Classification and Automatic Annotation of Tandem Repeat Proteins in RepeatsDB

by

Soroush Mozaffari

Supervisor: Prof. Silvio Tosatto

A Master's Thesis Submitted to

The Department of Pharmaceutical and Pharmacological Sciences

University of Padova

In Fulfillment of the Requirements

For the Degree of Master of Science in

Pharmaceutical Biotechnologies

2022 - 2023

Contents

Abstract

Protein tandem repeats are crucial structural elements in various biological processes, playing essential roles in cell adhesion, protein-protein interactions, and molecular recognition. These repetitive regions have sparked considerable interest in structural biology and bioinformatics, leading to the development of specialized resources like RepeatsDB. RepeatsDB is a comprehensive, curated database of annotated tandem repeat protein structures, offering a valuable resource for researchers. In this study, we systematically analyzed protein tandem repeats in RepeatsDB, with a primary focus on Alpha-Solenoids and Beta-Propellers, to enhance the existing classification system and provide a more profound understanding of protein tandem repeats. Our investigation commenced with an initial statistical analysis to elucidate the diversity and population status of distinct repeat groups within the database, as well as their respective degree of annotation. This approach proved instrumental in addressing the challenges associated with numerous entries that had a missing annotation. We conducted a structural analysis using pairwise structural alignment and explored dimensionality reduction and visualization techniques to uncover novel structural relationships. These findings improved our understanding of protein structural comparisons and informed a refined classification system. We utilized the density-based clustering algorithm, DBSCAN, to establish structural similarity ranges for Clan members and provide computational support for defining Clan boundaries. This method proved effective in detecting outlier entries and refining existing clans, leading to the proposal of new repeat groups. Additionally, we implemented a supervised classification experiment using the K-Nearest Neighbors (KNN) algorithm, which facilitated the automatic annotation of previously unannotated entries. This study introduces an automatic annotation methodology that significantly improves the performance of RepeatsDB curators and can be extended to other bioinformatics applications. The findings contribute to a more comprehensive understanding of protein tandem repeats and offer valuable insights for future research in structural biology and bioinformatics.

1. Introduction

1.1 The Central Dogma of Biology: Understanding the Blueprint of Life

The Central Dogma of molecular biology is a fundamental principle describing the unidirectional flow of genetic information in living organisms, from DNA to RNA to protein. Initially proposed by Francis Crick in 1958 [1], this concept has subsequently established itself as a cornerstone in the field of molecular biology.

The Central Dogma encompasses three stages. Firstly, transcription involves the conversion of DNA-encoded information into RNA, a process taking place in the nucleus of eukaryotic cells and the cytoplasm of prokaryotic cells. Secondly, translation entails the decoding of RNA-encoded information into protein, occurring within the ribosomes of all cells. Lastly, the protein undergoes folding and post-translational modifications to achieve its final, functional form.

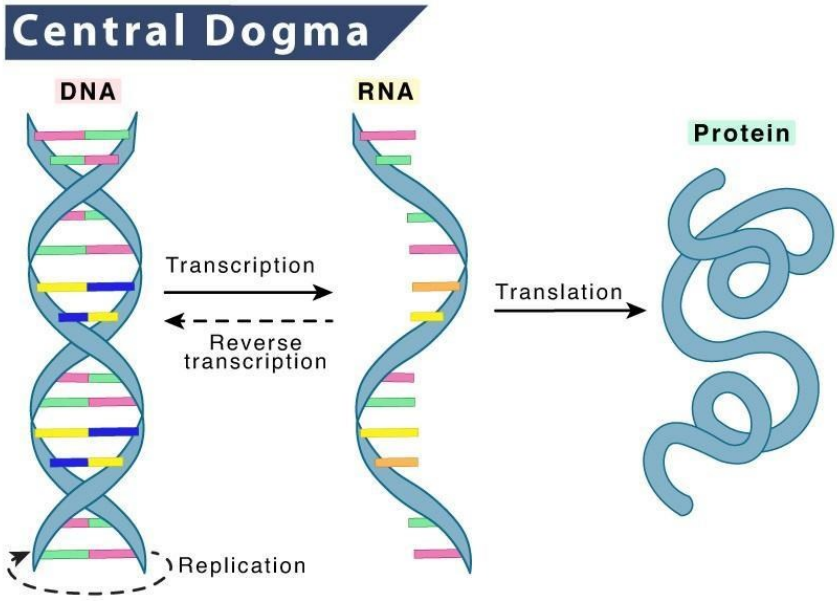


Fig. 1.1: Diagram of the central dogma of biology

Although the Central Dogma provides a fundamental framework for comprehending the flow of genetic information, exceptions do exist. For instance, certain viruses employ RNA as their genetic material, while others utilize reverse transcription to revert RNA back into DNA. Moreover, post-

transcriptional and post-translational modifications can modify the ultimate structure and function of proteins.

Despite these deviations, the Central Dogma remains an instrumental tool for understanding the molecular underpinnings of life. It elucidates the transmission of genetic information across generations and the emergence of genetic variation through mutations and other genetic recombination events. Over time, these alterations may result in the development of new species exhibiting distinct traits and functions. Consequently, the Central Dogma serves as a foundation for understanding the molecular basis of evolutionary change.

Additionally, the Central Dogma has significant implications for biotechnology and genetic engineering. By comprehending the transfer of genetic information from DNA to protein, scientists can manipulate this process to synthesize novel proteins with desired functionalities. For instance, genetic engineering has facilitated the production of insulin for diabetes treatment [2] and the creation of crops with enhanced resistance to pests and environmental stressors [3].

By exploring the properties and interactions of amino acids, scientists can harness the power of the Central Dogma to design innovative solutions for various challenges in biotechnology and genetic engineering, further expanding our capacity to improve human health and address global concerns.

1.2 Amino Acids: Structure and Classification

Amino acids, the fundamental building blocks of proteins, play a crucial role in numerous biological processes within living organisms. These organic compounds consist of an amino group (-NH2), a carboxyl group (-COOH), and a distinct side chain specific to each amino acid. The organization of atoms within amino acids is vital for their functionality and contribution to protein synthesis.

1.2.1 Structure of Amino Acids

Each amino acid features a central carbon atom (alpha carbon) connected to four distinct groups: an amino group, a carboxyl group, a hydrogen atom, and a side chain (R-group). The unique properties and functions of amino acids are determined by their R-groups.

The amino group (-NH2) consists of a nitrogen atom bonded to two hydrogen atoms, while the carboxyl group (-COOH) is made up of a carbon atom double-bonded to an oxygen atom and single-bonded to a hydroxyl group (-OH). A hydrogen atom (H) is also connected to the alpha carbon.

The R-group, or side chain, is responsible for the chemical properties of each amino acid. R-groups can range from a simple hydrogen atom, as in glycine, to a more intricate arrangement of atoms, as seen in tryptophan. The functional groups within the R-group determine whether an amino acid is nonpolar, polar, acidic, or basic, which in turn affects its solubility, polarity, and reactivity.

1.2.2 Classification of Amino Acids

Based on the characteristics of their R-groups, amino acids can be classified into four main categories:

- Nonpolar amino acids: These amino acids possess hydrophobic R-groups that do not interact with water. Examples include alanine, valine, and leucine.
- Polar amino acids: These amino acids have hydrophilic R-groups that interact with water. Examples are serine, threonine, and cysteine.
- Acidic amino acids: These amino acids feature negatively charged R-groups. Aspartic acid and glutamic acid are examples.
- Basic amino acids: These amino acids contain positively charged R-groups. Lysine, arginine, and histidine serve as examples.

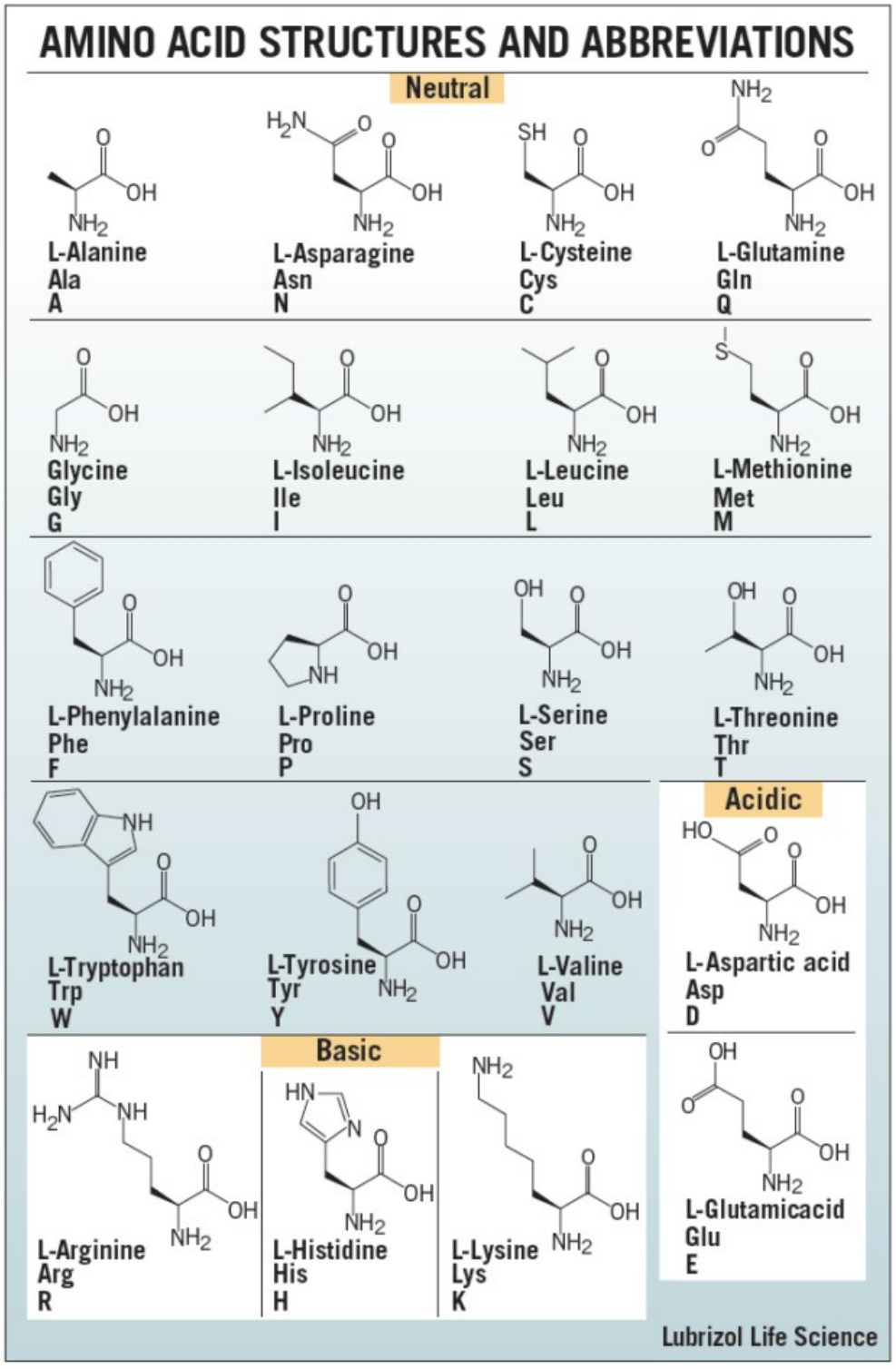


Fig. 1.2: Molecular structure and classification of amino acids

1.3 Primary Structure of Proteins

1.3.1 Peptide Bonds

Peptide bonds are covalent linkages that connect amino acids, forming the protein backbone. A peptide bond is established through a condensation reaction between the carboxyl group of one amino acid and the amino group of another. The resultant structure, which constitutes the protein's backbone, exhibits partial double bond characteristics due to resonance between the carbonyl group and the nitrogen atom. This resonance imparts planarity and rigidity to the peptide bond, significantly impacting higher-order protein structures. Additionally, peptide bonds possess a dipole moment, with a slightly negative carbonyl group and a slightly positive nitrogen atom.

