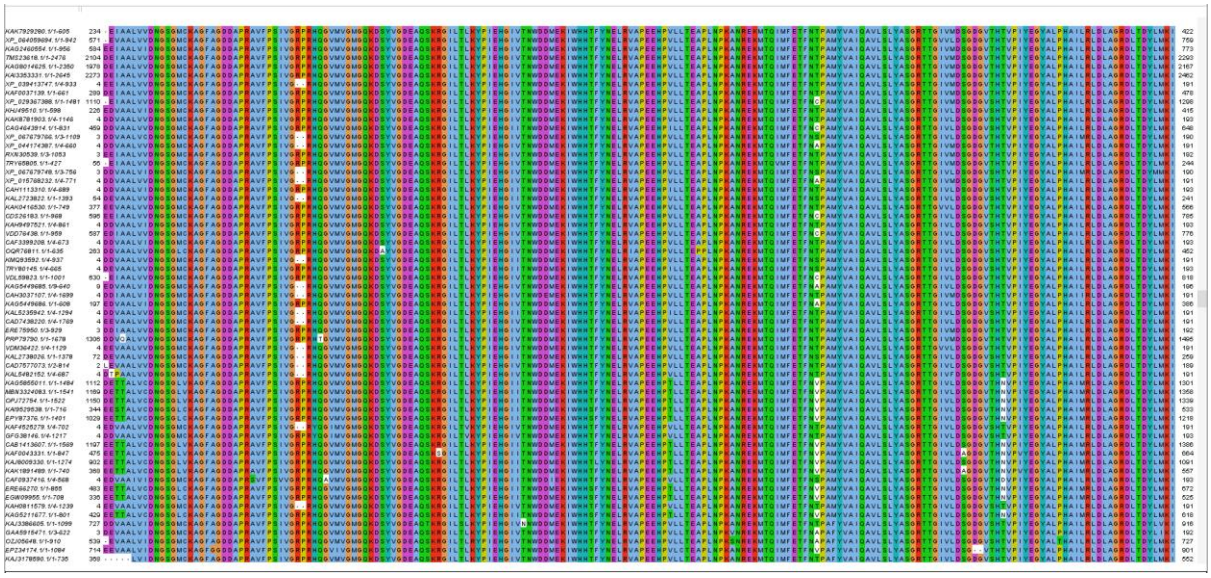


*Figure 24: Multiple Sequence Alignment of ACTB Using ClustalW (MEGA) Visualized in Jalview. **Blue** regions represent hydrophobic residues, **magenta** represents the negatively charged residue, **Green** represents the presence of polar residues, **Red** indicates positively charged residues, **Orange** represents glycines, **yellow** represents the proline residues.*

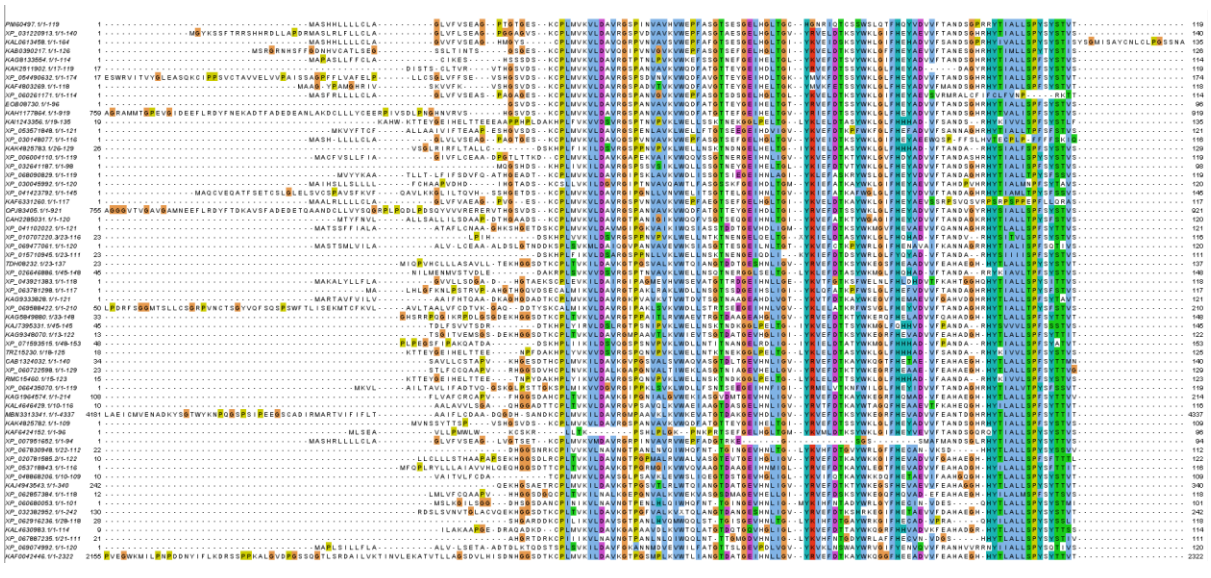


*Figure 25: Multiple Sequence Alignment of TTR Using ClustalW (MEGA) Visualized in Jalview. **Blue** regions represent hydrophobic residues, **magenta** represents the negatively charged residue, **Green** represents the presence of polar residues, **Red** indicates positively charged residues, **Orange** represents glycines, **yellow** represents the proline residues.*

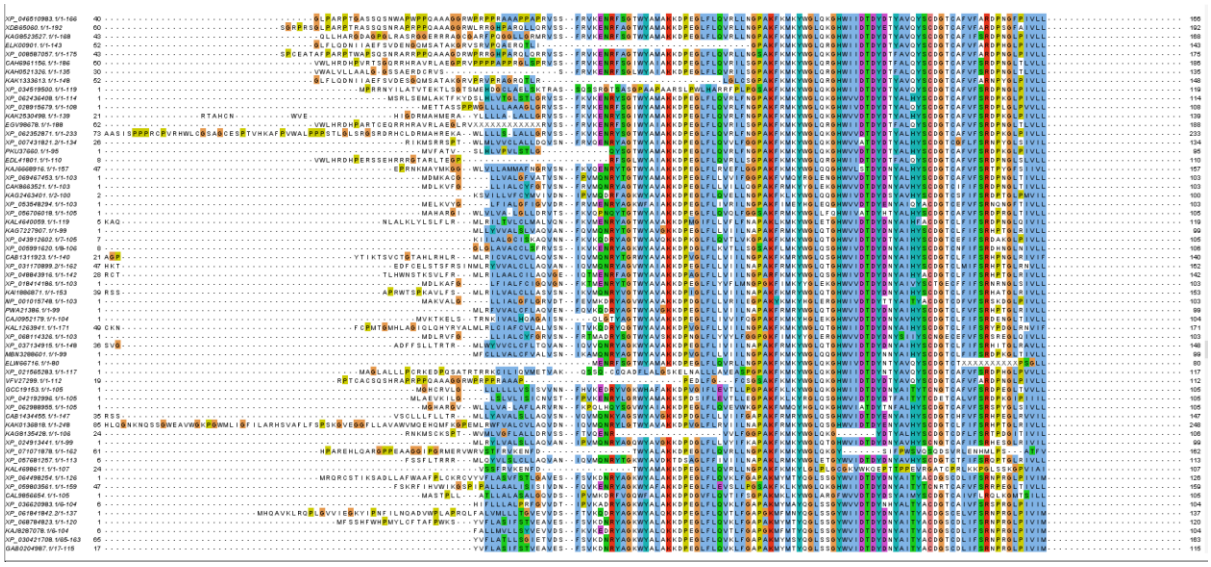


*Figure 26: Multiple Sequence Alignment of RBP4 Using ClustalW (MEGA) Visualized in Jalview. **Blue** regions represent hydrophobic residues, **magenta** represents the negatively charged residue, **Green** represents the presence of polar residues, **Red** indicates positively charged residues, **Orange** represents glycines, **yellow** represents the proline residues.*

Figure 23 to 26 is a multiple sequence alignment (MSA) results generated using the Clustal algorithm in MEGA and visualized in Jalview, the sequences of all four proteins PFN-1, Actin, TTR, and RBP4 were color-coded using the Clustal format to highlight residue conservation and chemical similarity. Across the alignments, certain colors appeared consistently, indicating conserved residue types with potential functional or structural importance. **Blue** was frequently observed, representing conserved **charged residues** such as Aspartate (D), Glutamate (E), Lysine (K), Arginine (R), and Histidine (H), which are crucial for electrostatic interactions and active sites. **Magenta** highlighted conserved **Cysteines (C)**, which may be involved in disulfide bond formation, contributing to protein stability. **Green** regions indicated the presence of **Serine (S)** and **Threonine (T)**, polar residues commonly involved in phosphorylation or hydrogen bonding. **Red** residues denoted **aromatic amino acids** such as Phenylalanine (F), Tyrosine (Y), and Tryptophan (W), often involved in stacking interactions or ligand binding. **Orange or brown** represented **Glycine (G)** and **Proline (P)**—unique in terms of structure, with glycine offering flexibility and proline introducing bends or constraints in the chain. **Yellow** highlighted conserved **aliphatic residues** like Isoleucine (I), Leucine (L), Valine (V), and Methionine (M), which are hydrophobic and often buried in the protein core. These consistent color patterns suggest the preservation of critical residues necessary for structural integrity or biological activity across homologous sequences in each protein family. The Clustal coloring scheme thus provided a

visually intuitive and informative representation of sequence conservation, aiding in the identification of potentially important residues for further functional or evolutionary analysis.

5. Multiple structure alignment

Multiple structure alignment using Dali, TM-align, and PROMALS3D revealed valuable insights into the similarities among the selected protein structures. Dali generated a heat map and dendrogram that visually represented pairwise structural similarities, clustering related proteins based on their 3D conformations. TM-align provided quantitative measures such as TM-scores and RMSD values, allowing for precise comparison of structural alignments and the extent of structural conservation. PROMALS3D, which incorporates both sequence and structural data, produced a structure-guided multiple sequence alignment and enabled superimposition of protein structures, highlighting conserved residues and regions important for functional or evolutionary significance. Together, these tools offered complementary perspectives, enhancing the reliability and interpretability of the structural alignment results.

5.1. Dali Results

A dendrogram generated from the DALI server provides a hierarchical representation of structural similarity among multiple protein structures. It is constructed based on Z-scores derived from pairwise structural alignments. The Z-score is a statistical measure indicating how similar two protein structures are, with values greater than 2 generally considered significant. Higher Z-scores suggest a strong structural resemblance, and such protein pairs are clustered closely in the dendrogram with shorter branch lengths. Conversely, lower Z-scores indicate weaker similarity and result in longer branch distances.

The heatmaps generated by the DALI server provide a visual matrix of pairwise structural similarities among the selected protein structures, using **Z-scores** as the comparative metric. Each cell in the heatmap represents the Z-score between two protein structures, with the color intensity indicating the level of similarity **darker shades (such as red or orange)** reflect higher Z-scores and thus stronger structural resemblance, whereas **lighter shades (such as yellow or white)** indicate lower similarity.

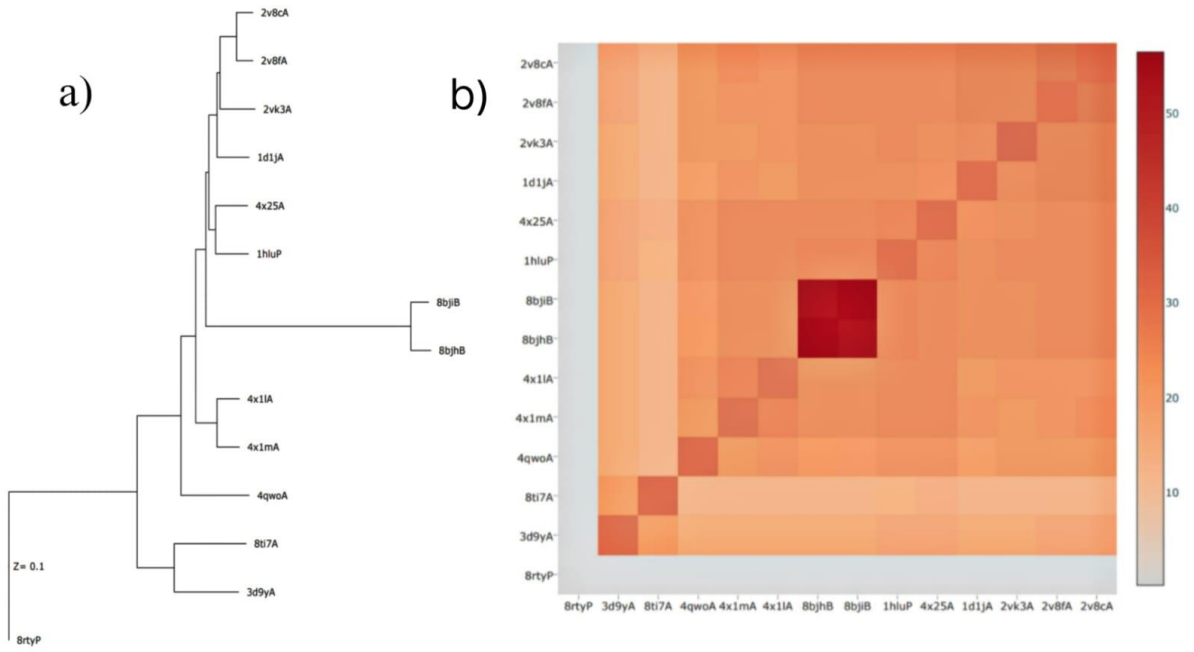


Figure 27: a) Dendrogram of PFN-1 protein b) heatmap of PFN-1 protein representing multiple structural alignment from DALI Server.

Figure 27a dendogram, the topmost cluster includes **2v8cA** *Mouse Profilin* and 2v8fA *Mouse Profilin IIa* are clustered in same branch indicating a high degree structural conservation and they belong to same species and are isoforms.

2vk3A *profilin2a***, 1d1jA** *human profilin 2***, 1hluP** *profilin (bos taurus***)** which are clustered closely on the dendrogram. This suggests that these three proteins exhibit a high degree of structural conservation, likely sharing similar folds or domains. The tight clustering indicates minimal deviation in their three-dimensional conformations.

Proteins like **8bjiB** and **8bjbA** form more distant branches, implying comparatively lower structural similarity to profilin-1.

At the base of the dendrogram lies **8rtyP** *F-actin and profilin complex*, which appears as the most structurally distinct protein in the dataset. Its early branching indicates that it has the least structural similarity to the other proteins analyzed in this alignment.

The figure 27b, heatmap demonstrates a significant cluster of high similarity scores centered around a subset of entries including **2v8cA**, **2v8fA**, **1hluP** and **2vk3A** suggesting that these structures share highly conserved three-dimensional conformations. Interestingly, **4x1mA** and **4x1lA** also show relatively strong alignment with the central cluster, indicating a

potential functional or evolutionary linkage despite possible divergence at the sequence level. These closely related variants represent isoforms or structurally resolved complexes of Profilin-1 under similar functional or conformational states.

In contrast, entries like **8bjiB**, **8bjhB**, and **8rtyP**, exhibit relatively lower Z-scores in comparison to the central cluster, implying moderate to low structural conservation. This variation is attributed to species-specific differences, conformational flexibility, or differences in bound ligands or interaction partners influencing structural deviations.

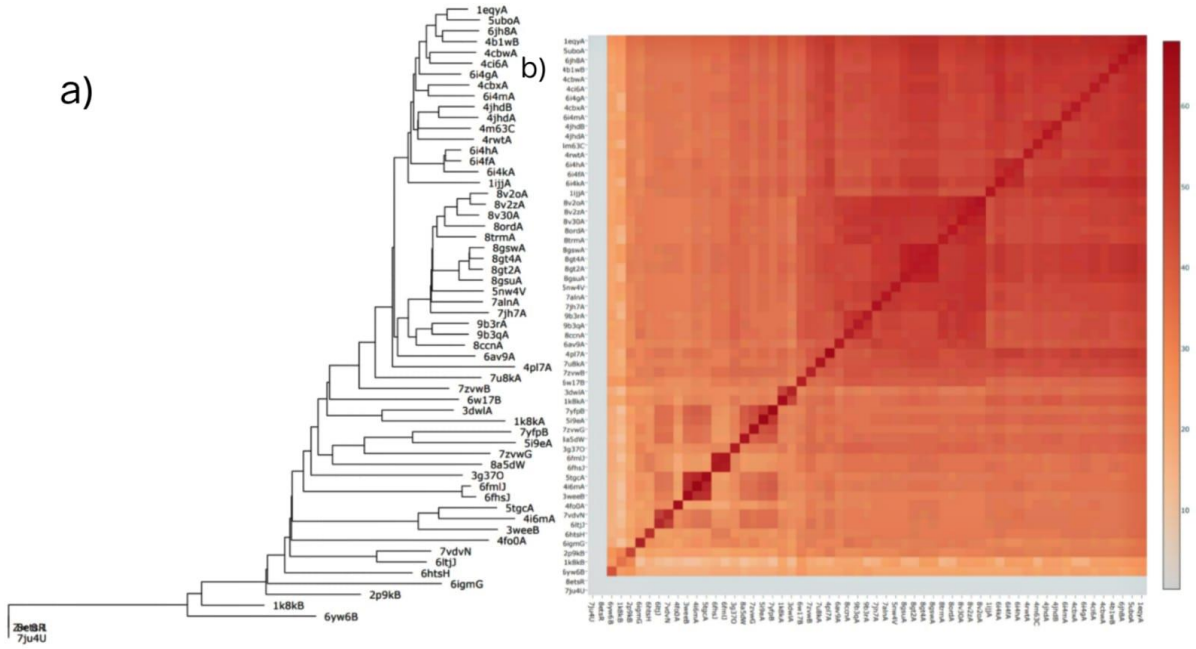


Figure 28: a) Dendrogram of ACTB protein b) heatmap of ACTB protein representing multiple structural alignment from DALI Server.

In figure 28a, dendogram, **1eqyA** *complex of rabbit muscle alpha-actin; human gelosin and* to **1ijj** *complex of rabbit skeletal muscle actin and latrunculin* clusters relatively deeply within a branch containing multiple actin-related proteins, including isoforms and homologous cytoskeletal proteins such as suggesting strong structural conservation among this subgroup.

8v2oA smooth muscle gamma-actin to 6av9A mical oxidized actin is another closely clustered proteins which formed a distant branch indicating divergence.

Notably, protein like **7ju4U** appear as outliers, branching distantly from the main cluster. This indicates significant structural divergence from canonical beta-actin, which may reflect functional specialization or evolutionary variation.

The figure 28b, heatmap of **actin** and its homologs, as generated by the DALI server, reveals a complex pattern of structural conservation and divergence among a large set of proteins. The diagonal line of deep red squares indicates perfect or near-perfect structural matches, representing self-alignments or very closely related isoforms. The surrounding gradient, from red to orange, reflects varying degrees of structural similarity based on Z-scores, with darker tones denoting higher structural conservation.

A prominent dense cluster of high similarity scores is observed in the central region of the heatmap, indicating a well-conserved core set of actin proteins and close homologs. These proteins, despite being from different organisms or structural states share a remarkably preserved tertiary structure, characteristic of the actin fold.

Notably, outside the central core, several proteins display intermediate Z-scores, suggesting partial structural similarity, possibly corresponding to actin-related proteins or divergent homologs. These variations might reflect adaptations for specialized functions while retaining core actin features. The peripheral and lower-similarity zones, shaded in lighter orange, likely represent more distantly related structural analogs or proteins with only partial domain-level homology.

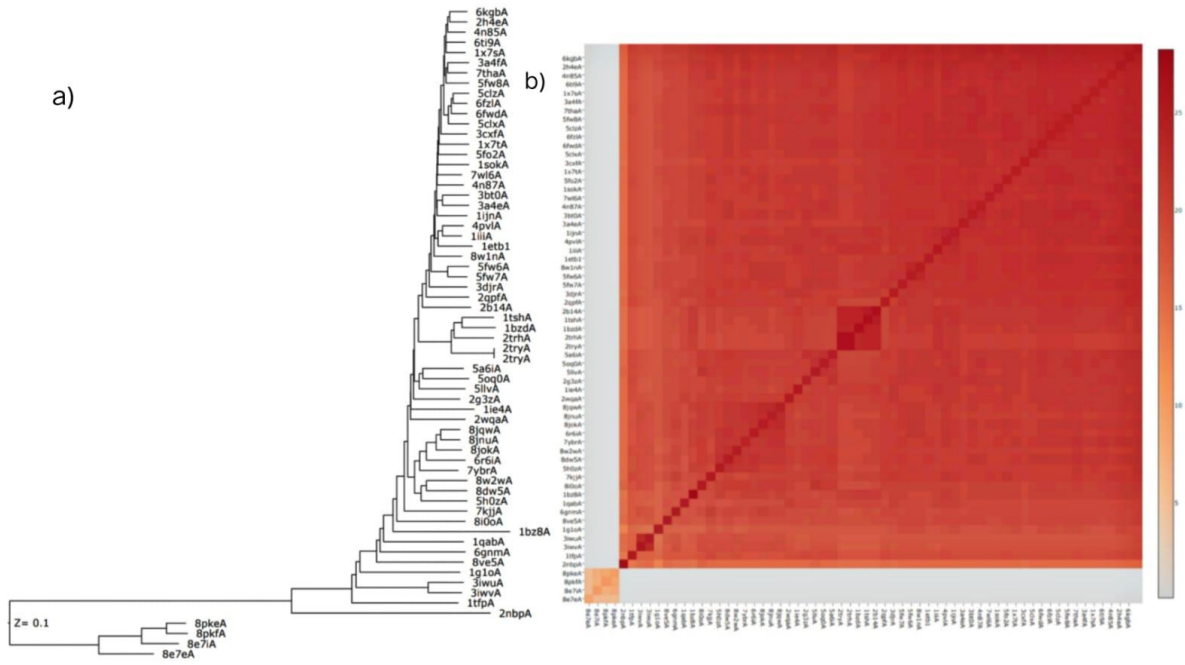


Figure 29: a) Dendrogram of TTR protein b) heatmap of TTR protein representing multiple structural alignment from DALI Server.

In figure 29a, dendrogram, compact cluster of proteins such from **6kgbA** *mutated transthyretin* to **2nbpA** *transthyretin in its monomeric state*, which suggests a high degree of structural conservation. These are likely homologous transthyretin isoforms or evolutionarily related proteins from different species, sharing the characteristic β-sheet-rich fold of transthyretin.

Proteins like **1tshA, 2tryA** and **1bz8A** *amyloidogenic transthyretin variants* are in the close clusters yet distant indicating the divergence from transthyretin

Additionally, there are distinct outliers, such as **8pkeA** *fibril*, **8pkfA** *amyloid fibril*, **8e7iA** *fibril* and **8e7eA**, which form distant branches on the dendrogram. These proteins show substantial structural divergence from transthyretin.

The figure 29b, heatmap for **transthyretin (TTR)** and its homologs provides a detailed visualization of structural conservation across a broad set of related protein structures. The color spectrum indicates Z-score values, with warmer shades (dark red) representing higher structural similarity. As expected, a distinct diagonal pattern indicates self-alignments and

highly similar entries, reflecting structural redundancy or isoforms of TTR under different conditions or mutations.

A notable cluster with deep red intensity is observed near the center of the heatmap, denoting a group of structures with high structural similarity, likely consisting of wild-type TTR and its close variants, including mutant forms associated with disease states like familial amyloid polyneuropathy (FAP). The tight clustering and strong Z-scores reflect the conserved beta-sheet-rich tertiary structure that is characteristic of TTR, underscoring its evolutionary and functional stability as a thyroid hormone and retinol transporter.

Beyond the central cluster, moderately high Z-score values are observed throughout the matrix, indicative of structural conservation among distant homologs or proteins with similar beta-sheet arrangements, such as other members of the transthyretin-like or lipocalin superfamily. The relatively fewer light orange cells suggest that most of the aligned proteins maintain a basic structural framework similar to TTR, though there is some variation likely due to sequence divergence or altered quaternary states.

Interestingly, a small corner of the heatmap (lower-left) shows a group of structures with distinctly lower Z-scores compared to the rest of the dataset. This deviation may represent distant structural analogs or functionally divergent proteins with only partial domain-level similarity.

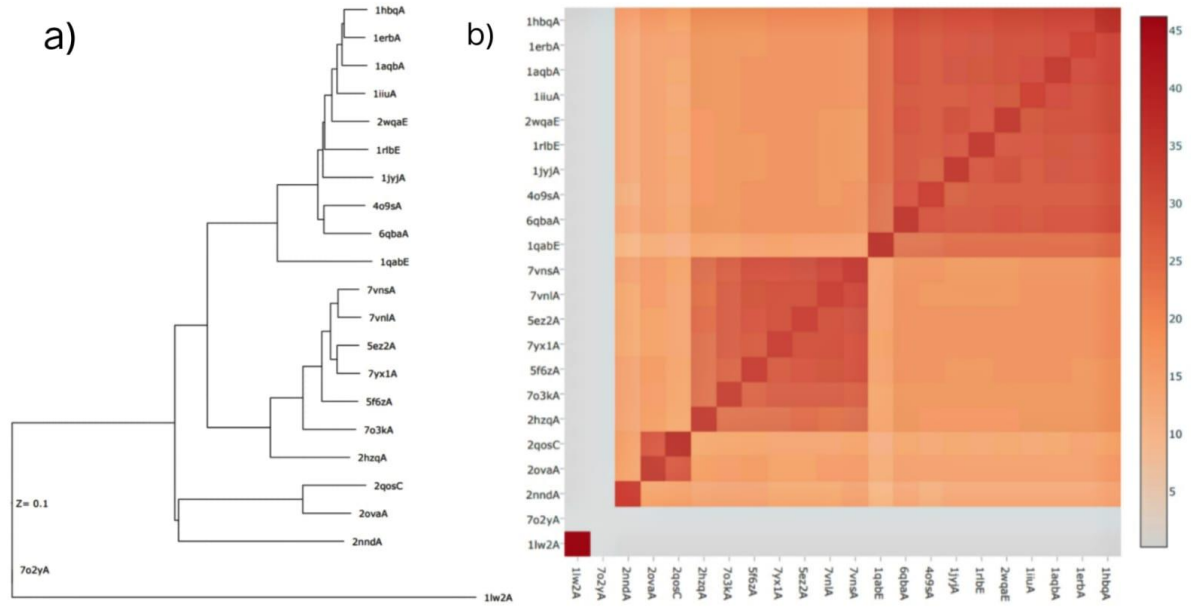


Figure 30: a) Dendrogram of RBP4 protein b) heatmap of RBP4 protein representing multiple structural alignment from DALI Server.

In figure 30a, dendrogram, **6qbaA** *complex of RBP4 and non-retinoid ligand*, **4o9sA** *complex of RBP4 and non-retinoid ligand*, and **2nnd** *functional site of RBP* clusters closely, indicating a high degree of structural conservation among members of the lipocalin family, to which RBP4 belongs. This close clustering confirms that these proteins share a conserved β-barrel fold that characterizes lipocalins, which are known for their ability to transport small hydrophobic molecules such as retinoids.

Notably, **1lw2A** and **7o2yA** appear as significant outliers, indicating lower structural similarity with the RBP4 core cluster. These proteins may share only minimal fold similarity, possibly representing distant homologs or analogs that have evolved similar structural motifs independently.

The figure 30b, heatmap generated using the DALI server for **Retinol Binding Protein 4 (RBP4)** and its homologs presents a structured overview of pairwise structural similarities. The diagonal line of deep red cells reflects perfect or near-perfect matches, as expected from self-alignments or nearly identical structures. The Z-score scale on the right denotes

increasing structural similarity from light orange (low similarity) to dark red (high similarity), enabling interpretation of evolutionary and structural conservation.

Two distinct clusters are evident in the heatmap. The upper right block primarily consists of classical RBP4 entries and closely related members, showing consistently high Z-scores, indicative of a conserved β-barrel structure typical of lipocalins. The presence of deep red shading across this cluster underscores the strong conservation of the retinol-binding fold across species or conditions.

In contrast, the lower left portion of the heatmap forms a second block of structures, displaying relatively strong internal similarity but less pronounced similarity to the main RBP4 cluster. This suggests that the entries in this cluster likely represent structurally related but functionally divergent proteins, such as other members of the lipocalin or transport protein families. The transitional zone between the two clusters displays moderate Z-scores, reflecting partial conservation, possibly restricted to domain-level or fold-level similarity rather than full-length alignment.

An outlier, **1lw2A**, shows minimal similarity with the rest, seen as a lighter zone and lack of integration into either cluster. This may represent either a structurally divergent analog or an incorrectly aligned entry, and would merit further manual verification.

5.2. TM - align Results

The structural alignments were performed using **TM-align**, and the resulting matrix provides both **TM-scores** a measure of structural similarity and **RMSD values** a measure of atomic deviation upon alignment. TM-scores range from 0 to 1, where a score above 0.5 generally indicates significant structural similarity, while RMSD values provide insights into the degree of structural divergence lower values indicate better alignment.