1.3.2 Protein Backbone

The primary structure of a protein is composed of its amino acid sequence, which is connected by a series of peptide bonds to form the protein backbone. This backbone is flexible thanks to the rotation of single bonds between the alpha carbon and the amino and carboxyl groups. Hydrogen bonds between the carbonyl group of one amino acid and the amide nitrogen of another stabilize the backbone, resulting in a recurring pattern of carbonyl groups and amide nitrogens known as the backbone hydrogen bond pattern. The protein backbone is crucial for the protein's structure and function, as it provides the framework for the protein's three-dimensional conformation.

1.4 Secondary Structure

The secondary structure of a protein encompasses the regular, repetitive patterns of local structure within the protein. These patterns significantly impact protein folding, stability, and functionality.

1.4.1 Types of Secondary Structure

Proteins primarily consist of two secondary structure types: alpha helices and beta sheets. Alpha helices are right-handed helical structures stabilized by hydrogen bonds between the carbonyl group of one amino acid and the amide nitrogen of another amino acid, four residues away in the sequence. Beta sheets are planar structures maintained by hydrogen bonds between the carbonyl

and amide groups of neighboring amino acids in the sequence. Depending on the orientation of the strands, beta sheets can be parallel or antiparallel.

1.4.2 Formation of Secondary Structure

Secondary structures arise from interactions between peptide bonds in the protein backbone, specifically between nearby amino acids. These interactions produce regular, repetitive patterns of structure primarily driven by hydrogen bonding between the backbone's carbonyl and amide groups. In alpha helices, hydrogen bonds form between carbonyl and amide groups four residues apart, while in beta sheets, hydrogen bonds form between adjacent carbonyl and amide groups.

1.4.3 Importance of Secondary Structure

The secondary structure is crucial for protein folding and stability. The regular, repetitive patterns allow the protein to adopt a specific conformation essential for function. The hydrogen bonding that stabilizes the secondary structure also contributes to the protein's overall stability. Furthermore, secondary structures can influence the protein's properties, such as solubility, flexibility, and capacity to interact with other molecules.

1.5 Super Secondary Structures

Super-secondary structures emerge from the interactions between secondary structure elements within larger structural units. These interactions can be facilitated by various mechanisms, including hydrogen bonding, hydrophobic interactions, and electrostatic interactions. In some instances, super-secondary structures may be stabilized by additional interactions, such as disulfide bonds or metal ions. The specific interactions that stabilize a super-secondary structure depend on the protein's amino acid sequence and its environment.

1.5.1 Importance of Super-Secondary Structures

Super-secondary structures contribute significantly to protein function and stability by providing additional structural organization, which can improve the protein's stability and functionality. Classifying proteins into structural families based on their super-secondary structures offers insights into their functions and evolutionary relationships [4].

1.6 Tertiary Structure

The tertiary structure of a protein represents its three-dimensional conformation, determined by the interactions between its amino acid residues. Tertiary structure is crucial for protein function and stability and is a key factor in the formation of higher-order structures, such as protein complexes.

1.6.1 Formation of Tertiary Structure

The protein's tertiary structure arises from the complex interplay of non-covalent interactions between its amino acid residues. These interactions can involve hydrogen bonding, electrostatic interactions, hydrophobic interactions, and van der Waals forces. The protein's amino acid sequence dictates the types and locations of residues involved in these interactions, ultimately determining the protein's three-dimensional conformation.

Proteins can adopt various shapes and configurations within their tertiary structure. Some common types include:

- **Globular proteins**, which possess a compact, rounded shape and are frequently involved in enzyme catalysis, transport, or signaling.
- **Fibrous proteins**, which exhibit an elongated, thread-like shape and often contribute to structural support, such as collagen.

The protein's tertiary structure is vital for its function and stability. The specific conformation determines the protein's capacity to interact with other molecules, including enzymes, receptors, and other proteins. The stability of the tertiary structure is essential for maintaining the protein's functionality under various conditions, such as changes in temperature, pH, and the presence of denaturants. Furthermore, the tertiary structure plays a crucial role in forming higher-order structures, such as protein complexes. The interactions between proteins within a complex are dictated by their tertiary structures and changes in tertiary structure can disrupt these interactions, potentially leading to altered protein function or even disease.

1.7 Quaternary Structure

The quaternary structure refers to the assembly of individual protein subunits to form a functional protein complex. Subunits can be identical (homodimers or homo-oligomers) or different (heterodimers or hetero-oligomers) and can come together in various ways, including linear chains, branched chains, and symmetric or asymmetric complexes [5]. Although not all proteins have quaternary structures, when present, they play a critical role in understanding protein function within complex biological systems.

A well-known example of a quaternary protein structure is hemoglobin, which is responsible for carrying oxygen in the blood. This tetrameric protein is composed of two alpha and two beta subunits. The assembly of these subunits creates a pocket essential for oxygen binding and protein function [6].

The quaternary structure can significantly impact a protein's function. For instance, some proteins have allosteric sites that can bind to small molecules, leading to changes in the protein's structure or function [7].

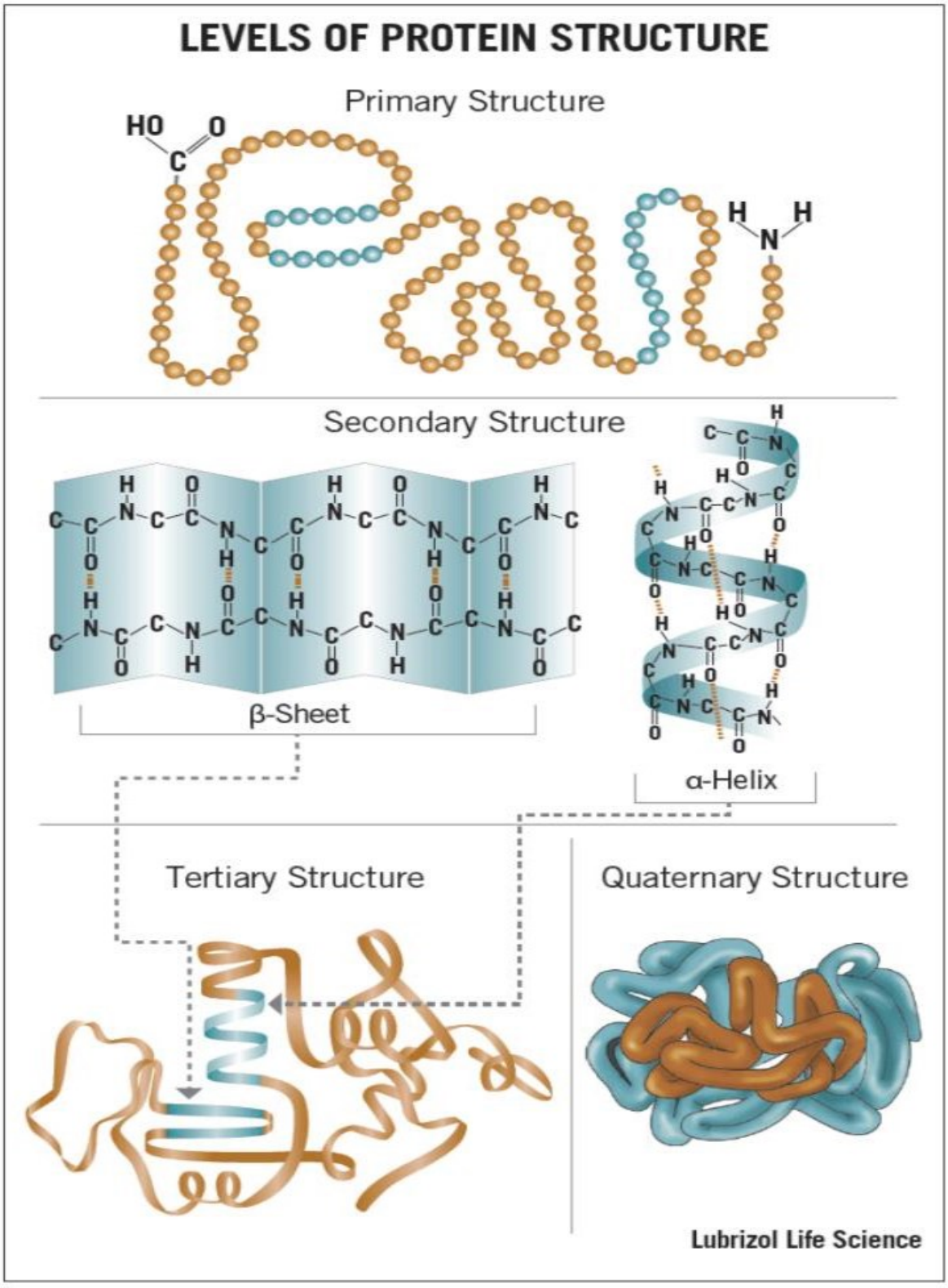


Fig. 1.3: Four types of protein structure

1.8 Protein Folding Landscape

The protein folding landscape is the energy landscape that guides protein folding, ultimately forming the tertiary structure. Influenced by factors such as denaturant concentration, temperature, and the presence of chaperones, the energy landscape can be graphically represented as a funnel-shaped curve. At the top, the unfolded or denatured state of the protein is present, while the native or folded state is at the bottom. As the protein folds, it transitions through a series of intermediate states, each with its own energy level and conformation.

The folding funnel is highly complex, and the protein folding process can be slowed down or halted by kinetic traps. These traps emerge when the protein encounters an energy minimum that is not the correct conformation for the folded protein. The protein can become trapped in these states, preventing proper folding.

Understanding the protein folding landscape is crucial for studying protein folding and misfolding, which can lead to protein aggregation and, in some cases, disease. Disorders such as Alzheimer's, Parkinson's, and cystic fibrosis are associated with protein misfolding. Gaining insights into the energy landscape can provide valuable information about the underlying causes of these diseases [8].

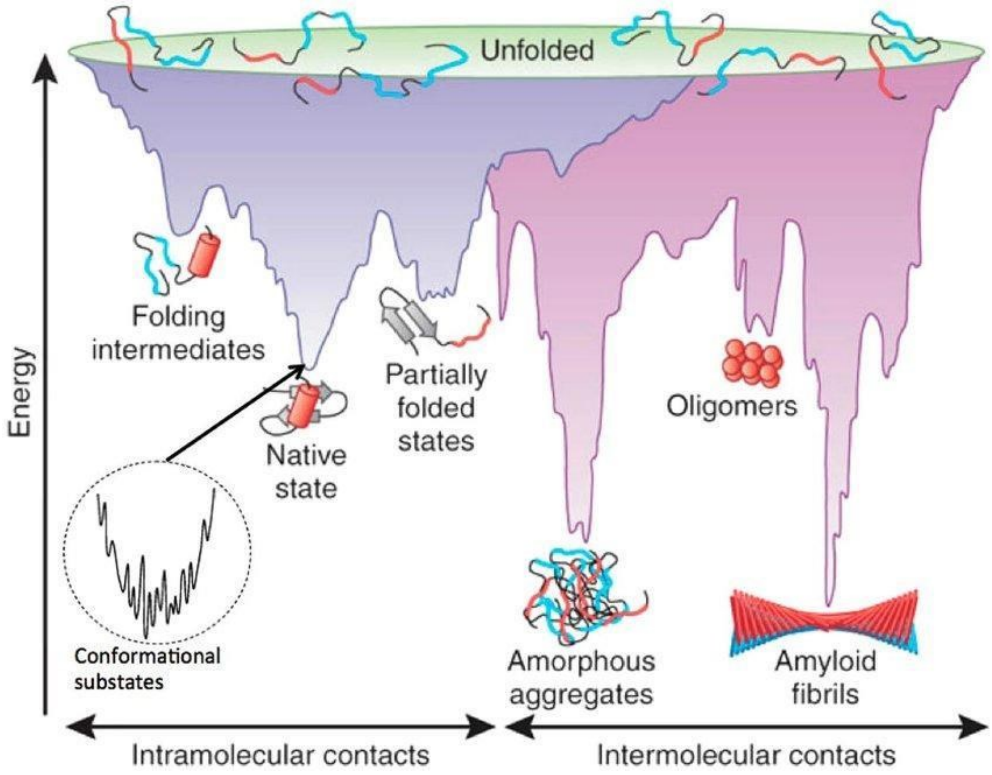


Fig. 1.4: A diagram illustrating the protein folding process and various states of protein folding

1.9 Globular VS Fibrous

As was previously mentioned in brief, proteins can be broadly classified into two main categories based on their shape and function: globular proteins and fibrous proteins.