PFN-1	8BJI_B	8BJH_B	4QWO_A	8IT7_A	2VK3_A	4X1L_A	4X25_A	4X1M_A	1HLU_P	2V8C_A	2V8F_A	2RTY_B	1D1J_A	3D9Y_A
8BJI_B	NIL	0.99266/ 0.62	0.18205/ 4.31	0.19540/ 4.54	0.20336/ 4.72	0.17481/ 4.49	0.18919/ 4.90	0.18602/ 4.10	0.95650/ 1.52	0.19679/ 4.64	0.19529/ 4.64	0.05011/ 3.22	0.19941/ 4.60	0.17352/ 5.21
8BJH_B	0.99266/ 0.62	NIL	0.18138/ 4.37	0.19690/ 4.46	0.20357/ 4.71	0.17770/ 4.80	0.18700/ 4.68	0.18729/ 4.12	0.95823/ 1.51	0.19853/ 4.52	0.19751/ 4.82	0.05405/ 3.01	0.19961/ 4.46	0.17433/ 5.27
4QWO_A	0.39393/ 4.31	0.39120/ 4.37	NIL	0.77376/ 2.27	0.92516/ 1.31	0.91334/ 0.97	0.92859/ 1.06	0.90839/ 1.18	0.41229/ 4.40	0.92413/ 1.30	0.91533/ 1.43	0.11179/ 2.31	0.91365/ 1.31	0.79288/ 1.79
8IT7_A	0.41843/ 4.54	0.42216/ 4.46	0.76380/ 2.27	NIL	0.78073/ 2.47	0.76865/ 2.15	0.80316/ 2.23	0.76738/ 2.24	0.42133/ 4.39	0.79569/ 2.25	0.79024/ 2.25	0.15666/ 2.94	0.78772/ 2.18	0.87398/ 1.40
2VK3_A	0.41438/ 4.72	0.41522/ 4.71	0.85200/ 1.31	0.73444/ 2.47	NIL	0.85730/ 0.99	0.90055/ 1.05	0.86229/ 1.01	0.41299/ 4.82	0.95095/ 0.80	0.94447/ 0.79	0.16788/ 2.10	0.93945/ 0.87	0.76019/ 1.84
4X1L_A	0.39031/ 4.49	0.39282/ 4.80	0.93404/ 0.97	0.79479/ 2.15	0.95415/ 0.99	NIL	0.98063/ 0.72	0.96958/ 0.44	0.40161/ 4.49	0.96607/ 0.83	0.95175/ 0.91	0.11853/ 2.36	0.94611/ 0.98	0.82181/ 1.64
4X25_A	0.39564/ 4.90	0.38851/ 4.68	0.90212/ 1.06	0.79272/ 2.23	0.95191/ 1.05	0.93022/ 0.72	NIL	0.92044/ 0.70	0.40968/ 4.59	0.96493/ 0.87	0.95073/ 0.99	0.10785/ 0.90	0.94075/ 1.15	0.80378/ 1.82
4X1M_A	0.41686/ 4.10	0.41751/ 4.12	0.92189/ 1.18	0.78801/ 2.24	0.95235/ 1.01	0.96205/ 0.44	0.96269/ 0.70	NIL	0.41094/ 4.18	0.96426/ 0.87	0.95224/ 0.92	0.11802/ 2.33	0.95462/ 0.98	0.81797/ 1.68
1HLU_P	0.93912/ 1.52	0.94079/ 1.51	0.18554/ 4.40	0.19309/ 4.39	0.19920/ 4.82	0.17662/ 4.49	0.18986/ 4.59	0.18335/ 4.18	NIL	0.19413/ 4.71	0.19581/ 4.71	0.05254/ 2.67	0.19526/ 4.39	0.16127/ 4.35
2V8C_A	0.41350/ 4.64	0.41350/ 4.52	0.86782/ 1.30	0.76113/ 2.25	0.97091/ 0.80	0.88523/ 0.83	0.93137/ 0.87	0.89036/ 0.87	0.40618/ 4.71	NIL	0.98011/ 0.56	0.10993/ 1.98	0.96352/ 0.99	0.78590/ 1.69
2V8F_A	0.40984/ 4.64	0.40912/ 4.82	0.86545/ 1.43	0.76061/ 2.25	0.97109/ 0.79	0.87868/ 0.91	0.92423/ 0.99	0.88575/ 0.92	0.41279/ 4.71	0.98714/ 0.56	NIL	0.10956/ 2.03	0.96728/ 0.85	0.77974/ 1.77
2RTY_B	0.41092/ 3.22	0.42234/ 3.01	0.37660/ 2.31	0.36348/ 2.94	0.41380/ 2.10	0.43580/ 2.36	0.41117/ 0.90	0.44205/ 2.33	0.41939/ 2.67	0.43794/ 1.98	0.44669/ 2.03	NIL	0.43048/ 2.27	0.30648/ 2.68
1D1J_A	0.41836/ 4.60	0.42025/ 4.46	0.86403/ 1.31	0.75839/ 2.18	0.96580/ 0.87	0.87382/ 0.98	0.91473/ 1.15	0.88824/ 0.98	0.41638/ 4.39	0.97033/ 0.99	0.96728/ 0.85	0.11018/ 2.27	NIL	0.78674/ 1.70
3D9Y_A	0.36517/ 5.21	0.36598/ 5.27	0.81505/ 1.79	0.91265/ 1.40	0.84465/ 1.84	0.82768/ 1.64	0.84884/ 1.82	0.82961/ 1.68	0.38750/ 4.35	0.85749/ 1.69	0.84518/ 1.77	0.15955/ 2.68	0.85337/ 1.70	NIL

Table 5: Pairwise TM-align structural similarity matrix of 15 representative PFN-1 protein structures. Each cell displays the TM-score (a measure of structural similarity) followed by the RMSD value (in Å) separated by a slash.

Table 5, resulting matrix includes TM-scores a measure of structural similarity and RMSD values (in Å), with TM-scores above 0.5 generally indicating significant similarity. Several protein pairs exhibited TM-scores greater than 0.9 and low RMSD values <1.2 Å, such as between 2V8F_A (*Profilin-2*) and 2V8C_A (*Mouse Profilin*) 0.98011 / 0.45 and 4X1M_A (*Profilin-1*) and 4X11_A (*Major capsid protein*) 0.96269 / 0.44, indicating strong conservation of the Profilin-1 fold. The majority of other comparisons yielded TM-scores ranging from 0.75 to 0.96 and RMSD values between 0.7 and 2.5 Å, further supporting the structural stability of Profilin-1 across various conformational states and crystallographic models. However, relatively lower TM-scores ~0.36–0.42 and higher RMSDs >4.5 Å were observed in alignments involving 3D9Y_A (*Profilin*), 8IT7_A (Phosphoglycerate mutase 1), and 4QWO_A (Profilin-like Protein) with others, such as 3D9Y_A vs 8BJI_B (actin-profilin) 0.36517 / 5.21, suggesting potential structural divergence or flexible regions. Overall, the TM-align analysis demonstrates that while Profilin-1 structures are largely conserved across homologs, a few entries exhibit conformational variation, possibly due to experimental conditions or ligand-bound states.

ACTIN	6FML_J	6FHS_J	3DWL_A	6HTS_H	4FO0_A	7YFP_A	519E_A	5TGC_A	1KBK_A	416M_A	8A5D_W	7ZVW_G	4PL7_A	4JHD_A	7VDV_N
6FML_J	NIL	0.99989 / 0.08	0.22086 / 6.49	0.85760 / 3.41	0.20924 / 7.20	0.24949 / 7.20	0.22767 / 7.24	0.22882 / 7.28	0.21581 / 7.48	0.22654 / 6.75	0.26029 / 2.95	0.07102 / 7.71	0.21418 / 7.35	0.21745 / 7.51	0.46494 / 3.65
6FHS_J	0.99989 / 0.08	NIL	0.21872 / 6.97	0.85806 / 3.40	0.20743 / 7.75	0.24940 / 6.92	0.23104 / 7.88	0.23179 / 6.67	0.21845 / 7.44	0.22224 / 6.73	0.26059 / 2.98	0.07104 / 7.70	0.19917 / 7.34	0.21882 / 7.26	0.46562 / 3.48
3DWL_A	0.26231 / 6.49	0.25845 / 6.97	NIL	0.27365 / 7.44	0.20946 / 7.82	0.82806 / 2.63	0.69953 / 3.33	0.70279 / 3.09	0.96946 / 1.15	0.72135 / 3.39	0.25979 / 4.90	0.05635 / 6.99	0.75081 / 2.80	0.84116 / 2.31	0.35985 / 4.87
6HTS_H	0.83110 / 3.41	0.83155 / 3.40	0.27365 / 7.44	NIL	0.23182 / 7.82	0.26693 / 7.63	0.23734 / 7.79	0.25030 / 7.06	0.24038 / 7.53	0.24301 / 6.97	0.24929 / 3.01	0.07255 / 7.45	0.23867 / 7.38	0.24215 / 7.46	0.47248 / 3.51
4FO0_A	0.19295 / 7.20	0.19177 / 7.75	0.79854 / 3.35	0.20946 / 7.82	NIL	0.89092 / 2.38	0.81839 / 2.73	0.78473 / 2.82	0.64533 / 3.09	0.81079 / 2.85	0.27714 / 2.85	0.06810 / 7.61	0.83418 / 1.75	0.89728 / 2.32	0.42431 / 4.30
7YFP_B	0.21903 / 7.20	0.21777 / 6.92	0.86452 / 2.63	0.24034 / 7.63	0.67061 / 2.38	NIL	0.90305 / 2.14	0.85474 / 2.37	0.90074 / 2.40	0.84711 / 2.60	0.29483 / 2.37	0.05578 / 6.92	0.89513 / 1.42	0.96597 / 1.27	0.14078 / 4.75
519E_A	0.23826 / 7.24	0.24135 / 7.88	0.81344 / 3.33	0.24240 / 7.79	0.69314 / 2.73	0.80342 / 2.14	NIL	0.83659 / 2.27	0.75468 / 2.92	0.80099 / 2.41	0.23819 / 4.35	0.05477 / 7.12	0.76351 / 2.06	0.79760 / 2.17	0.47319 / 3.64
5TGC_A	0.24514 / 7.28	0.24917 / 6.67	0.79406 / 3.09	0.26193 / 7.06	0.64508 / 2.82	0.78545 / 2.37	0.81004 / 2.27	NIL	0.75244 / 3.17	0.92066 / 0.88	0.34429 / 3.36	0.05819 / 7.09	0.77960 / 2.42	0.77858 / 2.51	0.36695 / 4.76
1KBK_A	0.23348 / 7.48	0.23662 / 7.44	0.86029 / 1.15	0.25378 / 7.53	0.79216 / 3.09	0.83928 / 2.40	0.78889 / 2.92	0.76270 / 3.17	NIL	0.77508 / 3.33	0.32399 / 3.19	0.05715 / 6.71	0.82399 / 2.64	0.91263 / 2.23	0.12062 / 3.88
416M_A	0.25170 / 6.75	0.24676 / 6.73	0.78029 / 3.39	0.26323 / 6.97	0.63833 / 2.85	0.81462 / 2.60	0.86556 / 2.41	0.96535 / 0.88	0.75234 / 3.33	NIL	0.30313 / 4.21	0.05762 / 7.13	0.81367 / 2.38	0.84225 / 2.65	0.38213 / 4.40
8A5D_W	0.06206 / 2.95	0.06200 / 2.98	0.09505 / 4.90	0.06679 / 3.01	0.05475 / 2.85	0.08314 / 2.37	0.06822 / 4.35	0.0922 / 3.36	0.08679 / 3.19	0.08968 / 4.21	NIL	0.46340 / 4.33	0.26970 / 4.76	0.08328 / 4.58	0.33382 / 1.59
7ZVW_G	0.34961 / 7.71	0.35019 / 7.70	0.35076 / 6.99	0.37140 / 7.45	0.29423 / 7.61	0.34014 / 6.92	0.29647 / 7.12	0.31837 / 7.09	0.32556 / 6.71	0.32831 / 7.13	0.02099 / 4.33	NIL	0.05502 / 6.46	0.05470 / 6.61	0.02106 / 4.21
4PL7_A	0.23967 / 7.35	0.22286 / 7.34	0.80215 / 2.80	0.26058 / 7.38	0.64116 / 1.75	0.87241 / 1.42	0.83738 / 2.06	0.82763 / 2.42	0.78790 / 2.64	0.80185 / 2.38	0.08848 / 4.76	0.33289 / 6.46	NIL	0.87760 / 1.32	0.48172 / 3.72
4JHD_A	0.24741 / 7.51	0.25045 / 7.26	0.87644 / 2.31	0.26989 / 7.46	0.67310 / 2.32	0.96851 / 1.27	0.89866 / 2.17	0.84914 / 2.51	0.84760 / 2.23	0.80787 / 2.65	0.23855 / 4.58	0.33425 / 6.61	0.90295 / 1.32	NIL	0.45925 / 3.41
7VDV_N	0.11462 / 3.65	0.11407 / 3.48	0.13505 / 4.87	0.11797 / 3.51	0.11141 / 4.30	0.41132 / 4.75	0.14240 / 3.64	0.12447 / 4.76	0.42719 / 3.88	0.12839 / 4.48	0.33382 / 1.59	0.47670 / 4.21	0.15495 / 3.72	0.14842 / 3.41	NIL

Table 6: Pairwise TM-align structural similarity matrix of 15 representative ACTB related protein structures. Each cell displays the TM-score (a measure of structural similarity) followed by the RMSD value (in Å) separated by a slash.

Table 6, assess the structural conservation among actin and actin-related proteins, TM-align was used to perform pairwise structural alignments on a selected set of 15 representative PDB structures. The resulting matrix provided TM-scores, which reflect the degree of structural similarity values >0.5 indicate significant similarity, and RMSD values, which measure the average atomic deviation upon superposition. The analysis revealed that a majority of the protein pairs exhibited high TM-scores ≥0.7 and low RMSD values, indicating a strong degree of structural conservation across these proteins. Notable examples include the alignments between **6FML_J** (*Actin related protein 5*) and **6FHS_J** (*Actin related protein 5*) (TM-score: 0.99989), **3DWL_A** (*Actin related protein 3*) and **6HTS_H** (*Actin related protein 5*) (0.85760), and **4PL7_A** (*Actin, Thymosin beta-4*) and **4JHD_A** (*Actin-5C*) (0.96957), highlighting conserved core architecture among actin family members. In contrast, a few protein pairs showed low TM-scores <0.3 and higher RMSD values >4 Å, such as 8A5D_W (*Actin related protein 4*) aligned with **3DWL_A** (*Actin related protein 3*) or **4FO0_A** (*Actin related protein 8*), suggesting structural divergence likely due to functional specialization or differences in domain organization. Overall, the TM-align results underscore a strong evolutionary conservation of actin-like fold architecture while also reflecting structural variability among functionally distinct members.

TTR	1bz8	1bzd	1etb	1g10	1ie4	1ii	1jp	1oo2	1qab	1sok	1tfp	1tsh	1x7s	1x7t	2b14
1bz8	NIL	0.92252/1.47	0.89827/1.37	0.83199/1.44	0.89177/1.31	0.88645/0.98	0.95649/1.21	0.87257/1.33	0.89066/1.43	0.88368/1.20	0.83233/1.54	0.92118/1.49	0.88621/0.98	0.88518/1.16	0.87945/1.06
1bzd	0.91544/1.47	NIL	0.92615/0.24	0.94092/0.88	0.92126/1.01	0.90285/0.23	0.90439/0.43	0.87507/0.93	0.90982/1.15	0.90664/0.38	0.84676/1.22	0.99361/0.35	0.90211/0.26	0.90713/0.35	0.89590/0.45
1etb	0.95785/1.37	0.99653/0.24	NIL	0.91979/0.85	0.98602/0.52	0.97201/0.21	0.97086/0.47	0.94047/0.91	0.97183/0.45	0.97445/0.39	0.90862/1.20	0.99729/0.21	0.97083/0.25	0.97434/0.38	0.96264/0.46
1g10	0.90776/1.44	0.85435/0.88	0.94314/0.85	NIL	0.93536/0.93	0.93425/0.86	0.92682/1.02	0.90411/1.17	0.92762/0.92	0.93375/0.95	0.87317/1.46	0.94206/0.87	0.93135/0.89	0.92485/0.97	0.92648/0.93
1ie4	0.93496/1.31	0.97399/1.01	0.96981/0.52	0.89771/0.93	NIL	0.95186/0.34	0.94803/0.59	0.92736/0.87	0.96468/0.62	0.95135/0.53	0.89684/1.17	0.97615/1.00	0.94966/0.39	0.95304/0.49	0.94258/0.53
1ii	0.96961/0.98	0.99673/0.23	0.99729/0.21	0.93425/0.86	0.99293/0.34	NIL	0.98715/0.47	0.95701/0.90	0.98693/0.47	0.99034/0.41	0.93237/1.19	0.99766/0.19	0.99727/0.21	0.94426/0.34	0.98874/0.44
1jp	0.95649/1.21	0.98918/0.43	0.98738/0.47	0.91918/1.02	0.98005/0.59	0.97875/0.47	NIL	0.93886/0.98	0.96550/0.68	0.93886/0.98	0.91505/1.34	0.98918/0.43	0.98104/0.42	0.98732/0.26	0.97436/0.54
1oo2	0.94474/1.33	0.95542/0.93	0.95618/0.91	0.89666/1.17	0.95839/0.87	0.94901/0.90	0.93886/0.98	NIL	0.94807/1.15	0.94392/0.93	0.90044/1.40	0.95603/0.91	0.94730/0.89	0.94317/0.95	0.94709/0.91
1qab	0.90458/1.43	0.93138/1.15	0.98533/0.45	0.86249/0.92	0.92451/0.62	0.91620/0.47	0.90476/0.68	0.88938/1.15	NIL	0.90993/0.59	0.86235/1.23	0.93408/1.33	0.91428/0.51	0.90789/0.63	0.90026/0.64
1sok	0.95783/1.20	0.99192/0.38	0.99110/0.39	0.82599/0.95	0.98359/0.53	0.98189/0.41	0.93886/0.98	0.94392/0.93	0.97137/0.59	NIL	0.92166/1.23	0.99132/0.37	0.98262/0.39	0.97984/0.45	0.97716/0.51
1tfp	0.91403/1.54	0.93823/1.22	0.33880/1.20	0.88026/1.46	0.94165/1.17	0.94049/1.19	0.93020/1.34	0.91524/1.40	0.93345/1.23	0.93695/1.23	NIL	0.98805/1.21	0.93695/1.22	0.92841/1.25	0.94008/1.22
1tsh	0.91412/1.49	0.99361/0.35	0.92681/.21	0.85528/0.87	0.92318/1.00	0.90361/0.19	0.90439/0.43	0.87562/0.91	0.91246/1.33	0.90614/0.37	0.84670/1.21	NIL	0.90275/0.23	0.90794/0.33	0.89579/0.45
1x7s	0.96930/0.98	0.99582/0.26	0.99604/0.25	0.93135/0.89	0.99052/0.39	0.99727/0.21	0.98948/0.42	0.95526/0.89	0.98469/0.51	0.99109/0.39	0.92925/1.22	0.99660/0.23	NIL	0.98600/0.29	0.99076/0.40
1x7t	0.95963/1.16	0.99246/0.35	0.99097/0.38	0.91717/0.97	0.98539/0.49	0.97484/0.34	0.98732/0.26	0.94317/0.95	0.96905/0.63	0.97984/0.45	0.91323/1.25	0.99344/0.33	0.97755/0.29	NIL	0.72082/0.46
2b14	0.96125/1.06	0.98827/0.45	0.98742/0.46	0.92648/0.93	0.98281/0.53	0.98874/0.44	0.98268/0.54	0.95505/0.91	0.97534/0.64	0.98553/0.51	0.93231/1.22	0.98812/0.45	0.99076/0.40	0.97915/0.46	NIL

Table 7: Pairwise TM-align structural similarity matrix of 15 representative transthyretin (TTR) protein structures. Each cell displays the TM-score (indicative of structural similarity) followed by the RMSD value (in Å), separated by a slash.

Table 7, matrix displays TM-scores and corresponding RMSD values (in Å), with the former indicating the degree of global structural similarity where scores >0.5 suggest meaningful structural alignment and the latter indicating spatial deviations. The analysis revealed a consistently high level of structural similarity across most protein pairs, with TM-scores frequently exceeding **0.9**, such as between **1BZ8_A** *(Transthyretin)* and **1ETB_A** *(Transthyretin)* (TM-score: 0.89827), **1I4E_A** *(Transthyretin)* and **1III_A** *(Transthyretin)* (0.95186), and **1TFP_A** *(Muscle fatty acid binding protein)* and **1TSH_A** *(Transthyretin)* (0.98805), coupled with low RMSD values typically below **1.5 Å**. These findings reflect a highly conserved fold characteristic of the TTR family. Very high TM-scores (≥0.97) were observed among multiple comparisons, notably **1X7T_A** *(Transthyretin)* with **1X7S_A** *(Transthyretin)* (0.99076) and **1TSH_A** *(Transthyretin)* with **1ETB_A** *(Transthyretin)* (0.99729), indicating near-identical core structures, likely corresponding to conserved β-sheet architecture central to TTR function. Overall, the TM-align results affirm the evolutionary conservation of TTR protein structures and underscore their structural rigidity despite possible sequence-level variability.

RBP4	4O9S_A	6QBA_A	1JYJ_A	1AQB_A	1HBQ_A	1ERB_A	7VNS_A	2NND-A	5E22_A	7VNL_A	7YX1_A	7O3K_A	1IW2_A	2QO5_C	2OVA_A
4O9S_A	NIL	0.97171/ 1.03	0.95157/ 1.27	0.96612/ 1.13	0.96600/ 1.13	0.95776/ 1.08	0.77038/ 2.35	0.63098/ 2.88	0.76927/ 2.35	0.76542/ 2.41	0.76706/ 2.36	0.75897/ 2.48	0.27672/ 5.62	0.70927/ 3.07	0.70531/ 3.04
6QBA_A	0.96631/ 1.03	NIL	0.94518/ 1.42	0.95542/ 1.34	0.96000/ 1.31	0.95098/ 1.14	0.77778/ 2.24	0.63186/ 2.80	0.77748/ 2.24	0.77444/ 2.26	0.77750/ 2.23	0.76514/ 2.45	0.31773/ 4.46	0.71220/ 2.96	0.70827/ 3.00
1JYJ_A	0.95687/ 1.27	0.95574/ 1.42	NIL	0.97805/ 0.80	0.97586/ 0.87	0.9706/ 0.85	0.76881/ 2.46	0.64415/ 2.73	0.76960/ 2.42	0.76763/ 2.44	0.77244/ 2.37	0.76221/ 2.55	0.25569/ 5.62	0.13981/ 4.63	0.71120/ 3.04
1AQB_A	0.96612/ 1.13	0.96074/ 1.34	0.97256/ 0.80	NIL	0.99149/ 0.48	0.98354/ 0.36	0.77351/ 2.35	0.74048/ 2.73	0.77249/ 2.35	0.77009/ 2.39	0.7725/ 2.38	0.76329/ 2.49	0.27883/ 6.20	0.71271/ 2.95	0.71044/ 3.03
1HBQ_A	0.96064/ 1.13	0.96000/ 1.31	0.96495/ 0.87	0.98590/ 0.48	NIL	0.97964/ 0.30	0.77438/ 2.31	0.73948/ 2.73	0.77376/ 2.32	0.77218/ 2.33	0.77310/ 2.31	0.7671/ 2.38	0.27881/ 5.99	0.71242/ 2.94	0.70785/ 3.06
1ERB_A	0.96858/ 1.08	0.96709/ 1.14	0.9761/ 0.85	0.99486/ 0.36	0.99689/ 0.30	NIL	0.77978/ 2.28	0.64534/ 2.75	0.77826/ 2.29	0.77659/ 2.31	0.77763/ 2.29	0.7690 / 2.45	0.27722/ 5.59	0.7183/ 2.87	0.71524/ 2.95
7VNS_A	0.79471/ 2.35	0.80650/ 2.24	0.78905/ 2.46	0.79802/ 2.35	0.80302/ 2.31	0.79629/ 2.28	NIL	0.67871/ 2.86	0.98123/ 0.96	0.99049/ 0.56	0.98248/ 0.95	0.94286/ 1.10	0.30529/ 5.56	0.76641/ 2.48	0.76689/ 2.50
2NND_A	0.72766/ 2.88	0.73249/ 2.80	0.74026/ 2.73	0.64122/ 2.73	0.63720/ 2.73	0.73749/ 2.75	0.76297/ 2.86	NIL	0.76207/ 2.77	0.76607/ 2.75	0.76157/ 2.79	0.76925/ 2.62	0.12157/ 5.01	0.80000/ 2.71	0.79077/ 2.08
5E22_A	0.79243/ 2.35	0.80607/ 2.24	0.78950/ 2.42	0.79698/ 2.35	0.80234/ 2.32	0.79476/ 2.29	0.98123/ 0.96	0.67781/ 2.77	NIL	0.99037/ 0.76	0.99562/ 0.32	0.9543/ 0.92	0.30300/ 5.86	0.75931/ 2.66	0.76129/ 2.56
7VNL_A	0.78950/ 2.41	0.80303/ 2.26	0.78784/ 2.44	0.79446/ 2.39	0.80060/ 2.33	0.79299/ 2.31	0.99040/ 0.56	0.68107/ 2.75	0.99037/ 0.76	NIL	0.99101/ 0.75	0.95132/ 0.97	0.30731/ 5.28	0.76269/ 2.63	0.76568/ 2.52
7YX1_A	0.79117/ 2.36	0.80622/ 2.23	0.79298/ 2.37	0.79697/ 2.34	0.80226/ 2.31	0.79438/ 2.29	0.98248/ 0.95	0.67754/ 2.79	0.99562/ 0.32	0.99101/ 0.75	NIL	0.95376/ 0.93	0.30503/ 5.70	0.76012/ 2.65	0.7615/ 2.41
7O3K_A	0.79513/ 2.48	0.80568/ 2.45	0.79478/ 2.55	0.79989/ 2.49	0.80430/ 2.38	0.7978/ 2.45	0.95632/ 1.10	0.69494/ 2.62	0.97112/ 0.92	0.96804/ 0.97	0.97056/ 0.93	NIL	0.30877/ 5.64	0.77473/ 2.48	0.77384/ 2.44
1IW2_A	0.12152/ 5.62	0.13145/ 4.46	0.11643/ 5.62	0.12685/ 6.20	0.12415/ 5.99	0.12349/5.59	0.13367/ 5.56	0.31660/ 5.01	0.13257/ 5.86	0.13375/ 5.28	0.13358/ 5.70	0.13335/ 5.64	NIL	0.13981/ 4.63	0.13924/ 4.50
2QO5_C	0.71642/ 3.07	0.72292/ 2.96	0.33459/ 4.63	0.71989/ 2.95	0.72313/ 2.94	0.7183 / 2.87	0.75052/ 2.48	0.69524/ 2.71	0.74369/ 2.66	0.74705/ 2.63	0.74450/ 2.65	0.74678/ 2.48	0.33459/ 4.63	NIL	0.92193/ 0.95
2OVA_A	0.74611/ 3.04	0.75299/ 3.00	0.74003/ 3.04	0.75185/ 3.03	0.75265/ 3.06	0.74931/ 2.95	0.78782/ 2.50	0.72021/ 2.08	0.78197/ 2.56	0.78655/ 2.52	0.76689/2.41	0.78225/ 2.44	0.34751/ 4.50	0.97135/ 0.95	NIL

Table 8: Pairwise TM-align structural similarity matrix of 15 representative RBP4 protein structures. Each cell displays the TM-score (indicating structural similarity) followed by the RMSD value (in Å).

Table 8, examine structural conservation among members of the retinol-binding protein 4 (RBP4) family, TM-align was used to perform pairwise structural alignments on 15 representative PDB structures. The resulting matrix includes TM-scores (a normalized measure of structural similarity) and RMSD values (in Å) for each protein pair. TM-scores above 0.5 are indicative of meaningful structural similarity, with values closer to 1.0 denoting nearly identical protein folds. The majority of protein pairs in this matrix exhibit TM-scores between 0.7 and 0.97, demonstrating a moderate to high level of structural conservation. Examples include **1JYJ_A** *(Plasma retinol binding protein)* vs **4O9S_A** *(Retinol-binding protein 4)* (TM-score: 0.95687, RMSD: 1.27) and **1AQB_A** (*Retinol-binding protein)* vs **1HBQ_A** *(Retinol-binding protein Bos taurus)* (0.99149 / 0.48), suggesting that key structural elements of the RBP4 fold are preserved across homologs. However, certain alignments, such as those involving **1IW2_A** *(Complement Protein C8gamma)* or **2Q05_C** ***(****tyr/ser protein)*, show lower TM-scores 0.27672 / 5.62 between **1IW2_A** *(Complement Protein C8gamma)* and **4O9S_A** *(Retinol-binding protein 4)*, reflecting either significant conformational variability or functional divergence. Overall, the TM-align analysis reveals that while a core β-barrel structure is largely conserved across RBP4 homologs, structural deviations exist in some members, likely reflecting differences in ligand binding, conformational states, or domain arrangements.

5.3. PROMALS3D Results

PROMALS3D was utilized to perform **pairwise sequence-structure alignments** between selected protein structures derived from CD-HIT clustering. By incorporating both sequence and 3D structural information, PROMALS3D provides more accurate alignments, particularly for distantly related homologs. The pairwise alignments revealed that conserved residues and structural motifs were consistently aligned, even in regions where sequence identity was low.