Globular proteins are known for their compact, approximately spherical shape, which generally results from the protein folding into a globular form in solution. These proteins tend to be water-soluble and possess a hydrophobic interior and a hydrophilic exterior, making them well-adapted to functions such as enzyme catalysis, signal transduction, and transport. The interior of a globular protein is often densely packed with hydrophobic amino acids, creating a stable environment for the protein to perform its function.

Myoglobin serves as an example of a globular protein, playing a role in the storage and transport of oxygen in muscle tissue. Composed of a single polypeptide chain, myoglobin has a compact, globular structure that enables efficient oxygen binding and release [9]. This shape allows the protein to be transported with ease through the blood and into muscle tissue.

In contrast, fibrous proteins are elongated and play a structural role in the body. These proteins are typically insoluble in water and consist of long, repetitive sequences of amino acids, providing a high degree of strength and stability. Fibrous proteins usually form long, thread-like structures well-suited for support and protection, such as keratin, which makes up hair and nails, and collagen, the primary structural protein in connective tissue.

Although the distinction between globular and fibrous proteins is not always strictly defined, these two categories represent fundamentally different protein types with distinct roles in the body [5].

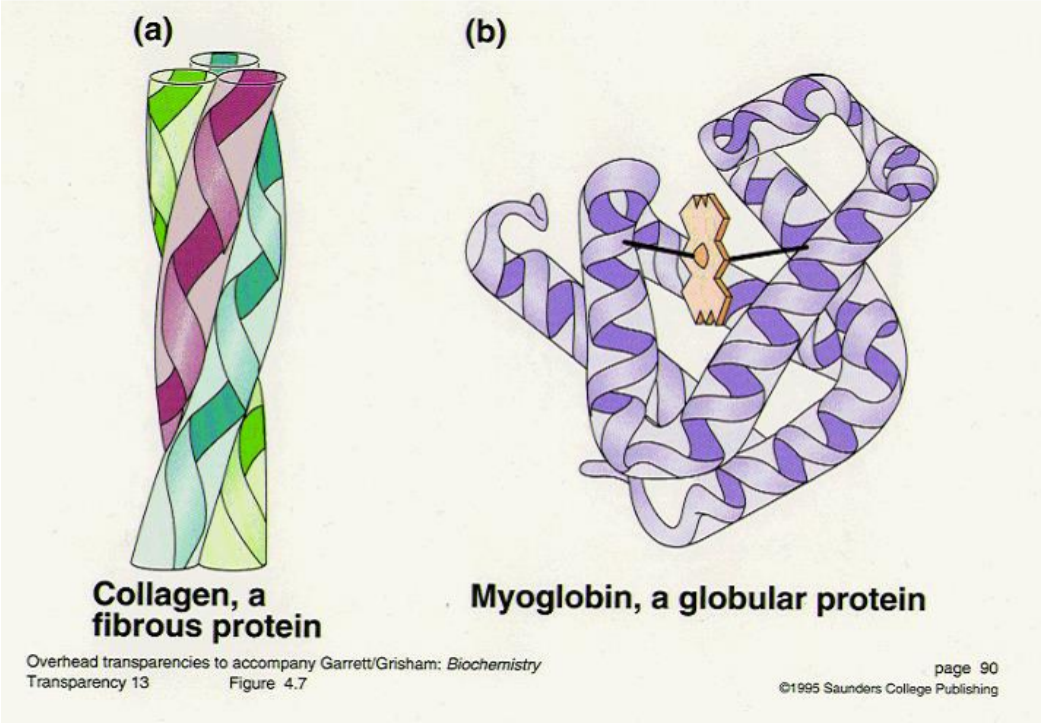


Fig. 1.5: A structural comparison of a fibrous protein (collagen) and a globular protein (myoglobin)

1.10 Tandem Repeats

Tandem repeats are short sequences of DNA or amino acids that are repeated in a head-to-tail fashion. In DNA, tandem repeats are classified based on their length and organization. Short tandem repeats (STRs) are usually composed of simple, repetitive motifs of 2-6 nucleotides, and are usually found in non-coding regions of the genome [10]. On the other hand, longer tandem repeats can span tens or hundreds of base pairs, and are often found in coding regions of the genome. These long tandem repeats can affect gene expression, splicing, and stability [11].

On the other hand, in proteins, based on the pioneering research of Andrey V. Kajava, a leading expert in the field of tandem repeat protein research, a new and more comprehensive classification of tandem repeat proteins was proposed in 2012. The updated classification takes into account the diverse structures and functions of these proteins and is summarized as follows: 1. Crystalline aggregates, 2. Fibrous, 3. Elongated, 4. Closed, and 5. Beads on a string [12]. Later on, each of these classes will be elaborated upon in detail.

Tandem repeat proteins are prevalent in a wide variety of organisms and are found in a significant proportion of protein sequences. In fact, it has been estimated that tandem repeat-containing proteins make up approximately 14% of all proteins [12].

Tandem repeats in DNA and proteins can have different functions and evolutionary consequences. In DNA, tandem repeats can affect gene expression and stability. For example, the expansion of tandem repeats in the promoter regions of genes can affect transcription, leading to diseases such as Fragile X syndrome [13]. In contrast, in proteins, tandem repeats are often involved in protein-protein interactions and structural stability. For example, the immunoglobulin superfamily of proteins contains tandem repeats that mediate protein-protein interactions [14].

1.10.1 Diverse Roles of Tandem Repeat Proteins

Tandem repeat proteins are associated with a diverse array of functions such as:

- **Structural Suppor**

Tandem repeat proteins play a crucial role in providing structural support to cells and tissues. They can form specific structural motifs, such as alpha-helices, beta-sheets, and beta-turns, which are important for maintaining the shape and stability of proteins. For example, the protein elastin, which contains tandem repeat domains, provides elasticity to tissues such as blood vessels and skin [15]. Similarly, the protein collagen, which contains repeating sequences of glycine, proline, and hydroxyproline, provides strength and stability to tissues such as bone and cartilage [16].

- **Signal Transduction**

Tandem repeat proteins also play an important role in signal transduction pathways. They can act as scaffolds for the assembly of signaling complexes, allowing for efficient and specific signaling

between cells. For example, the protein beta-catenin, which contains tandem repeat domains, is a signaling protein that plays an important role in cell adhesion and the regulation of gene expression [17].

- **Immune System Function**

Tandem repeat proteins can also act as antigens that stimulate antibody production and activate immune responses. One such protein is C4BP, which contains tandem repeat domains and serves as a regulator of the complement system involved in immune response regulation. [18].

- **Genetic Markers**

The high degree of variability in tandem repeat proteins can also contribute to genetic diversity within populations. This makes them useful in a number of applications, including as genetic markers in population studies and in biotechnology. For example, the number of repeats in the protein D4Z4 has been linked to facioscapulohumeral muscular dystrophy (FSHD), a neuromuscular disorder [19].

- **Vesicle Formation**

Tandem repeat proteins are crucial for the regulation and formation of vesicles, small membrane-bound compartments that transport and deliver proteins and other molecules within cells. One example of such vesicles is clathrin-coated vesicles, formed via clathrin-mediated endocytosis. Clathrin proteins, which contain specific types of tandem repeat domains, are essential components of this process [20]. They assemble into a lattice on the cytoplasmic side of the plasma membrane and recruit other proteins and adaptors to form a coated pit. The coated pit then invaginates and eventually buds off from the membrane to form a clathrin-coated vesicle [21]. The figure below illustrates the constituent elements of the clathrin-coated vesicles.

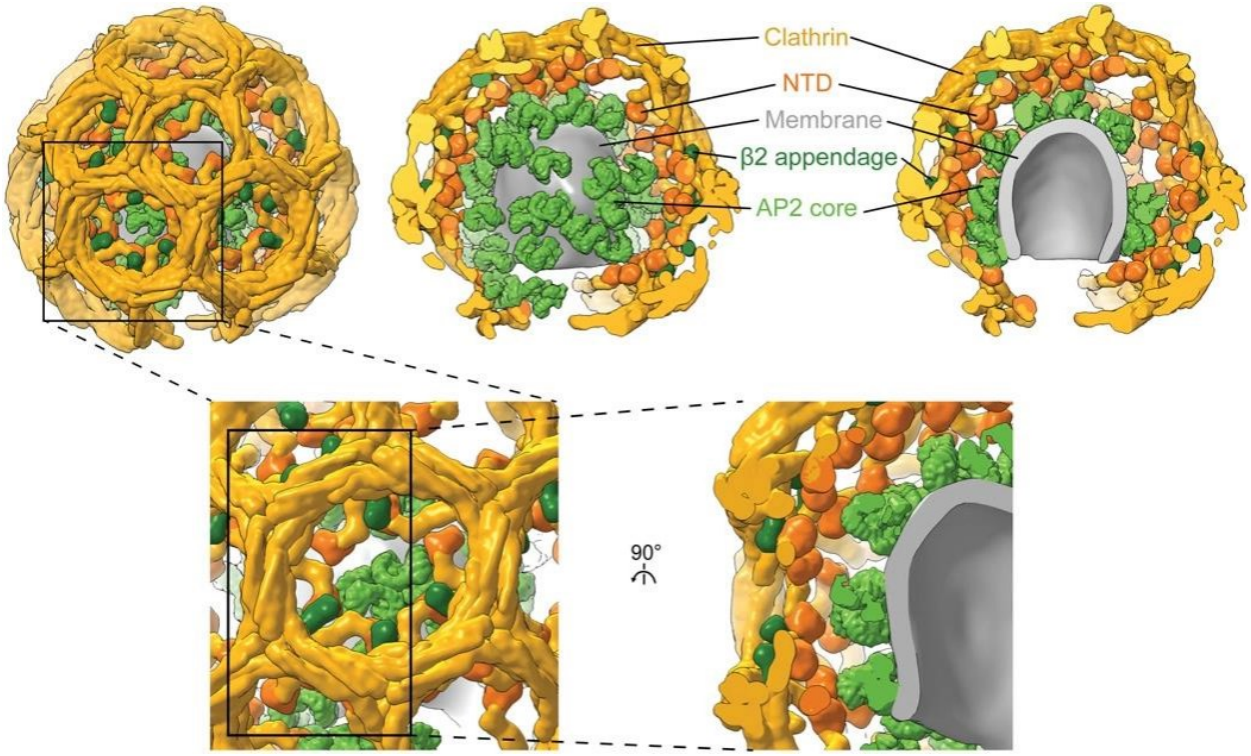


Fig. 1.6: A depiction of a clathrin-coated vesicle, highlighting its structural components and assembly

1.11 Composition of Tandem Repeat Proteins

Tandem repeats in proteins are consisted of three main elements:

- **Repeat Region:** The entire segment of the protein where the tandem repeats occur. Its length can significantly vary, and it can be found anywhere within the protein sequence.
- **Repeat Units:** The individual segments repeated in the region, often similar in structure but also can be significantly degenerate on the sequence level [22]. They can be single residues or a stretch of multiple residues.
- **Insertions:** They represent unique structural segments situated within a repeat unit or between two units. They disrupt the symmetrical pattern of the repeat sequence and pose challenges for both automated identification and structural analysis of repeated sequences [23].

The following figure represents these 3 main elements:

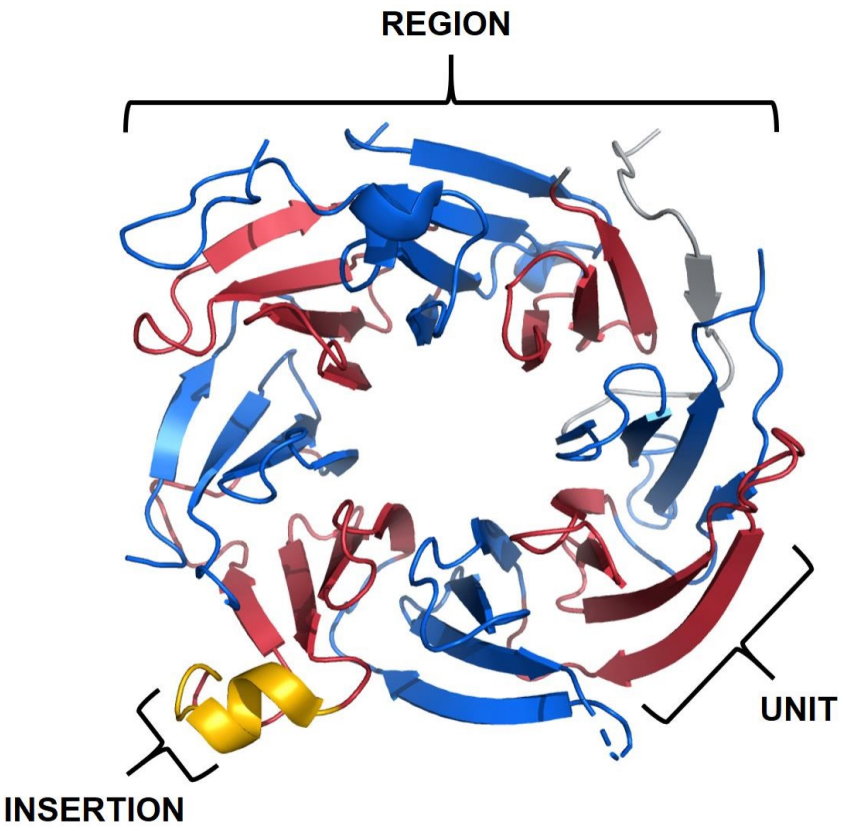


Fig. 1.7: A schematic representation of a protein tandem repeat, detailing its key elements

1.12 Classification of Protein Tandem Repeats

Protein tandem repeats were previously classified solely based on repeat lengths [24]. However, as previously mentioned, a more refined classification has been proposed by Andrey V. Kajava, due to the discovery of new 3D structures. The new classification takes into account additional factors such as repeat secondary structure composition, topologies, and interfaces[12]. Briefly, the classification is organized as it follows:

- **Class I: Crystalline aggregates**

This category consists of proteins and peptides that have repeating units of 1 or 2 residues. These repeating units can form various types of crystal structures that can grow to any size and can be detrimental to living organisms. In proteomes, regions that contain these repeating units are typically hydrophilic and are prone to be unfolded.

On the other hand, the formation of crystalline aggregates can also be detrimental to cells and tissues. Many neurodegenerative diseases, such as Alzheimer's and Parkinson's, are associated with the accumulation of insoluble protein aggregates in the brain. These aggregates are thought to disrupt normal cellular processes and contribute to disease pathology [25].

Further research is needed to fully understand the mechanisms underlying the formation of crystalline aggregates in proteins and to develop new materials and therapies based on this knowledge.

- **Class II: Fibrous**

This category comprises two prominent fibrous structures: collagen and alpha-helical coiled coils. Collagen is composed of a tripeptide sequence Gly-X-Y, where the residues X and Y can vary but commonly consist of proline or hydroxyproline. The chains adopt an elongated polyproline II conformation and form a triple helix structure in the 3D conformation [26].

The alpha-helical coiled coils are identified by a repetitive $(abcdefg)_n$ motif, where hydrophobic residues occupy positions “a” and “d”, and polar residues occupy other positions. The apolar residues “a” and “d” are spaced at intervals of 3-4 residues. Each chain folds into an alpha-helix and intertwines around the axis of the coiled-coil structure [27].

- **Class III: Elongated**

Elongated repeats are composed of two sub-categories of Solenoidal structures and Non-solenoidal structures. In the first group, repeats of 5–40 residues adopt a solenoidal structure, where the polypeptide chain is wound in a spiral fashion. As a consequence, solenoid proteins tend to exhibit elongated structures, which are distinct from the globular shape commonly observed in other proteins [28].

An example of this is the ‘‘β-spiral’’ structure of the adenovirus fiber shaft having 15 residue repeats. The β -spiral folds exhibit greater complexity when compared to the solenoidal fold due to the presence of elongated central beta-strands, which maintain the trimer through inter-chain hydrogen bonding. Furthermore, the structure is stabilized by interactions involving the nonpolar side chains and short peripheral beta-strands [29].

The individual units of Elongated repeats are interdependent for preserving the overall structure. However, unlike class II proteins, these repeating units can still possess a stable structure independently without assembling into an oligomeric form [12].

- **Class IV: Closed**

Proteins belonging to this category exhibit a fixed number of repeats owing to their circular or "closed" structures. The length of these repeats overlaps with structures classified under both classes 3 and 5 with, proximately, ranging from 40 to 60 residues. The WD repeat is one of the most well-known structures belonging to this class. These proteins fold into a closed beta-propeller structure, where each repeat is a four-stranded antiparallel beta-blade. A single repeat by itself is not stable, but multiple repeats can form a stable beta-propeller structure where the first and last blades interact to close the symmetric structure [30].

- **Class V: Beads on a string**

This category of structures consists of repetitive units that are large enough to independently fold into stable domains, typically containing over 50-60 residues. The protein structure with these domains follows a "beads on a string" arrangement, where the beads correspond to globular domains. The smallest domains of this type start at around 30 residues but require disulfide bonds or metal ion coordination to stabilize their structure. Zn-finger domains are a classic example of such structures, being the most common DNA-binding motif that is stabilized by zinc metal ion coordination. The flexible connections between Zn-finger domains and the diversity of their binding surfaces enable specific interactions with various DNA motifs [31].

In the following figure, several structures and domains associated with the 5 (I-V) main classes of protein tandem repeats are illustrated.

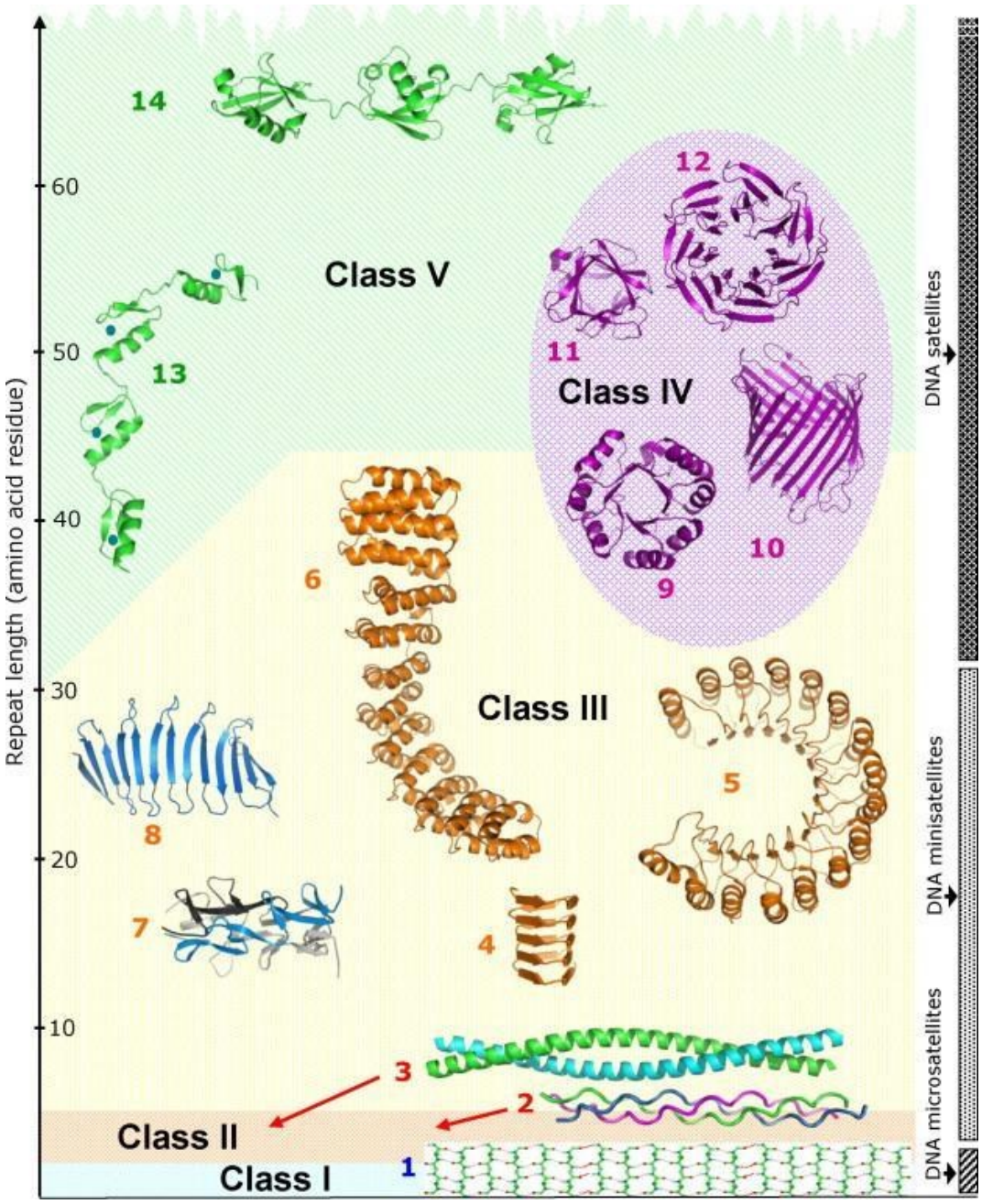


Fig. 1.8: The structural classification of the repetitive proteins based on the length of their repeats. The examples shown are: class I, 1 – crystalline b-structure of poly-alanine, class II, 2 – triple-helix of collagen, 3 – a-helical coiled coil, class III, 4 – b-solenoid, 5 – a/b-solenoid of LRR protein, 6 – a-solenoid, 7 – trimer of b-spirals, 8 – single-layer antiparallel b-structure, class IV, 9 – TIM-barrel, 10 – transmembrane b-barrel, 11 – b-trefoil structure, 12 – b-propeller structure, and Class V, 13 – Zn-finger domains, 14 – polyubiquitin chain.

1.13 Identification of Protein Tandem Repeats

Over recent years, numerous algorithms and software have emerged to identifying tandem repeats in proteins. These methods are primarily based on either the protein sequence or its structure.

1.13.1 Sequence-based methods

Sequence-based methods can be broadly categorized into five distinct types, each with its own advantages and drawbacks:

1. The first type employs Fourier Transform analysis to detect periodicities in amino acid sequences. As a de novo or ab initio method, it does not depend on prior knowledge of potential repeats. This approach excels in identifying long tandem repeat arrays without insertions or deletions, but struggles with short repeats and those containing indels [32].

2. The second type, exemplified by programs like XSTREAM[33] and T-REKS [34], utilizes short string extension algorithms. These methods can detect tandem repeats with indels and are particularly adept at ab initio identification of short repeats (fewer than 15-20 residues). With linear time complexity, they are fast and well-suited for large-scale repeat searches.

3. The third type identifies repeats through sequence-sequence alignment algorithms that compare the protein sequence to itself. Web servers such as RADAR [35] and TRUST [36] embody this approach, which is effective for ab initio detection of long repeat arrays. However, it often fails to detect short repeats, and its $O(n^2)$ time complexity renders it relatively slow and ill-suited for large-scale analysis.

4. The fourth type relies on pre-generated alignments of repeats, which are used to build Hidden Markov Models (HMMs) or sequence profiles. These HMM profiles are then compared to the query sequence to find the best and multiple hits [37]. Web servers such as Pfam, SMART, REP, TPRpred, PROSITE, and BiSMM server employ this method. While it excels in detecting long and highly imperfect tandem repeats, its dependence on prior alignments precludes it from automated ab initio large-scale analysis.

5. The fifth type of sequence-based method involves HMM-HMM or profile-profile comparisons, as seen in the HHrepID server [38]. It constructs an HMM from a multiple sequence alignment of homologous proteins and searches for sub-optimal alignments within the HMM. This highly sensitive approach allows efficient ab initio detection of highly divergent "covert" tandem repeats.