Figures 31 to 34 is the pairwise sequence-structure alignments generated using PROMALS3D were visualized using **Jalview**, which enabled a clearer interpretation of conserved and variable regions across the aligned protein pairs. The alignment was saved in **Clustal format**, preserving alignment structure and annotations. In Jalview, the sequences were colored based on **percentage identity**, allowing visual assessment of residue conservation across aligned positions. In this coloring scheme, **dark blue** indicates **100% identity**, meaning that the residues are identical in all sequences at that position. **Lighter shades of blue** (e.g., cyan or light blue) represent **high but not complete identity**, showing partial conservation, while **white or uncolored positions** reflect low or no conservation. This color gradient helps highlight conserved motifs and structurally important regions, especially those maintained across evolution despite sequence divergence. The combined use of PROMALS3D and Jalview thus facilitated a detailed evaluation of both sequence and structural conservation within the aligned protein pairs.

Promals3D sequence-structure based pairwise alignment results:

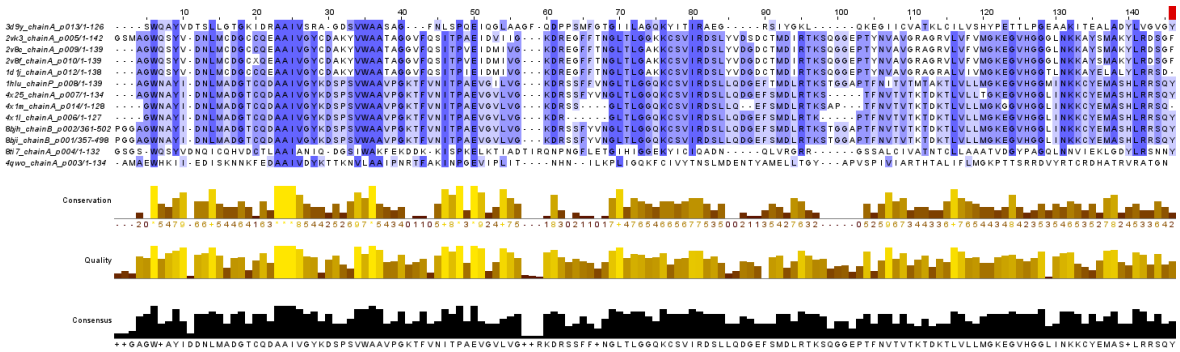


Figure 31: Sequence-Structure Alignment of PFN-1 Clustered Protein Representatives from PROMALS3D. Dark blue shades represent higher conservation, and lighter colors reflect low conservation.

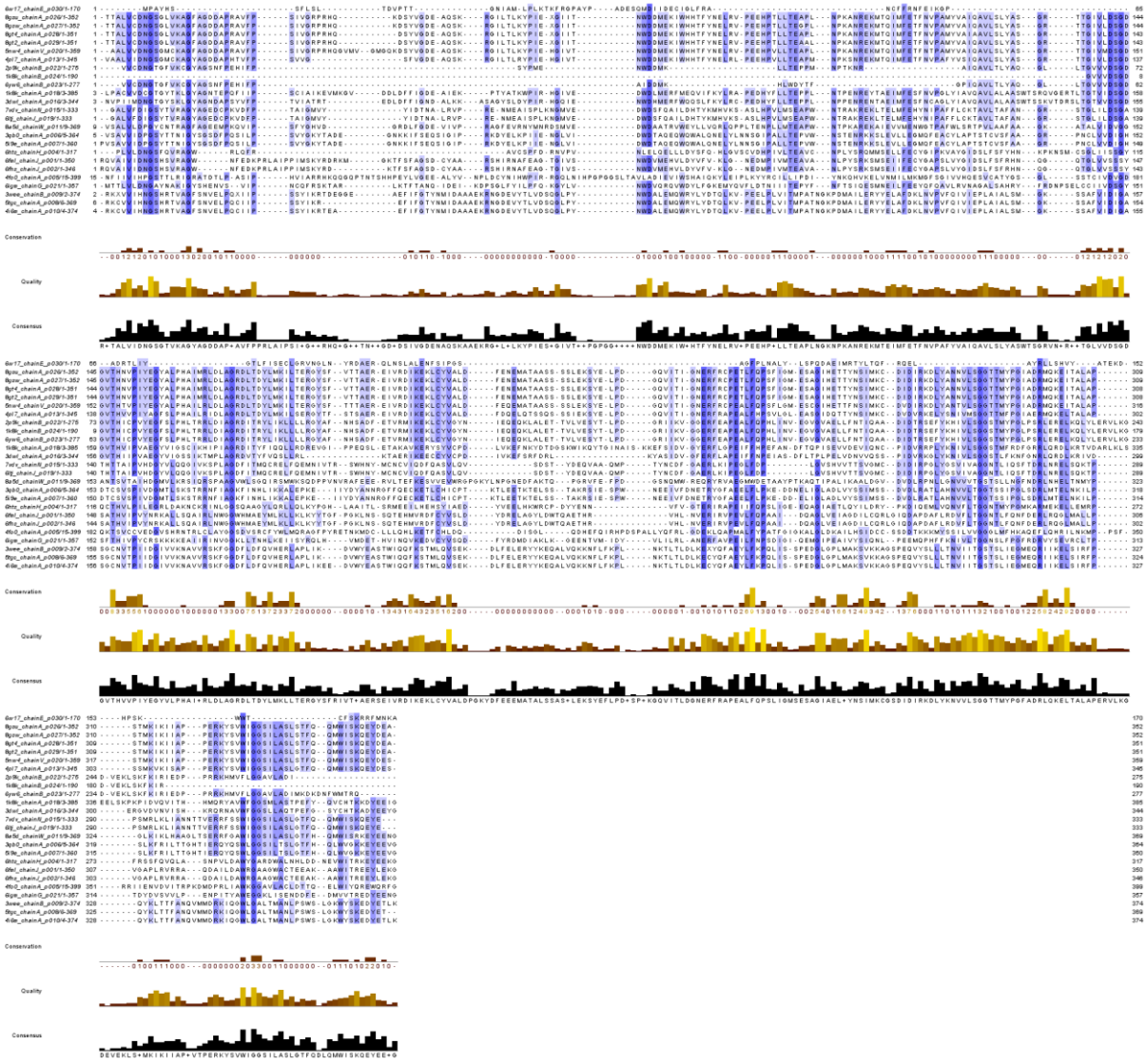


Figure 32: Sequence-Structure Alignment of ACTB Clustered Protein Representatives from PROMALS3D. Dark blue shades represent higher conservation, and lighter colors reflect low conservation.

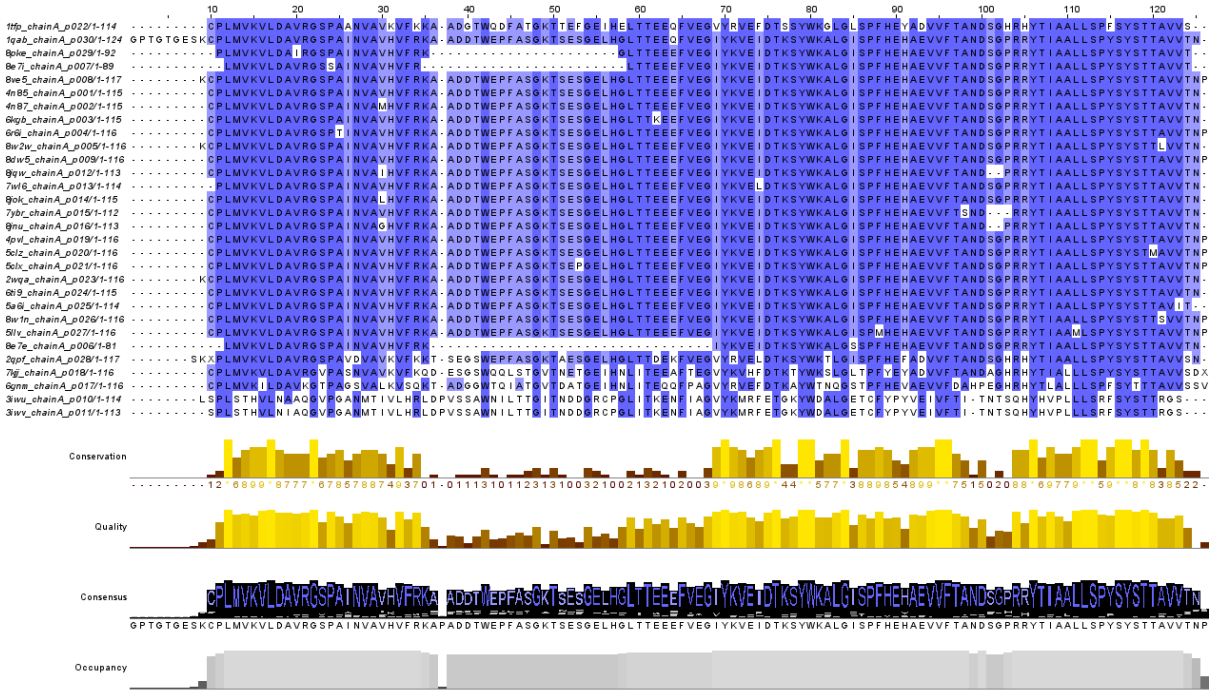


Figure 33: Sequence-Structure Alignment of TTR Clustered Protein Representatives from PROMALS3D. Dark blue shades represent higher conservation, and lighter colors reflect low conservation.

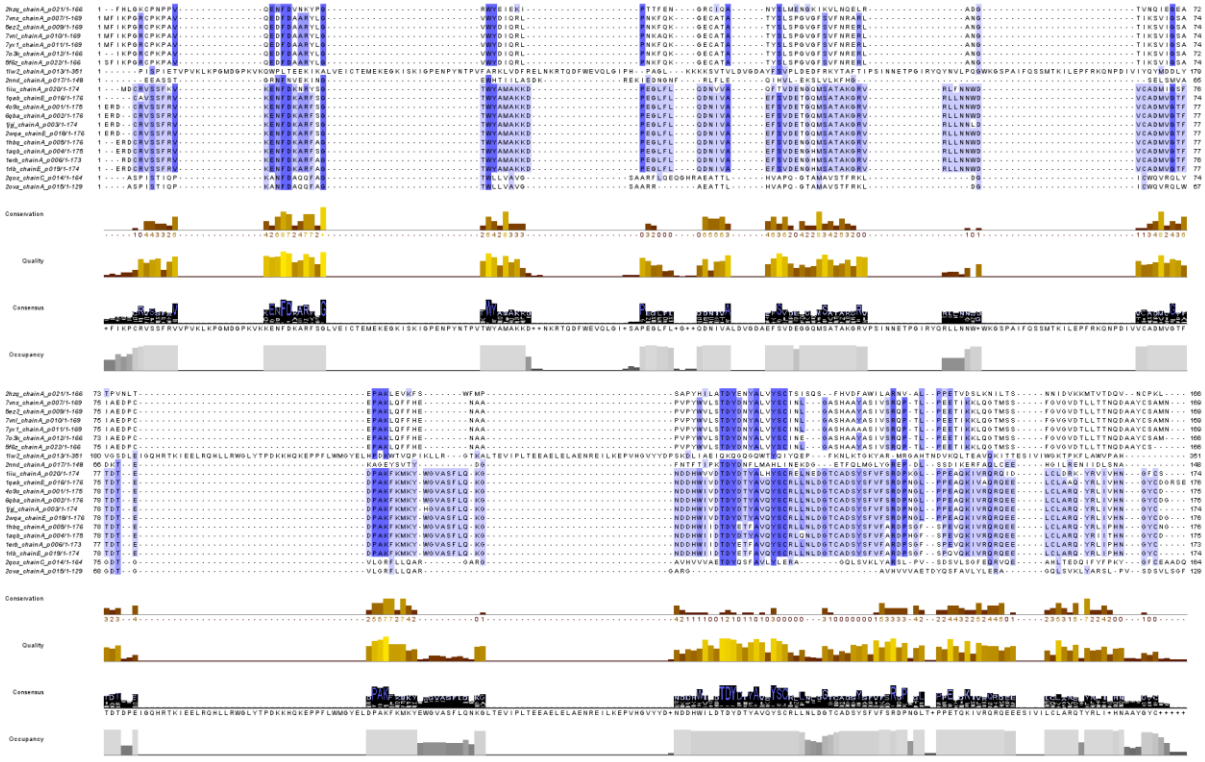


Figure 34: Sequence-Structure Alignment of RBP4 Clustered Protein Representatives from PROMALS3D. Dark blue shades represent higher conservation, and lighter colors reflect low conservation.

PROMALS3D Superimposed structures

The representative PDB structures obtained from CD-HIT clustering were used as input for structural alignment using PROMALS3D. The resulting superimposed models revealed that, despite considerable sequence variation, the core structural frameworks of the proteins remained well conserved. The alignment prominently highlighted conserved secondary structure elements, such as α-helices and β-sheets, particularly within homologous regions. In the superimposed visualization, the **main reference structure is shown in green**, allowing for clear comparison against the aligned proteins.

The figure 35 is superimposed models revealed conserved core structures across the protein pairs, confirming structural conservation despite evolutionary divergence. In the visualizations, structurally aligned regions are closely overlaid, especially in functional domains such as active sites, interaction interfaces or ligand-binding pockets. The sequence-structure alignments further supported this observation, with conserved residues aligning well in spatially equivalent positions. These results reinforce the evolutionary and functional relevance of the studied proteins, demonstrating how structurally important features are preserved even when sequences differ significantly.

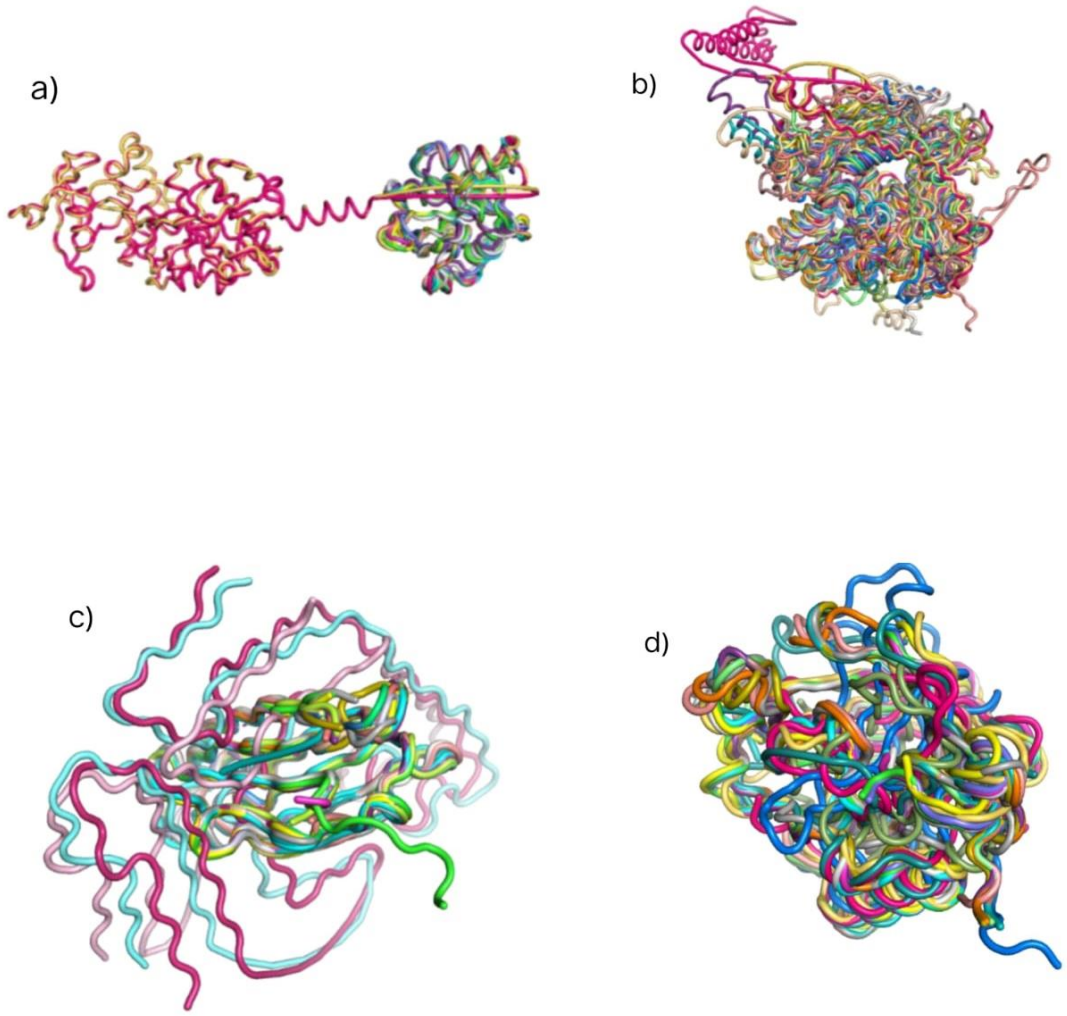


Figure 35: Superimposed Protein Structures of a) PFN-1 b) ACTB c) TTR, d) RBP4 generated using PROMALS3D based on CD-HIT cluster representatives and main reference structure is shown in green.

6. Visualization of Phylogeny Tree using iTOL

The phylogenetic tree visualized using iTOL (Interactive Tree Of Life) illustrates the evolutionary relationships among sequences derived from diverse taxonomic groups. Organisms are color-coded based on their classification, allowing for clear visual distinction between clades.

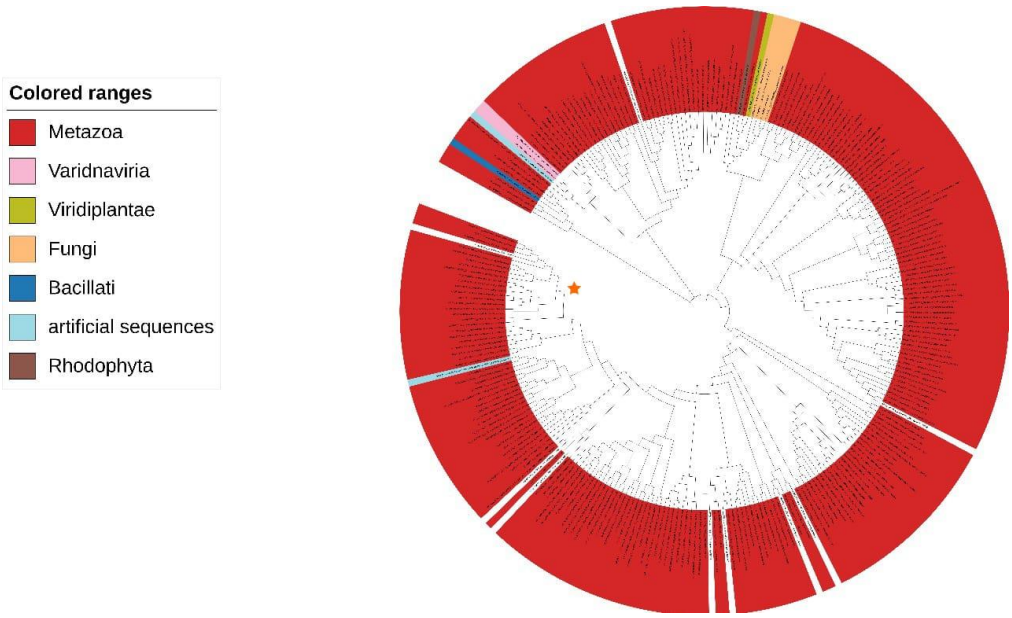


Figure 36: Phylogenetic tree of PFN-1 visualized using iTOL.

Figure 36 is a circular phylogenetic tree visualized through iTOL illustrates the taxonomic distribution of homologs of **Profilin-1** across various domains of life. The dataset is overwhelmingly dominated by sequences from the **Metazoa** group (shown in red), indicating the strong evolutionary conservation and functional significance of Profilin-1 in multicellular animals. This widespread presence within Metazoa suggests that Profilin-1 has a conserved role in actin dynamics across diverse animal phyla. A few sequences are also observed in **Fungi** (orange), **Viridiplantae** (green), and **Rhodophyta** (brown), reflecting limited but notable cross-kingdom conservation. Additionally, the presence of sequences labeled as **artificial** (light blue) and rare occurrences in groups like **Varidnaviria** and **Bacillati** indicate either synthetic constructs, database artifacts, or distant homologs. The overall topology of the tree supports the hypothesis that Profilin-1 originated early in eukaryotic evolution and diversified extensively within Metazoa, maintaining a conserved structural and functional role essential for cellular processes such as cytoskeleton regulation and signal transduction.

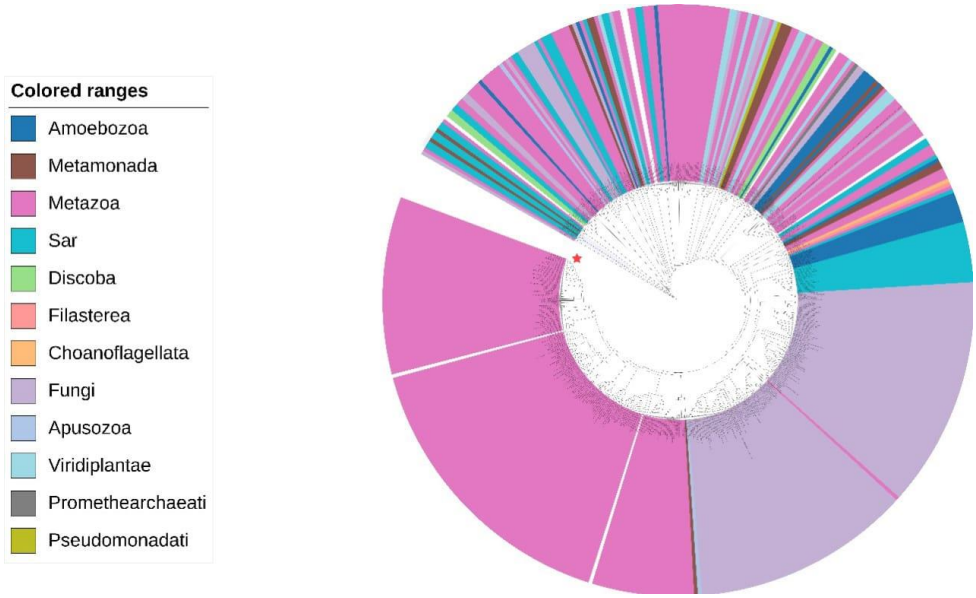


Figure 37: Phylogenetic tree of ACTB visualized using iTOL.

Figure 37 is a circular phylogenetic tree generated using iTOL depicts a diverse distribution of homologous sequences across multiple taxonomic groups. The largest proportion of sequences is clustered under the **Metazoa** clade (represented in pink), indicating a strong representation of multicellular animals in the dataset. This is followed by a significant grouping under the **Fungi** clade (lavender), reflecting the evolutionary conservation of the protein across eukaryotic lineages. Smaller, yet notable distributions are seen in groups such as **SAR** (cyan), **Viridiplantae** (light blue), **Discoba** (light green), and **Amoebozoa** (blue), highlighting the presence of homologs in diverse protist and plant lineages. The presence of taxa such as **Filasterea**, **Choanoflagellata**, **Apusozoa**, and **Metamonada** suggests that the protein is evolutionarily ancient, likely present before the divergence of major eukaryotic supergroups. The tree structure demonstrates evolutionary relationships and potential gene duplication or loss events across lineages. Overall, the distribution supports the hypothesis of a conserved functional role for the studied protein across a broad evolutionary spectrum.

The interaction between Profilin-1 and Actin is believed to be evolutionarily ancient, suggesting that this functional association originated early in the history of eukaryotes. Its strong conservation across diverse taxa highlights its essential role in cytoskeletal regulation and cellular processes throughout evolution.

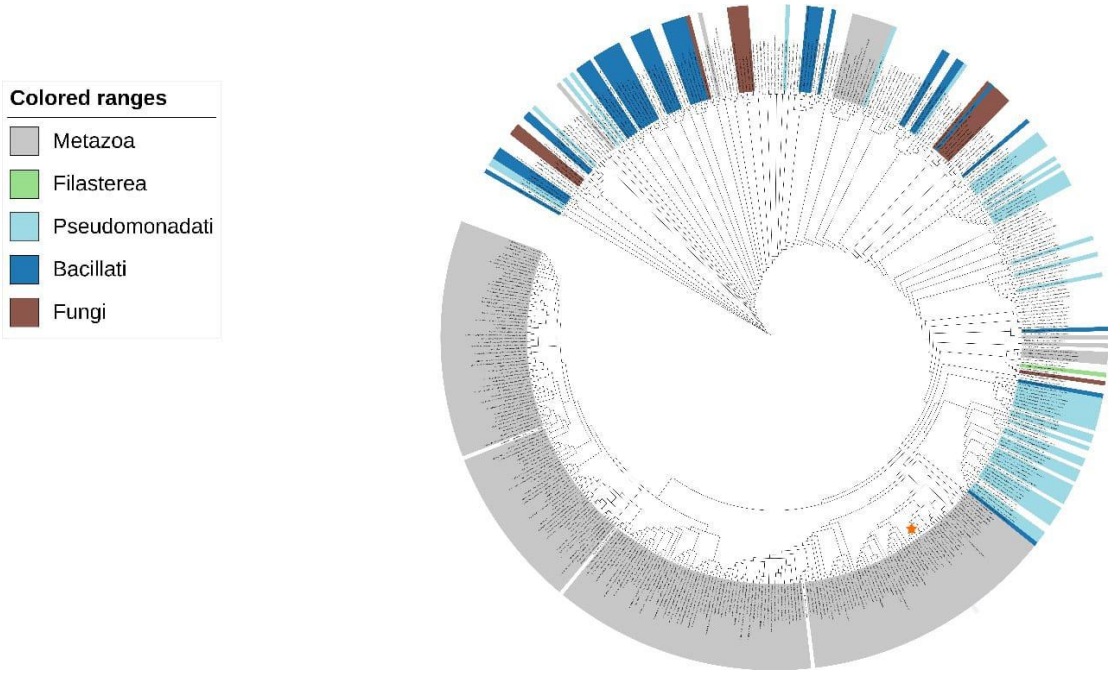


Figure 38: Phylogenetic tree of TTR visualized using iTOL.

Figure 38 is a circular phylogenetic tree for the Transthyretin (TTR) protein, visualized using iTOL, reveals a broad taxonomic distribution, with a dominant presence across **Metazoa** (gray) indicating its evolutionary conservation in multicellular animals. Significant representation is also observed among **Pseudomonadati** (light blue), **Bacillati** (blue), and **Fungi** (brown), reflecting the widespread evolutionary presence of TTR-like sequences even among bacteria and fungi. A smaller yet distinct cluster from **Filasterea** (green) further highlights the protein's presence in unicellular relatives of animals, suggesting a possibly ancient evolutionary origin. The tree displays a clear separation of clades based on taxonomy, and the diversity across domains suggests that transthyretin or its homologs may have evolved to fulfill essential, possibly divergent, functions in various organisms. The consistent representation across multiple lineages supports its role as a structurally conserved and potentially functionally adaptable protein.

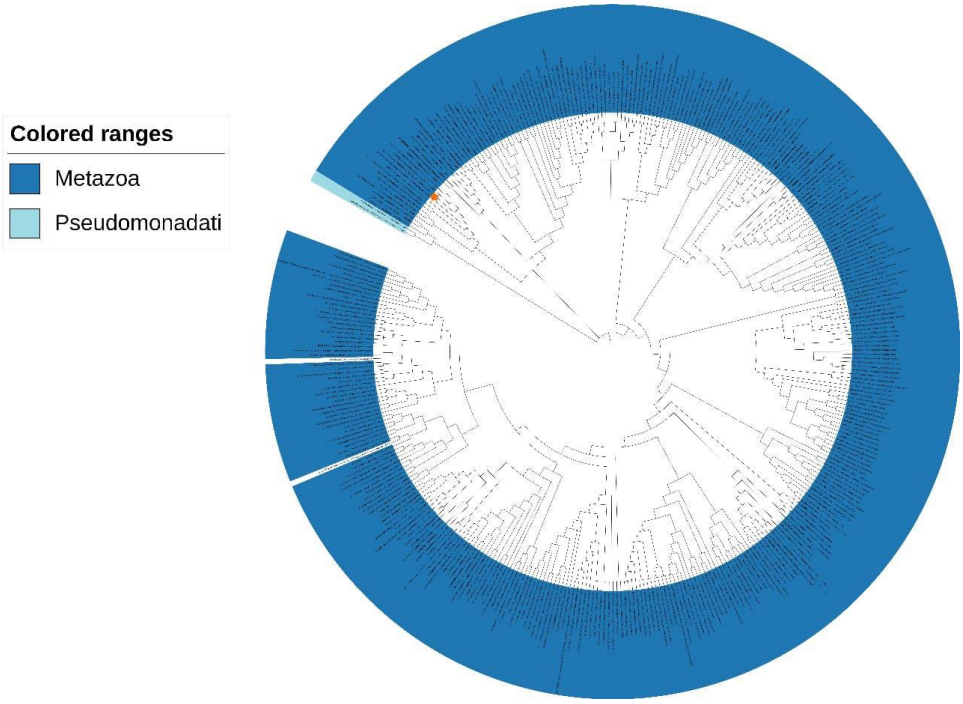


Figure 39: Phylogenetic tree of RBP4 visualized using iTOL.