1.13.2 Structure-based methods

Structure-based methods aim to identify the periodicity of protein tandem repeats using various strategies, such as detecting 3D symmetries or examining the regularity of structural features. These methods differentiate between repeat and non-repeat structures and sometimes annotate TRPs with unit positions.

Examples of structure-based methods include ConSole and TAPO. ConSole identifies protein modularity using contact maps [39], while TAPO detects periodicities in atomic coordinates and other structural representations [40].

Another method, ReUPred [41], identifies structural units by comparing the protein structure to a manually curated library of TRP units, designed to represent the diversity of repeat units in terms of their conformational space. This allows for the identification and classification of modular substructures in TRPs.

1.14 Tandem Repeat Protein Databases

In recent years, numerous databases have been established to support the systematic organization, classification, and analysis of tandem repeat proteins (TRPs). A prime example of such a database is RepeatsDB.

RepeatsDB is an extensive database devoted to the classification and structural analysis of TRPs. It consolidates a vast array of structural and functional information for each protein entry, making it an invaluable resource for researchers examining TRP structure-function relationships and their potential applications across diverse scientific fields. The database employs a hierarchical classification system consisting of four primary levels: Class, Topology, Fold, and Clan. A numerical index is assigned to each tier to ensure a standardized representation of protein structures [42].

In addition to RepeatsDB, other databases have been developed to catalog and analyze TRPs, such as DbStRiPs [43]. These databases offer complementary resources and tools to enable a comprehensive understanding of TRPs and their functions. However, in this study, our analyses are focused solely on the resources provided by RepeatsDB, which will be elaborated upon in detail through the following sections.

1.14.1 RepeatsDB

RepeatsDB is a database that contains information about tandem repeat structures in proteins. The initial dataset is extracted from the Protein Data Bank (PDB), which is a large repository of experimentally determined protein structures. Repeat candidates are identified from the reduced PDB dataset using ReUPred, which uses a geometric approach to imitate the work of a human curator. The resulting dataset is stored in the database as 'predicted' entries, which then undergo a classification and curation process.

The predicted entries are manually curated using a two-level annotation system. In the first manual annotation level ('manually classified'), an entry is classified into a structural repeat class and subclass. This classification is based on previous work where five classes of repeat structures are proposed, which are then further divided into subclasses. Class assignment is based mainly on repeat unit length, and subclass assignment is based on secondary and tertiary structure features.

In the second manual annotation level ('detailed'), information is provided about the start and end positions of the repeat units, repeat regions, and/or insertions. A repeat unit is defined as the smallest structural building block that is repeated to form a repeat region. A repeat region is defined as a group of at least three repeat units. Proteins with two repeat units are not included because they significantly complicate classification, as many typical globular domains have this type of architecture. Insertions, that as previously described are non-repeated segments of structure that occur either inside a repeat unit or between two of them, are also annotated. These are particularly interesting because they break the repeat symmetry and represent a challenge both for automatic detection and for the analysis of repeat structures.

Several curators annotate each protein undergoing manual classification by consensus. For first-level annotations, at least 75% of the curators have to agree for a protein to be included. Otherwise,

it is excluded and placed on a reserve list for future annotation. The rationale for this choice is that ambiguous cases are generally difficult to classify, but may occasionally represent a novel repeat class. For second-level annotations, the threshold for consensus is at least 65% agreement (typically two of three curators). In case of discrepancy, an expert arbitrates the final annotation based on the alternative proposals.

Proteins with detailed annotations are also used to search for similar sequences in proteins from the PDB. Any PDB chain with at least 40% sequence identity and a coverage of at least 80% of the classified protein, belonging to the initial list of predicted entries, is added to the 'classified by similarity' annotation level. The similarity thresholds are implemented to exclude possible false positives.

1.14.2 Classification of TRPs in RepeatsDB

The classification of Tandem Repeat Proteins (TRPs) currently consists of a hierarchical organization that includes four primary tiers, which are delineated as follows:

(I) 'Class' characterizes the overarching protein architecture, with particular emphasis on the general shape, the interplay between repetitive elements, and the oligomerization state, which are contingent on the length of the repeating sequence.

(II) 'Topology', previously designated as 'subclass', differentiates between various configurations of the polypeptide chain, as well as the secondary structure types present within each repetitive unit.

(III) 'Fold' serves as a refinement of the 'topology' tier, providing further distinction based on the arrangement of secondary structures and the overall structural characteristics, such as the degree of twisting, within a specific repeat.

(IV) 'Clan' refers to a subcategory of 'fold' that assembles protein structures that share a common sequence/structure motif within the repeating unit or a portion thereof.

By incorporating these four hierarchical tiers, the classification system for TRPs enables a comprehensive understanding of their structural properties, as well as their relationships with one another. Furthermore, in this classification system, each level of the hierarchical organization has

a numerical index associated with it. This numerical index serves as an identifier for the specific attributes within each classification tier, enabling a standardized representation of the protein structure. Consequently, this systematic approach facilitates communication and comparison between various TRP structures.

The numerical indices assigned to each of the four tiers can be represented as a series of four numerical values separated by periods, following the format: Class. Topology. Fold. Clan. An example of this is shown in the figure below.

Fig. 1.9: A visual representation of the classification hierarchy of an entry in RepeatsDB, accompanied by their corresponding numeric indices

1.15 Where the Problem Arises

The classification of TRPs within RepeatsDB, while comprehensive and valuable, is not without its limitations. Two key areas where the database may be improved are one, the potential flaws in the classification system, and two, the limited degree of statistical support provided to elucidate the classifications. Addressing these issues is crucial to furthering our understanding of TRPs and enhancing the database's utility for the scientific community.

1.15.1 Challenges with the classification of repeats in RepeatsDB

Although the hierarchical classification system employed by RepeatsDB (Class, Topology, Fold, and Clan) has proven effective in organizing and representing TRPs, it is not devoid of flaws. One potential limitation lies in the reliance on the expert-driven manual curation of protein structures. While this approach ensures accurate classification, it may also introduce biases, be time-

consuming, and may not scale efficiently with the rapid influx of newly discovered protein structures. Also, many entries lack a complete annotation, mostly, at the level of Fold and/or Clan, mostly due to the fact that these entries might not completely fit into certain levels of the existing classification.

1.15.2 Limited statistical support for the classification system

Another limitation of the RepeatsDB classification system is the absence of robust statistical support for the assigned categories. While the hierarchical organization allows for a clear and concise representation of TRP structures, it lacks quantitative measures to justify the classification decisions, such as statistical significance or confidence intervals. This makes it difficult to assess the validity and reliability of the classification system and may hinder the interpretation and comparison of different TRP structures.

1.15.3 Addressing these limitations

To overcome the limitations of the current classification system in RepeatsDB, it is essential to develop and implement novel methodologies that leverage automated approaches, particularly machine learning techniques. These methods can significantly improve the classification and clustering of tandem repeat protein structures, enhancing the accuracy, scalability, and objectivity of the database.

Machine learning algorithms offer a promising avenue for addressing the challenges associated with manual curation and the lack of statistical support in the current classification system. By training models on structural data, machine learning techniques can learn complex patterns and relationships, ultimately providing a more robust and data-driven classification framework. Employing these techniques could assist in the manual curation of TRPs and potentially facilitate a more scalable and unbiased classification process.

1.16 The Expanding Role of Machine Learning in Bioinformatics

Machine learning has brought about groundbreaking changes in various facets of the healthcare sector, with bioinformatics standing out as one of the most significantly impacted fields. As a discipline that applies computational methodologies to biological data, bioinformatics has reaped immense rewards from machine learning algorithms. These advancements have propelled progress

in genomics, proteomics, drug discovery, and personalized medicine. A few examples of machine learning applications across different fields of healthcare and biotechnology include:

- **Genomics and Personalized Medicine**

Machine learning has made a considerable impact on bioinformatics, particularly in the domain of genomics, which encompasses the study of an organism's entire DNA set. Machine learning algorithms have found utility in a multitude of genomics tasks, such as:

Genome Sequencing: Machine learning enhances genome sequencing by detecting sequencing inaccuracies, forecasting gene structures, and assembling raw sequence information into whole genomes.

Gene Expression Analysis: Researchers employ machine learning to scrutinize gene expression data, pinpointing patterns and connections that can help comprehend the fundamental molecular mechanisms of diseases and devise targeted treatments.

Genotype-Phenotype Associations: Machine learning algorithms have the capacity to discern correlations between genetic variations and specific phenotypes, like disease susceptibility or drug response. This invaluable information facilitates the development of personalized medicine, enabling the customization of treatments based on an individual's genetic composition. [44].

- **Proteomics**

Proteomics, the large-scale study of proteins, represents yet another field where machine learning has made considerable strides. By investigating protein structures, interactions, and functions, researchers can glean insights into cellular processes and pinpoint potential drug targets. Some notable machine learning applications in proteomics encompass:

Protein Structure Prediction: Machine learning algorithms are capable of forecasting the three-dimensional structure of proteins based on their amino acid sequences. This crucial information aids in drug design and elucidating protein function.

Protein-Protein Interaction (PPI) Prediction: Identifying PPIs is essential for comprehending cellular processes and disease mechanisms. Machine learning approaches can predict potential interactions by analyzing protein sequence, structure, and functional data.

Post-Translational Modification (PTM) Prediction: PTMs play a pivotal role in modulating protein function. Machine learning models can predict the sites and types of PTMs, thus enabling the investigation of protein regulation and signaling pathways. [45].

- **Drug Discovery**

Machine learning has surfaced as a formidable instrument in the drug discovery process, aiding researchers in streamlining the identification and development of new therapeutic compounds. Some of the key applications are:

Target Identification: Machine learning algorithms can analyze extensive biological data to pinpoint potential drug targets, such as proteins or genes implicated in disease pathways.
Compound Screening: Virtual screening methods, powered by machine learning, can predict the binding affinity of compounds to specific targets. This substantially reduces the time and cost associated with experimental screening.
ADMET Prediction: Machine learning models can forecast the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of drug candidates. This allows researchers to select compounds with a higher probability of success in clinical trials [46].

1.17 Machine Learning Applications

Machine learning algorithms can be effectively applied to TRP structures, offering significant improvements in the classification, clustering, and visualization of high-dimensional data. The following approaches can be utilized to enhance the understanding and analysis of TRP structures:

1.17.1 Supervised Classification

Supervised machine learning algorithms, such as the k-Nearest Neighbors (KNN) [47], can be employed to facilitate the automatic classification of TRP structures. In this approach, the algorithm is trained using labeled data, which consists of TRP structures with known classifications. The KNN algorithm classifies a new, uncharacterized TRP structure by analyzing the classifications of its k-nearest neighbors in the feature space. By leveraging the information from neighboring structures, the KNN algorithm can effectively predict the class membership of previously uncharacterized TRPs, providing an automated and scalable classification process.

1.17.2 Unsupervised Clustering

Unsupervised machine learning algorithms, particularly density-based methods such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [48], can be used to cluster TRP structures and refine the classification system. These algorithms identify clusters of TRPs with similar structural features without relying on predefined class labels. Density-based algorithms group TRPs based on the density of data points in the feature space, allowing for the identification of clusters with arbitrary shapes and the separation of noise from meaningful data. By revealing underlying patterns and relationships within the dataset, unsupervised clustering methods can provide a more comprehensive understanding of the structural landscape of TRPs and contribute to refining the classification system.

1.17.3 Dimensionality Reduction

Dimensionality reduction techniques, such as Principal Component Analysis (PCA) or Multidimensional Scaling (MDS) [49], can be employed to visualize high-dimensional TRP data in lower-dimensional spaces. These methods transform high-dimensional data into a lower-dimensional representation while preserving the essential structure and relationships between data points. By projecting the TRP data into a lower-dimensional space, researchers can more intuitively explore the relationships between TRP structures and identify potential outliers, novel structural motifs, or trends in the data. This approach facilitates a more accessible and interpretable understanding of the complex relationships between TRP structures.