Figure 39 is a circular phylogenetic tree of **Retinol Binding Protein 4 (RBP4)**, generated using iTOL, reveals a highly concentrated distribution within the **Metazoa** (blue) clade, which dominates the dataset. Nearly all sequences are derived from metazoan organisms, indicating the conserved and lineage-specific nature of RBP4 within multicellular animals. Only a minimal number of sequences are assigned to the **Pseudomonadati** (light blue) group, suggesting either distant homologs or possible annotation artifacts. This tight clustering within Metazoa supports the idea that RBP4 plays a specialized and conserved role in animal physiology, likely associated with vitamin A transport and metabolism, a function not widely distributed outside of metazoan systems. The tree structure reflects the evolutionary constraint on this protein's sequence and functional domain within animals.

The interaction between Transthyretin and Retinol Binding Protein 4 is evolutionarily conserved among metazoans, indicating that this functional partnership likely emerged early during vertebrate evolution and preserved due to its critical physiological role. This association underscores their critical roles in transport and metabolism of thyroid hormones and retinol reinforcing their interdependence in vertebrates.

7. Visualization of Functionally important residues

Functionally important residues identified from **PDBe** and **PDBsum** were visualized using **Jalview** to assess their conservation across homologous sequences. The selected alignments included sequences retrieved from CD-HIT clusters at a 0.7 identity threshold, specifically those containing ***Homo sapiens* counterparts**, ensuring relevance to human protein function. In Jalview, the alignments were colored using the **"Annotation (Conservation)"** color scheme, which shades residues based on their degree of conservation across all sequences in the alignment. In this scheme, **highly conserved residues appear in darker shades (typically dark red), while variable positions are represented with lighter tones**, allowing for immediate visual recognition of evolutionary conservation. This visualization helped confirm that many of the functionally significant residues mapped from structural databases are also conserved across species, supporting their functional importance.

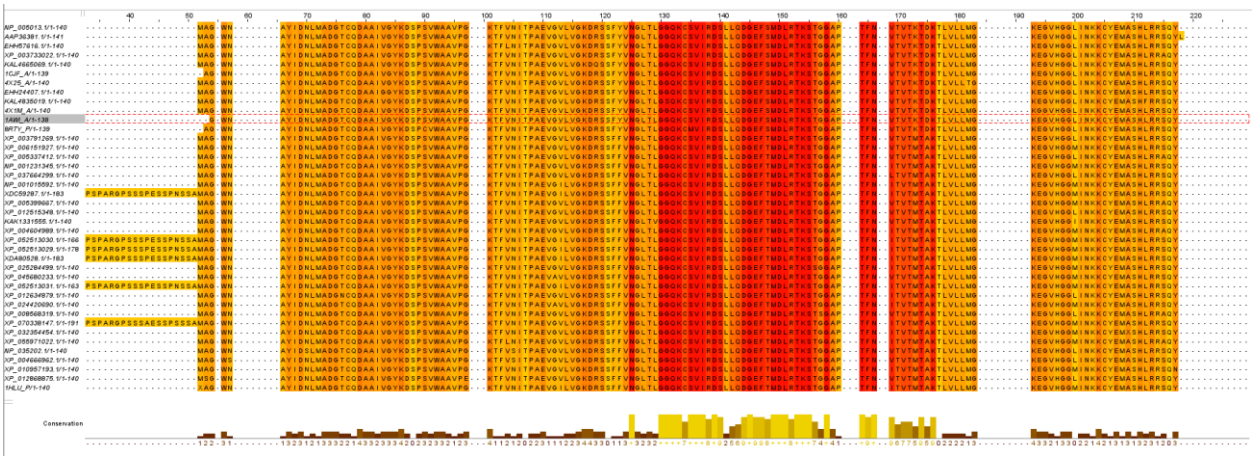


Figure 40: Visualization of conserved residues of PFN-1 using Jalview for clustered protein sequences retrieved at a 0.7 CD-HIT cut-off, highlighting conserved residues using the ***"Annotation (Conservation)"*** *color scheme. The alignment includes* Homo sapiens *sequences to facilitate functional and evolutionary comparisons.*

Figure 40 shows the following residues **Trp2, Asn3, Ile7, Lys24, Asp25, Ser26, Val29, Trp30, Ala31,** and **Ala32** which were residues of interest are found to be conserved across *Homo sapiens* counterparts, highlighting their potential functional importance.

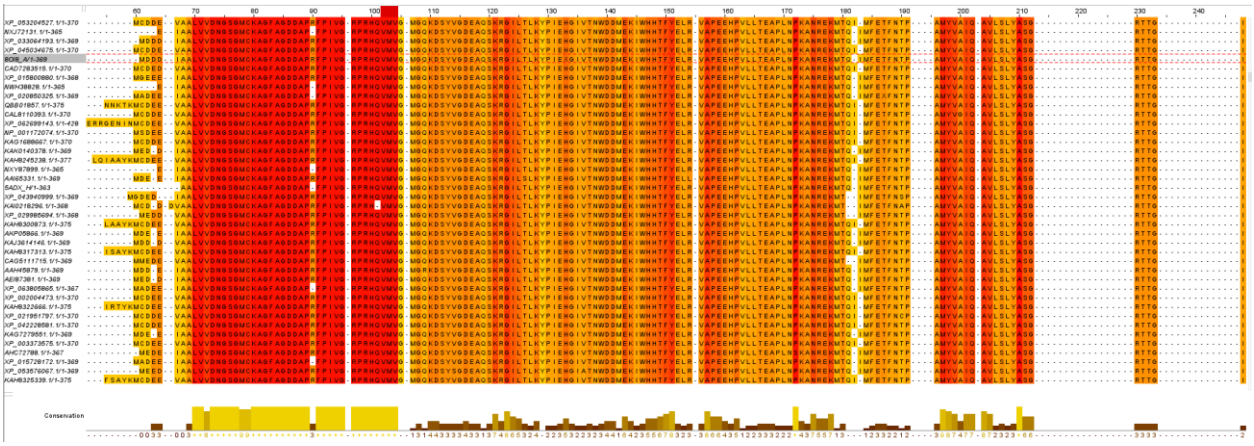


Figure 41: Visualization of conserved residues of ACTB using Jalview for clustered protein sequences retrieved at a 0.7 CD-HIT cut-off, highlighting conserved residues using the ***"Annotation (Conservation)"*** *color scheme. The alignment includes* Homo sapiens *sequences to facilitate functional and evolutionary comparisons.*

Figure 41 is the following residues **Ile5, Phe20, Gly22, Asp24, Arg27, Leu93, Val95, Ile233, Asp286, Val287, Gly12, Met16, Lys17, Pro31, Ile33, Gly155, Phe199, Tyr68, Arg205,** and **Val209** which were residues of interest are found to be conserved across *Homo sapiens* counterparts, highlighting their potential functional importance.

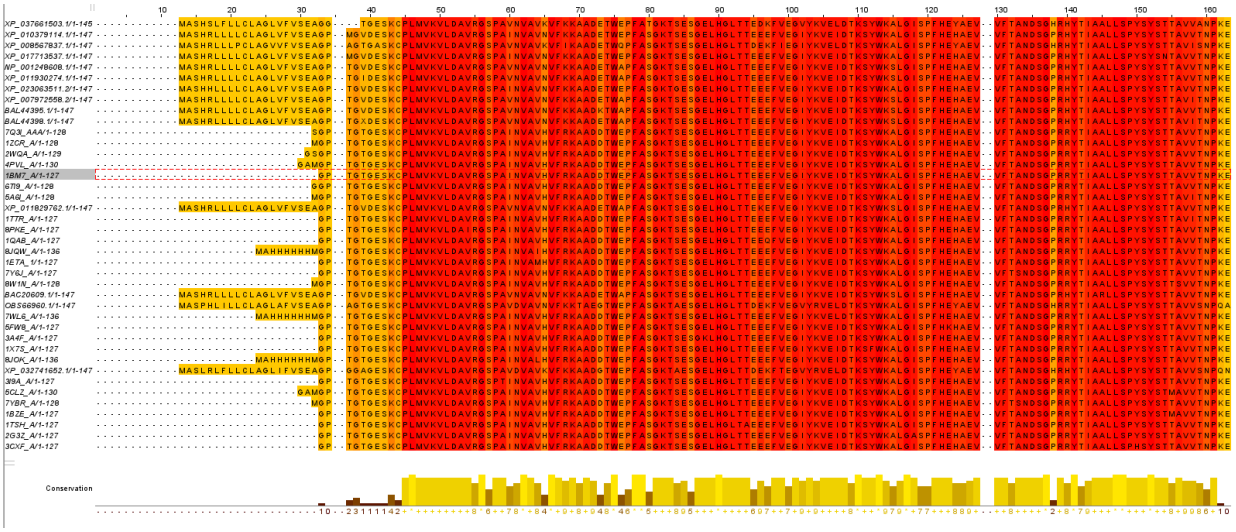


Figure 42: Visualization of conserved residues of TTR using Jalview for clustered protein sequences retrieved at a 0.7 CD-HIT cut-off, highlighting conserved residues using the ***"Annotation (Conservation)"*** *color scheme. The alignment includes* Homo sapiens *sequences to facilitate functional and evolutionary comparisons.*

Figure 42 is the following residues **Lys15, Leu17, Thr106, Ala108, Leu110, Ser117, Thr119, Val121, Ala19, Pro24, Ile68, Phe87, His90, Glu92, Ile93, Asp94, Thr96, Tyr105,**

Ile107, Ser112, Ala117, and **Val124** which were residues of interest are found to be conserved across *Homo sapiens* counterparts, highlighting their potential functional importance.

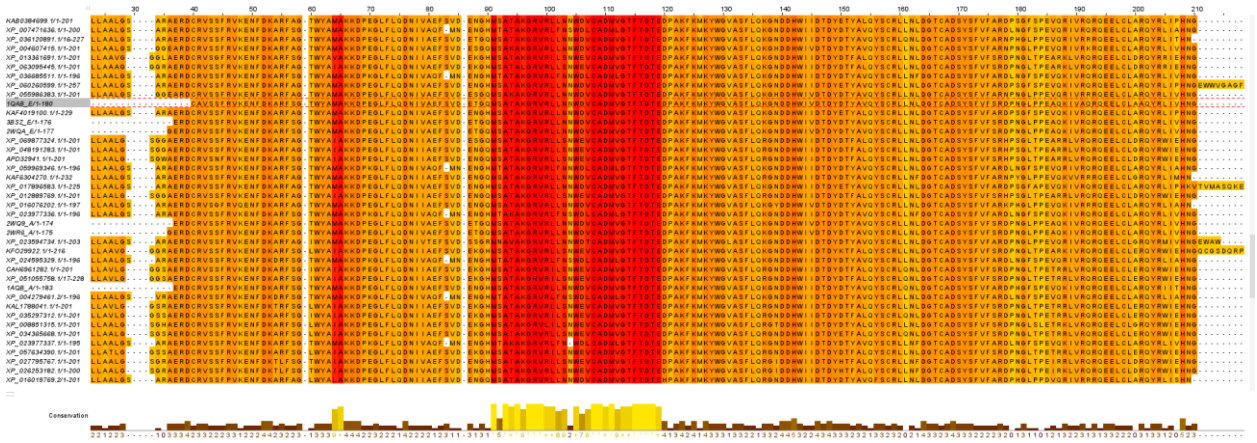


*Figure 43: Visualization of conserved residues of RBP4 using Jalview for clustered protein sequences retrieved at a 0.7 CD-HIT cut-off, highlighting conserved residues using the "**Annotation (Conservation)**" color scheme. The alignment includes* Homo sapiens *sequences to facilitate functional and evolutionary comparisons.*

Figure 43 is the following residues **Ala52, Ala54, Val58, Met70, Val71, Phe74, Tyr130,** and **Glu176** which were residues of interest are found to be conserved across *Homo sapiens* counterparts, highlighting their potential functional importance.

Mutant residues

Several of the identified mutant residues overlapped with the previously predicted functionally important residues, indicating their critical role in maintaining protein structure or function. Notably, these overlapping residues were associated with known disease implications, reinforcing the idea that mutations at conserved and functionally essential positions are more likely to disrupt protein activity and lead to pathological outcomes.

Residue no	**Residue name**	**Mutated residue**	**Effect**	**Disease**
4	Tryptophan	Cysteine	Somatic; MODERATE impact	Transitional Cell Papillomas and Carcinomas
17-27	Cysteine - Glutamine	Deletion	Somatic; MODERATE impact	Adenomas and Adenocarcinomas
134	Histidine	Leucine	Somatic; MODERATE impact	Cystic, Mucinous and Serous Neoplasm

Table 9: Mutant variants of PFN-1 protein from UNIPROT database.

Residue no	**Residue name**	**Mutated residue**	**Effect**	**Disease**
2	Aspartic acid	Alanine	Somatic; MODERATE impact.	Cystic, Mucinous and Serous Neoplasms
2	Aspartic acid	Asparagine	Somatic; MODERATE impact.	Cystic, Mucinous and Serous Neoplasms - Ductal and Lobular Neoplasms
2	Aspartic acid	Tyrosine	Baraitser-Winter syndrome 1	Baraitser-Winter syndrome 1 (BRWS1)
3	Aspartic acid	Phenylalanine	Baraitser-Winter	Baraitser-Winter

			syndrome 1	syndrome 1 (BRWS1)
3	Aspartic acid	Asparagine	Somatic; MODERATE impact.	Adenomas and Adenocarcinomas
5	Isoleucine	Phenylalanine	Somatic; MODERATE impact.	Plasma Cell Tumors
7	Alanine	Threonine	Intellectual disability Baraitser-Winter syndrome 1	Baraitser-Winter syndrome 1 (BRWS1) - Intellectual disability
16	Methionine	Threonine	Somatic; MODERATE impact.	Plasma Cell Tumors
20	Phenylalanine	Leucine	Somatic; MODERATE impact.	Transitional Cell Papillomas and Carcinomas
24	Aspartic acid	Asparagine	ACTB-related BAFopathy	ACTB-related BAFopathy
31	Proline	Serine	Baraitser-Winter syndrome 1	Baraitser-Winter syndrome 1 (BRWS1)
58	Alanine	Threonine	Somatic; MODERATE impact.	Adenomas and Adenocarcinomas
58	Alanine	Valine	Intellectual disability	Intellectual disability
66	Leucine	Proline	Somatic; MODERATE impact.	Adenomas and Adenocarcinomas
68	Tyrosine	Cysteine	Somatic; MODERATE impact.	Adenomas and Adenocarcinomas

205	Arginine	Glutamine	Baraitser-Winter syndrome 1 Inborn genetic diseases	Baraitser-Winter syndrome 1 (BRWS1) - Inborn genetic diseases
205	Arginine	Tryptophan	Baraitser-Winter syndrome 1	Baraitser-Winter syndrome 1 (BRWS1)
209	Valine	Leucine	Baraitser-winter syndrome 1 (brws1) Baraitser-Winter syndrome 1 Somatic; MODERATE impact.	Baraitser-Winter syndrome 1 (BRWS1) - Transitional Cell Papillomas and Carcinomas
209	Valine	Methionine	Baraitser-winter syndrome 1 (brws1) Baraitser-Winter syndrome 1 Inborn genetic diseases ACTB-related BAFopathy	ACTB-related BAFopathy - Baraitser-Winter syndrome 1 (BRWS1) - Inborn genetic diseases

Table 10: Mutant variants of ACTB protein from UNIPROT database.