1.18 The Importance of Classification and Annotating TRPs

The biological significance of protein tandem repeats was acknowledged previously, but their study has been fairly limited due to their inherent complexity and variability. However, with advancements in bioinformatics and computational biology, the potential of TRPs is increasingly recognized across various fields within life sciences, such as drug discovery, biotechnology, and diagnostics.

Classification and annotation of TRPs are essential for understanding their biological roles and harnessing their potential across these diverse applications. Classification groups these proteins based on characteristics such as structural properties or functional roles, facilitating a systematic

exploration of TRPs. This reliable classification system aids in identifying novel targets and applications in drug discovery, biomarker identification, and understanding disease mechanisms.

Comprehensive annotation of TRPs provides invaluable information with wide-ranging implications. Annotations offer insights into the structure, function, and interactions of TRPs, which are critical factors in drug design, biotechnology, molecular diagnostics, and personalized medicine. Understanding the precise three-dimensional structure of a TRP can guide the design of effective molecules or biotechnological tools, while functional knowledge of a TRP can help predict biological outcomes, inform disease diagnosis, and develop targeted therapies.

In general, reliable classification and comprehensive annotation of TRPs are pivotal steps in harnessing the potential of these proteins across life sciences. Such endeavors aid in identifying new targets and applications, guiding molecular design, and establishing a foundation for future therapeutic strategies, biotechnological advances, and diagnostic innovations. As bioinformatics tools continue to evolve, continued investment in these efforts is essential to fully explore the potential of TRPs and their impact on the scientific community.

2. Materials and Methods

This section provides a comprehensive overview of the resources and methodologies employed in this study, segmented into three primary areas: Summary of Materials, Data Collection and Analysis, and Strategy Formulation.

2.1 Summary of Materials

A concise summary of the resources employed in the current research is as follows:

- **Data Source**

Data for this research was obtained from RepeatsDB (version 3.0), encompassing a total of 15,296 entries. The release, dated September 15, 2022, was utilized for this study.

- **Programming Languages**

Python programming language (version 3.9) served as the computational tool for the execution of data manipulation, analysis, and visual representation operations.

- **Analytical Tools**

Several Python libraries were employed for the data analytics process. The Pandas library (version 2.0) was instrumental for data structuring and manipulation, whereas the Biopython library (version 1.81) facilitated the handling of biological data. Further, select modules from the scikit-learn library (version 1.2.2) were deployed for the application of machine learning algorithms and statistical techniques. Namely, the KNeighborsClassifier (KNN) was used for supervised classification, DBSCAN for unsupervised clustering, and MDS for multidimensional scaling.

- **Visualization Tools**

Data visualization was realized using the Matplotlib (version 3.7.1) and Plotly (version 5.14.1) Python libraries. These tools were pivotal in generating interactive and static visualizations to elucidate patterns and trends within the data.

- **Protein Comparison Tools**

TM-align and mTM-align were the software tools of choice for the structural comparison of protein tandem repeats. These tools enabled efficient and accurate alignment of protein structures, contributing to the comparative analysis.

2.2 Data Collection and Analysis

By performing a comprehensive statistical analysis on the data stored in RepeatsDB, we can uncover crucial information regarding the various properties of these protein repeats. This knowledge can be instrumental in making data-driven decisions and devising targeted experiments. By leveraging these insights, we can increase the efficiency and effectiveness of the future experiments, potentially saving time, resources, and reducing the likelihood of pursuing dead ends.

With having this in mind, initially, the RepeatsDB database is acquired from the designated website and subsequently parsed and processed into a Pandas dataframe using Python. Following this, an extensive statistical analysis is conducted based on various attributes associated with the repeat regions. These attributes encompass a diverse set of characteristics, including the length of the regions, the number of units within each region, the reviewed status, curator ID (if available), and other less pertinent information.

In this section, we conduct a thorough examination of various features and properties of the database entries, primarily focusing on the aspects of reviewed status, classifications, and the extent of annotation. It is important to emphasize, again, that the classification of repeats is executed across four distinct levels: Class, Topology, Fold, and Clan. To facilitate a more comprehensive analysis that balances between over-generalization and excessive attention to minute details, we opt to group the repeats at the Topology level. Consequently, their numerical representation in subsequent analyses will consist of only the first two numerical indices.

For instance, Alpha-Solenoids, a topology that falls under the Elongated class, and all the folds and clans that it encompasses, will be denoted as 3.3 in our analysis. This approach enables a more streamlined and focused examination of the data.

2.3 Strategy Formulation

This section presents a comprehensive description of the modified strategy employed for the analysis, clustering, and classification of protein tandem repeats. The approach is delineated in a meticulous, step-by-step fashion, ensuring a thorough understanding of the methodology.

2.3.1 Focusing on Key Areas

The diversity of protein tandem repeats requires a focused research approach as conducting an in-depth analysis of each group within this diversity is considerably time-consuming for this study, rendering it an impractical objective.

The primary aim, therefore, is to identify groups characterized by both a high overall number of repeats and a significant proportion of manually reviewed entries, due to their greater reliability compared to unreviewed ones. This focus on key areas, identified by their prevalence and reliability, aims to enhance the value of the study.

The data analysis carried out in the previous section sets the stage for this targeted approach. By identifying the distribution and characteristics of the repeats, it allows for the selection and prioritization of key areas for deeper investigation.

2.3.2 Resolving the Issue of redundancy

Upon examination of the RepeatsDB entries, it becomes evident that a substantial number of entries (Region_IDs) are associated with the same protein (Uniprot ID), primarily due to the redundancy of protein structures (PDB IDs) linked to specific proteins. Consequently, this redundancy within the dataset may adversely affect the validity of experimental outcomes.

To address this issue, for each unique protein (Uniprot ID), a single unique Region ID was randomly selected, and the remaining Region IDs related to that specific protein segment were excluded from the dataset. It is essential to note that, on occasion, two distinct repeat regions may reside on the same polypeptide chain, necessitating caution to avoid the omission of one of the

repeat regions. By implementing this strategy, potential biases introduced by structural redundancy can be mitigated, thereby enhancing the integrity of subsequent analyses.

2.3.3 Structural Comparison of Protein Repeats

After establishing the experimental population and eliminating noise and redundancies, the subsequent stage involves comparing the protein repeats to precisely discern the relationships between each entry and the entire dataset. Consequently, a distance matrix can be generated, serving as the foundation for the application of machine learning techniques in this investigation.

Given that this study is focused on the structure of proteins, the method for comparing repeats and calculating their respective distances relies on Protein Structure Alignment as the primary approach for assessing structural similarities.

Here we briefly discuss the logistics behind protein structure alignment and TM-align as the main used algorithm for this purpose.

TM-Align: TM-Align is a structure alignment algorithm that employs a heuristic iterative approach to align protein structures [1]. The algorithm consists of three main steps: (i) initial alignment using secondary structure matching, (ii) optimal superimposition of structures based on the TM-score, and (iii) dynamic programming for the final alignment. The TM-score is a metric that measures the global similarity between two protein structures, with values ranging from 0 (no similarity) to 1 (identical structures). TM-Align is known for its speed, accuracy, and ability to handle large-scale structure comparisons [50]

While TM-Align is primarily used for pairwise comparisons, mTM-Align, an extension of TM-Align, extends its capabilities to handle multiple structure alignments, making it more suitable for large-scale analyses involving multiple protein structures [51].

TM-score: One of the key advantages of both algorithms is their reliance on TM-score as a measure of structural similarity. The TM-score is less sensitive to local structural variations and is more robust in assessing global structural similarity than traditional measures like RMSD (root-mean-square deviation).

2.3.4 Comparison Levels

The analysis of structural similarities among protein repeats can be performed at various levels. In this study, three distinct methodologies were implemented, each possessing unique advantages and disadvantages, rendering them suitable for application in diverse scenarios. The following figure illustrates an example of 3 different levels of structural units that were used during the course of this study.

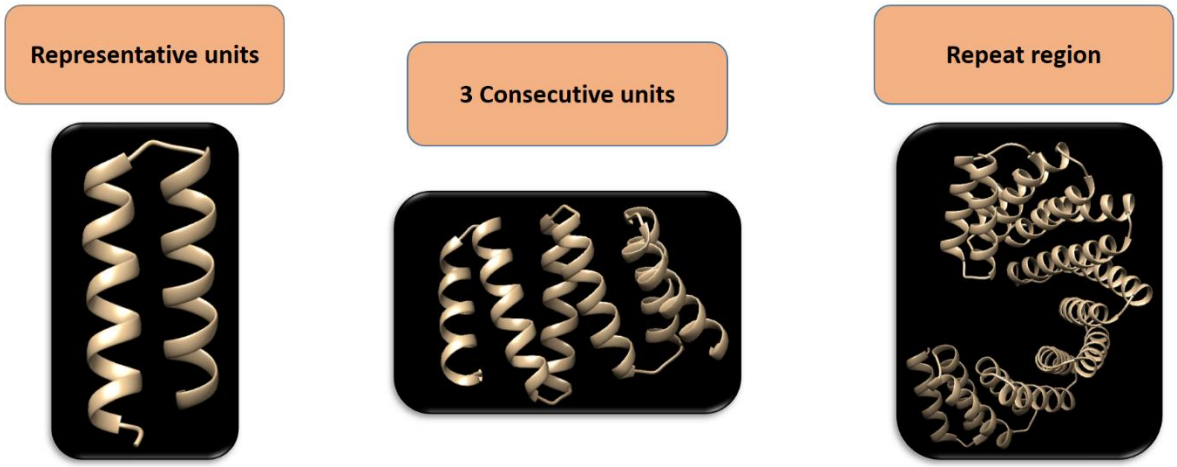
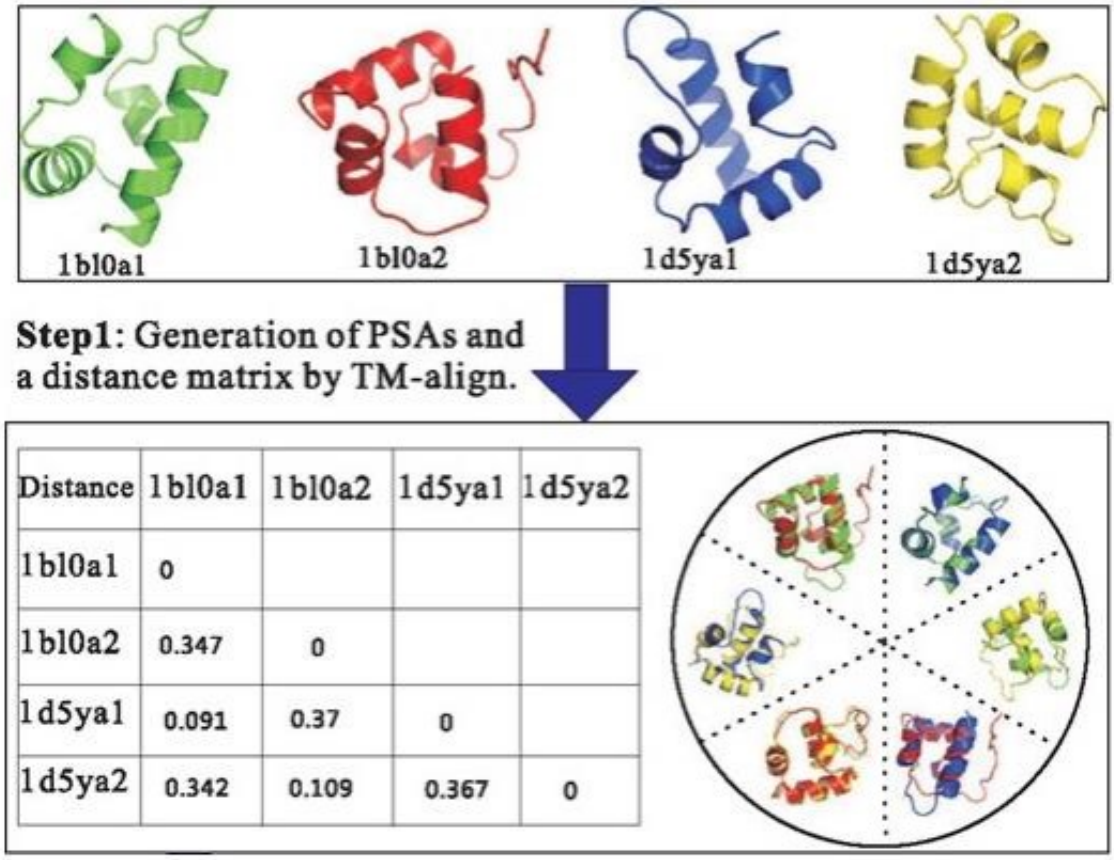


Fig. 2.1: An illustration of three distinct levels of structural comparison

- **Representative Unit**

Initially, for each repeat region, all the units within that region were extracted from the PDB files and aligned using mTM-align. The algorithm's output is a similarity matrix of all the units within a region. By analyzing this matrix, the unit with the highest average similarity to the other units was identified as the "Representative Unit." This unit is considered to best represent the overall structure of the units within that region.