Residue no	**Residue name**	**Mutated residue**	**Effect**	**Disease**
93	Isoleucine	Valine	AMYLD1; amyloid polyneuropathy	Hereditary amyloidosis - Amyloidosis, hereditary systemic 1 (AMYLD1)
94	Aspartic acid	Tyrosine	Somatic; MODERATE impact	Adenomas and Adenocarcinomas
95	Serine	Tyrosine	AMYLD1; amyloid	Hereditary amyloidosis

			polyneuropathy	- Amyloidosis, hereditary systemic 1 (AMYLD1)
117	Alanine	Glycine	AMYLD1; amyloid polyneuropathy	Amyloidosis, hereditary systemic 1 (AMYLD1)
117	Alanine	Serine	AMYLD1; amyloid polyneuropathy	Amyloidosis, hereditary systemic 1 (AMYLD1)

Table 11: Mutant variants of TTR protein from UNIPROT database.

Residue no	**Residue name**	**Mutated residue**	**Effect**	**Disease**
130	Tyrosine	Asparagine	Congenital ocular coloboma	Congenital ocular coloboma - Microphthalmia
176	Glutamic acid	Lysine	Somatic; MODERATE impact	Nevi and Melanomas

Table 12: Mutant variants of RBP4 protein from UNIPROT database.

Table 8 to 12 are the mutant variants of PFN, Actin, TTR and RBP4 from UNIPROT database

Critical Residues

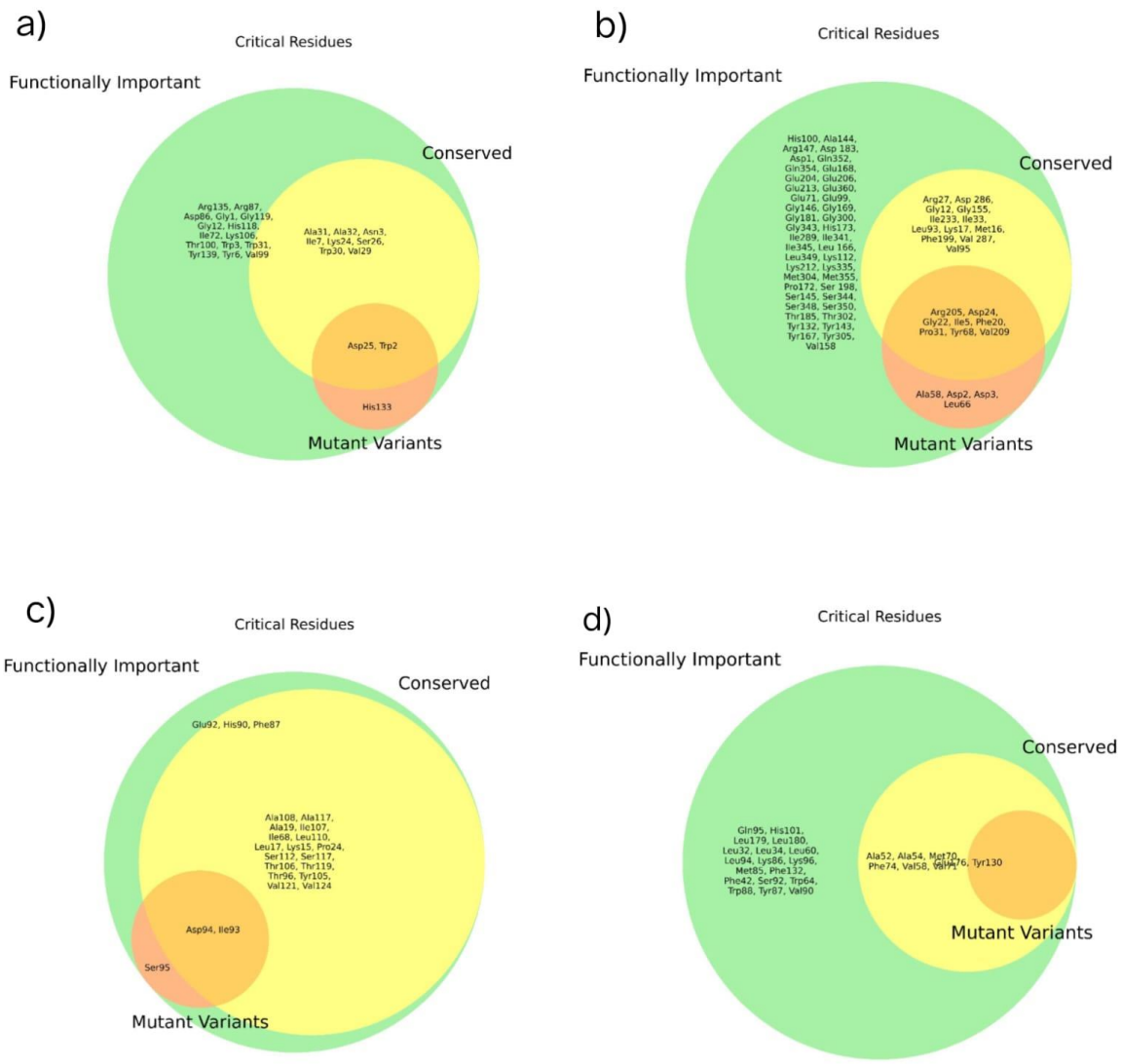


Figure44: Venn diagram illustrating the overlap among functionally important residues (light-green), conserved residues (yellow), and known mutant variants (salmon) in the a) PFN-1, b) ACTB, c) TTR, d) RBP4 proteins. The intersections (orange) highlight residues sharing multiple attributes.

Figure 44 is the Venn diagram that illustrates the distribution and overlap of functionally important residues, conserved residues, and mutant variants in the PFN-1, Actin, TTR, RBP4 proteins. Residues classified as **functionally important** depicted in **light-green** which are likely involved in essential biochemical interactions. The **conserved residues** highlighted in **yellow**, are preserved across multiple species, suggesting a fundamental role in maintaining structural or functional integrity. **Mutant variants** shown in **salmon** represent sites were pathogenic mutations. Notably, residues such as **Asp25** and **Trp2** in PFN-1, **Gly22, Ile5, Phe20, Asp24, Pro31, Tyr68, Arg205, Val209** in Actin, **Ile93, Asp94** and **Ala117** in TTR and **Gly176** and **Tyr130** in RBP4 proteins are situated at the intersection of all three categories highlighted in **orange**, indicating that they are conserved, functionally significant, and mutated making them critical hotspots of interest. Their concurrent involvement in multiple roles highlights their potential impact on protein stability and function, and underscores their relevance as potential targets for therapeutic intervention or biomarker development. Overall, this diagram underscores the intricate relationship between evolutionary conservation, functional importance, and pathogenic variation within the proteins.

Discussion

The present study explored the evolutionary and functional conservation of two distinct protein-protein interaction (PPI) pairs: Profilin-1 with Beta-actin and Transthyretin with Retinol Binding Protein 4 (RBP4) through an integrated bioinformatics approach. The results reveal substantial conservation at both sequence and structural levels across homologous proteins, highlighting the functional importance of these interactions in essential cellular and physiological processes.

Structural classification using SCOPe confirmed that the selected proteins belong to well defined folds and superfamilies, strengthening their evolutionary and functional significance. Protein-protein interaction networks retrieved from the STRING database and validated by PyMOL and PDBe interface analysis highlighted specific residues responsible for binding. These residues were found to cluster within conserved domains, supporting their roles in enabling interaction specificity and stability. Homologous sequence analysis, followed by CD-HIT clustering, demonstrated that a 70% sequence identity threshold provided an optimal balance between redundancy reduction and diversity preservation. This ensured that meaningful evolutionary patterns could be extracted without compromising on sequence variability.

Multiple sequence alignment using ClustalW (MEGA) consistently yielded the highest completeness scores, emphasizing its effectiveness in generating biologically relevant alignments. Jalview visualizations further revealed that highly conserved residues tended to be chemically and functionally important, often coinciding with regions involved in protein-protein interfaces or active sites. Structural alignments conducted using TM-align, PROMALS3D, and dendrograms generated through DALI provided a quantitative and visual summary of structural DALI revealed a consistent conservation in three-dimensional structure. Heatmaps and relatedness, with most homologs clustering closely, except for a few structural outliers. Such findings highlight the preservation of structural conservation across divergent sequences, particularly in biologically essential proteins like actin and transthyretin.

Phylogenetic analyses visualized using iTOL highlighted clear clade formation across taxonomic lineages, particularly among metazoans. This clustering emphasizes evolutionary constraints imposed on these proteins due to their central roles in cytoskeletal regulation and vitamin A transport. Interestingly, the presence of divergent sequences from fungi, plants, and other non-metazoan groups suggests ancient evolutionary origins and potential functional analog.

Conclusion

Overall, this study investigates the protein–protein interactions between proteins belonging to distinct structural folds. Transthyretin (TTR), classified within the all-beta class and forming part of the transthyretin-like family, was examined alongside its interacting partner Retinol Binding Protein 4 (RBP4), which also falls under the all-beta class but belongs to a different structural lineage. Similarly, Beta-Actin, a member of the actin-like ATPase fold within the alpha/beta (α/β) class, was analyzed in relation to its binding partner Profilin-1 (PFN-1), which belongs to the profilin-like family of the alpha and beta (α+β) class. This comparative analysis highlights how proteins from varied structural classes engage in biologically significant interactions.

By using various analyses like sequence-based, structure-based and phylogenetic approaches, it was found that conserved residues in these proteins also have significant functional importance. Mutations of these were also found to have implications in disease conditions.

References

1. Lodish H, Berk A, Kaiser CA, Krieger M, Bretscher A. Molecular Cell Biology. W H Freeman & Company; 2012. 1154 p.

2. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. EMBO J. 2003 Jul 15;22(14):3486–92.

3. Keskin O, Tuncbag N, Gursoy A. Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. Chem Rev. 2016 Apr 27;116(8):4884–909.

4. Shao J, Welch WJ, Diprospero NA, Diamond MI. Phosphorylation of profilin by ROCK1 regulates polyglutamine aggregation. Mol Cell Biol. 2008 Sep;28(17):5196–208.

5. Read TA, Cisterna BA, Skruber K, Ahmadieh S, Lindamood HL, Vitriol JA, et al. The actin binding protein profilin 1 is critical for mitochondria function [Internet]. bioRxiv. 2023. Available from: http://dx.doi.org/10.1101/2023.08.07.552354

6. Witke W. The role of profilin complexes in cell motility and other cellular processes. Trends Cell Biol. 2004 Aug;14(8):461–9.

7. Drazic A, Aksnes H, Marie M, Boczkowska M, Varland S, Timmerman E, et al. NAA80 is actin's N-terminal acetyltransferase and regulates cytoskeleton assembly and cell motility. Proc Natl Acad Sci U S A. 2018 Apr 24;115(17):4399–404.

8. dos Remedios C, Chhabra D. Actin-Binding Proteins and Disease. Springer Science & Business Media; 2008. 362 p.

9. Dominguez R, Holmes KC. Actin structure and function. Annu Rev Biophys. 2011;40:169–86.

10. Herbert J, Wilcox JN, Pham KT, Fremeau RT Jr, Zeviani M, Dwork A, et al. Transthyretin: a choroid plexus-specific transport protein in human brain. The 1986 S. Weir Mitchell award. Neurology. 1986 Jul;36(7):900–11.

11. Colon W, Kelly JW. Partial denaturation of transthyretin is sufficient for amyloid fibril formation in vitro. Biochemistry. 1992 Sep 15;31(36):8654–60.

12. Peterson PA. Studies on the interaction between prealbumin, retinol-binding protein, and vitamin A. J Biol Chem. 1971 Jan 10;246(1):44–9.

13. Newcomer ME, Ong DE. Plasma retinol binding protein: structure and function of the prototypic lipocalin. Biochim Biophys Acta. 2000 Oct 18;1482(1-2):57–64.

14. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2007 Nov 13;36(suppl_1):D419–25.

15. Mallika V, Bhaduri A, Sowdhamini R. PASS2: a semi-automated database of Protein Alignments Organised as Structural Superfamilies. Nucleic Acids Res. 2002 Jan

1;30(1):284–8.

16. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2018 Nov 22;47(D1):D607–13.

17. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct 5;215(3):403–10.

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

19. Iyer MS, Bhargava K, Pavalam M, Sowdhamini R. GenDiS database update with improved approach and features to recognize homologous sequences of protein domain superfamilies. Database (Oxford). 2019 Apr 3;2019:baz042.

20. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010 Jan 6;26(5):680–2.

21. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014 Jan;42(Database issue):D304–9.

22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

23. Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, et al. PDBe: Protein Data Bank in Europe. Nucleic Acids Res. 2014 Jan;42(Database issue):D285–91.

24. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci. 2018 Jan;27(1):129–34.

25. Rosignoli S, Paiardini A. Boosting the Full Potential of PyMOL with Structural Biology Plugins. Biomolecules [Internet]. 2022 Nov 27;12(12). Available from: http://dx.doi.org/10.3390/biom12121764

26. Lassmann T, Sonnhammer ELL. Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005 Dec 12;6:298.

27. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772–80.

28. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 19;32(5):1792–7.

29. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct 11;7:539.

30. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate

multiple sequence alignment. J Mol Biol. 2000 Sep 8;302(1):205–17.

31. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005 Apr 22;33(7):2302–9.

32. Holm L. Dali server: structural unification of protein families. Nucleic Acids Res. 2022 Jul 5;50(W1):W210–5.

33. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008 Apr;36(7):2295–300.

34. Wong TKF, Kalyaanamoorthy S, Meusemann K, Yeates DK, Misof B, Jermiin LS. A minimum reporting standard for multiple sequence alignments. NAR Genom Bioinform. 2020 Apr 14;2(2):lqaa024.

35. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018 Jun 1;35(6):1547–9.

36. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2;49(W1):W293–6.

37. Procter JB, Carstairs GM, Soares B, Mourão K, Ofoegbu TC, Barton D, et al. Alignment of Biological Sequences with Jalview. Methods Mol Biol. 2021;2231:203–24.