Subsequently, the representative units from each region were aligned pairwise using TM-align to generate a comprehensive similarity matrix of representative units in an all-vs-all manner. The process of creating a distance matrix with TM-align is illustrated in the figure below.



Distance	1bl0a1	1bl0a2	1d5ya1	1d5ya2
1bl0a1	0			
1bl0a2	0.347	0		
1d5ya1	0.091	0.37	0	
1d5ya2	0.342	0.109	0.367	0

Fig. 2.2: A schematic representation of the distance matrix generation process using TM-align

- **3-Consecutive Units**

First, the representative units of regions are found using a method similar to the one mentioned earlier. Then, besides the representative unit, the two closest units to it are also picked and taken from the PDB structures. So, if the unit is in the middle, the units before and after it are included. If it's at the N-terminus or C-terminus, two units upstream or downstream are added, depending on the location.

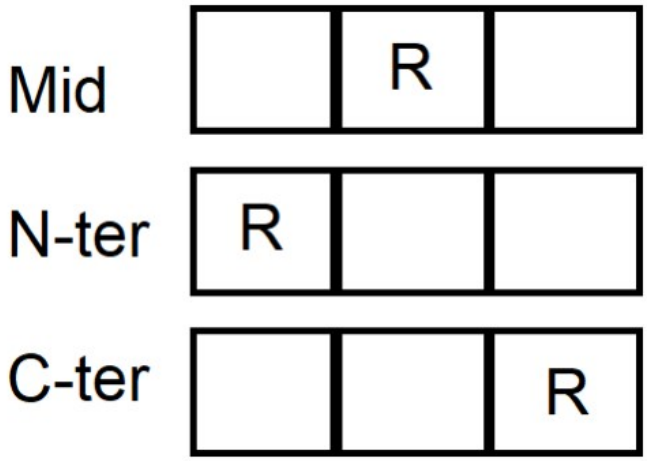


Fig. 2.3: Three combinations of 3-consecutive units in relation to the representative unit (R)

- **Whole Region**

At this level, for each repeat, the whole region is extracted from the associated PDB structures and is aligned with all the other repeat regions in an all-vs-all PSA (pairwise structural alignment) to generate the mother distance matrix. This approach is particularly suitable for the comparison of repeats that do not differ drastically in the number of their entailed units.

2.3.5 Visualization of High Dimensional Data by MDS

Prior to delving into the application of clustering and classification algorithms on the distance matrices generated in previous steps, it is essential to visualize the data to obtain a more intuitive comprehension of the information being processed. Heatmaps, a widely utilized data visualization technique, display the magnitude of a phenomenon as color in two dimensions and are often employed for visualizing distance matrices.

Despite the accuracy of heatmaps in representing raw data, they can prove difficult to interpret, particularly when dealing with matrices that span thousands of rows and columns (Fig X).

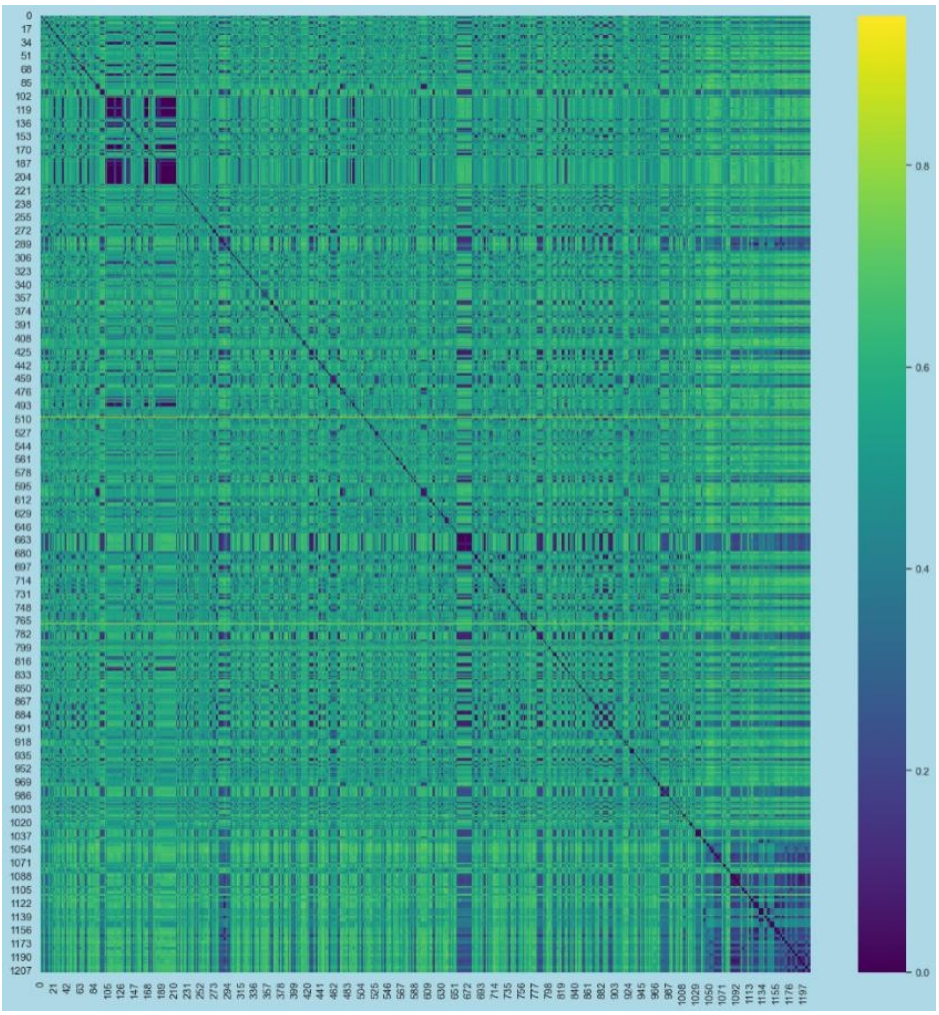


Fig. 2.4: A visual example of a large distance matrix

Consequently, alternative visualization approaches may be required. Dimensionality reduction techniques offer a promising solution in such situations. These algorithms can be beneficial by effectively reducing the complexity of the data, enabling easier identification of patterns and relationships within the dataset. By projecting high-dimensional data onto a lower-dimensional space, Dimensionality reduction algorithms enhance interpretability and facilitate a more comprehensive understanding of the underlying data structure.

Multidimensional Scaling: In this study, the main algorithm used for this purpose is called Multidimensional Scaling (MDS). It is a statistical technique that, as previously explained, is used for visualizing and analyzing complex, high-dimensional data by reducing it to a lower-dimensional representation. This method aims to preserve the original distances between data

points as closely as possible while transforming them into a more easily interpretable format, typically a two or three-dimensional space [52]. The figure below is a simple illustration of how this technique is applied.

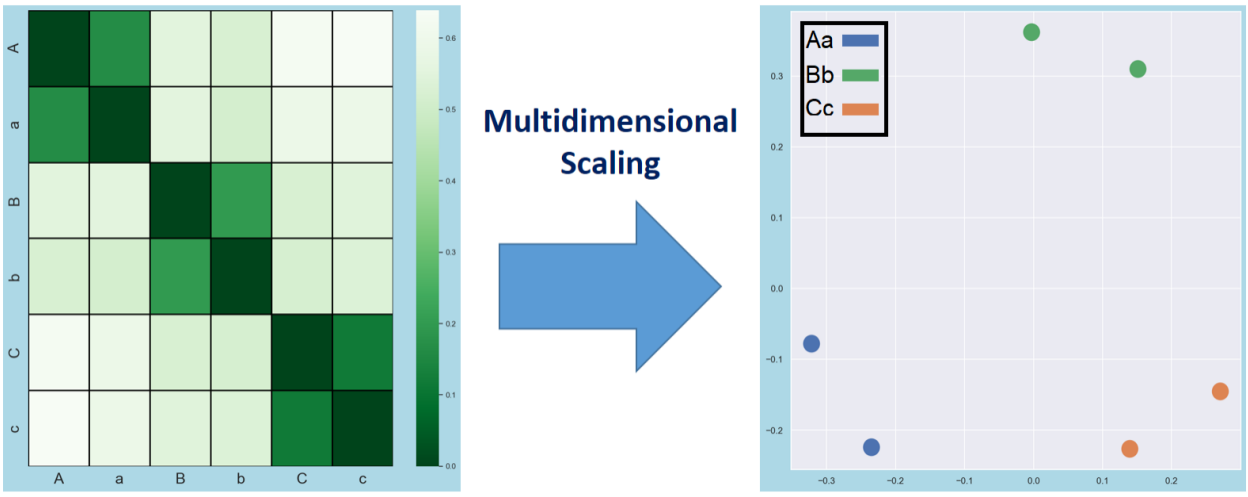


Fig. 2.5: How MDS enables a facilitated visual representation of a distance matrix

2.3.6 Clustering

Employing various unsupervised clustering algorithms to group the data can provide insights into which repeats exhibit sufficient structural similarity to form distinct clusters. By comparing the algorithmically defined clusters with the existing classification, we can assess the extent of their alignment and agreement. In cases of discrepancies, further investigation into the root causes can help determine whether the classifications can be rectified or enhanced.

Although there is a multitude of clustering methods and configurations available. However, In the context of clustering a large distance matrix without prior assumptions about the number of clusters and with the goal of identifying outliers, not all are well-suited to the objectives of this study or the specific data type under examination. Here, we will discuss 3 of the widely utilized clustering techniques and provide justification for selecting one method that is most appropriate for achieving the study's goals.

- **K-means Clustering**

K-means is a partitioning algorithm that requires the predetermined number of clusters (k) as input [53]. This requirement may not align with the aim of clustering without any assumptions about cluster quantity. Additionally, K-means is sensitive to the initial placement of cluster centroids and is generally not effective at identifying outliers or handling noise, limiting its suitability for the given task. Most importantly, it requires continuous, vector-based data as input and cannot be directly applied to a precomputed distance matrix.

- **Hierarchical Clustering**

Hierarchical clustering generates a tree-like structure (dendrogram) to represent the hierarchy of data clusters. The number of clusters can be determined by cutting the tree at a specific level. While this method does not necessitate prior knowledge of the number of clusters, it may be sensitive to noise and outliers, making it less suitable for the outlined purpose. Furthermore, its computational complexity may hinder its performance when applied to a very large dataset.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

DBSCAN is a density-based clustering algorithm that does not require a predefined number of clusters, making it well-suited for the task. It groups points based on their density, enabling the detection of clusters of varying shapes and sizes. Additionally, DBSCAN classifies less dense regions as noise, effectively separating outliers. However, it is essential to note that the algorithm's performance may be impacted by its sensitivity to the choice of parameters.

Overall, among the widely used clustering methods, DBSCAN emerges as the most suitable algorithm for clustering a large distance matrix of structural similarities between protein repeats without prejudice about the number of clusters and for identifying outliers.

In the figure below, it is shown how different clustering algorithms, including K-means and DBSCAN, handle the clustering task in different cases and conditions. It is obvious that DBSCAN proves to give out the best results in the given examples.

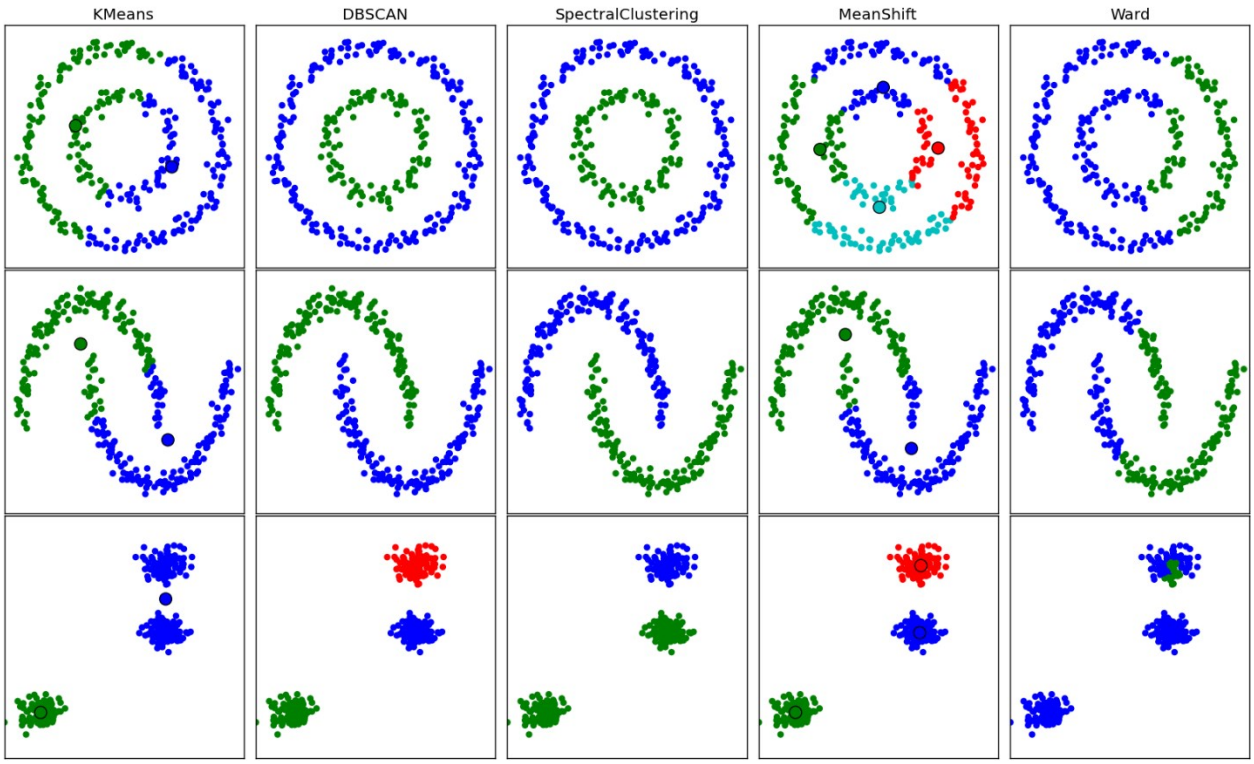


Fig. 2.6: A side-by-side comparison of clustering results obtained from five distinct algorithms

2.3.7 DBSCAN parameters

The DBSCAN algorithm offers a variety of adjustable parameters that can be fine-tuned to accommodate the specific data and analysis objectives, ensuring optimal and trustworthy results. The following are some of the most critical parameters relevant to our study:

- **Eps:** Epsilon (Eps) denotes the radius around a data point, defining the neighborhood within which other data points are considered neighbors. The choice of Eps directly influences the algorithm's ability to identify dense regions and separate clusters from noise. An appropriate Eps value is essential for accurate cluster formation and depends on the scale and distribution of the data. During the course of the analysis, a multitude of experiments were conducted utilizing an extensive range of Eps values. The objective was to examine and ascertain the extent of structural similarity exhibited by the members of a Clan in relation to one another.
- **Metric:** The metric parameter refers to the distance measure used to compute the similarity or dissimilarity between data points. Common distance metrics include Euclidean,

Manhattan, and cosine distances. The choice of an appropriate metric is crucial for capturing the underlying structure of the data and directly affects the quality of the clustering results. However, as we already have our distance matrices prepared, this parameter was set to "precomputed".

- **Min_samples:** It represents the minimum number of data points required to form a dense region or cluster. This parameter is used to differentiate between core points and noise in the dataset. A suitable min_samples value balances sensitivity to noise and cluster formation, ensuring that the algorithm can accurately identify meaningful groups within the data. Given that some Clans in the dataset comprise only a limited number of members, it is essential to ensure that they, too, have the potential to constitute individual clusters. Consequently, the min_samples parameter was set to a value of 2 in order to accommodate this constraint.

2.3.8 Classification by KNN and the K Parameter

Having established a refined and reliable classification through the previous step, we are now well-equipped to proceed with the classification process. In this phase, representative samples were designated for each group defined within the classification system, corresponding to the Clan level. Subsequently, the K-Nearest Neighbors (KNN) algorithm was employed to classify and label the extensive collection of Reviewed but Not-fully Classified entries. In the following discussion, we elucidate the functioning of KNN and the parameters that would be most suitable for our experiment.

K-Nearest Neighbors: is a non-parametric, instance-based supervised learning algorithm employed for classification tasks. It operates by computing the distances between a query instance and all labeled instances in the training set. The algorithm then selects the 'k' closest instances (neighbors) and assigns the majority class among these neighbors to the query instance.

The figure below pictures how KNN functions in a simple instance through 4 main steps.

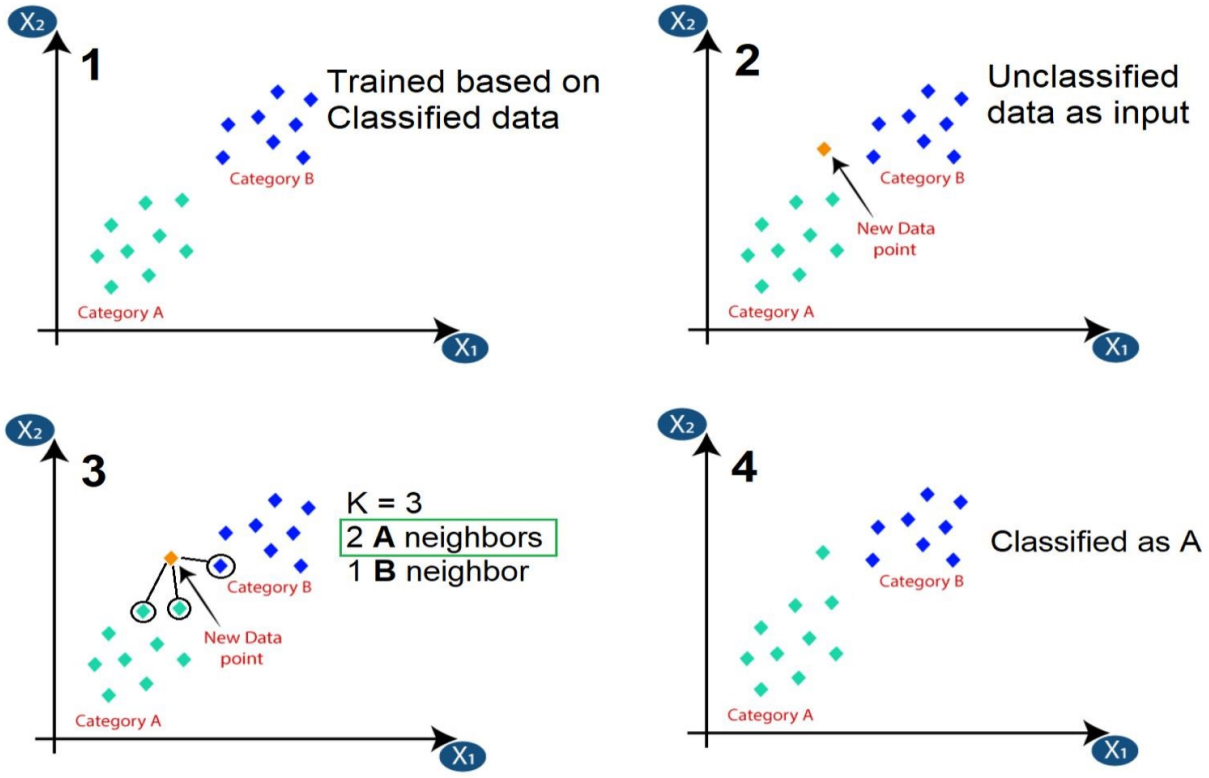


Fig. 2.7: A simple diagram demonstrating the K-nearest neighbors (KNN) algorithm in action

It is important to note that the choice of 'k' is a very crucial parameter that can significantly impact the classification results.

The number of neighbors, 'k' influences the algorithm's sensitivity to noise and its ability to generalize. A small value of 'k' (e.g., k=1) may result in a highly flexible model, making it susceptible to overfitting and noise. Conversely, a large value of 'k' may lead to an overly smooth decision boundary, potentially increasing classification errors due to underfitting. Selecting an appropriate 'k' value involves balancing the trade-off between overfitting and underfitting.

The following figure illustrates how different 'k' values can affect the classification result drastically.

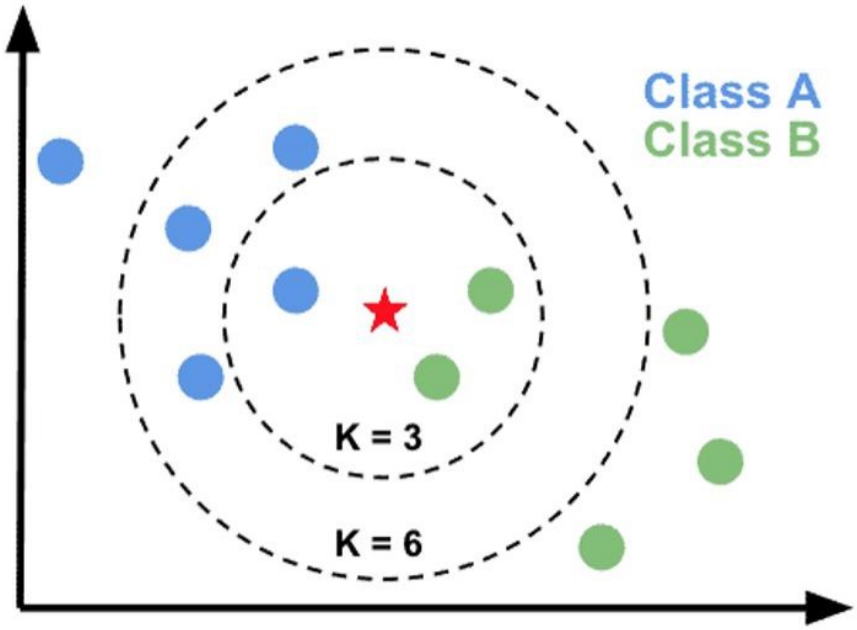


Fig. 2.8: An example of the impact of the K parameter on classification results in KNN

The Choice of 'K': In the present study, two representative samples were selected for each well-defined and refined group derived from the clustering analysis. A series of tests were conducted to determine the optimal 'k' parameter for the K-Nearest Neighbors (KNN) algorithm. As a result, a 'k' value of 3 emerged as an appropriate choice for this experiment.

As previously discussed, an excessively small 'k' value, such as 1, may compromise the model's reliability, rendering it susceptible to overfitting and noise. Furthermore, a 'k' value of 2 can introduce complexities when a sample is equidistant from two neighbors with distinct labels. In this context, a 'k' value of 3 represents the smallest viable option that accommodates the experimental conditions effectively.

The rationale behind this choice lies in the observation that, in the majority of cases, two of the three nearest neighbors will belong to the same group, as they represent the same group's units. This configuration enables the KNN model to operate seamlessly with minimal complexity while maintaining an adequate level of reliability.

2.3.9 Distance threshold

A notable limitation of the K-Nearest Neighbors (KNN) algorithm is its inherent tendency to assign a label to every input, irrespective of its dissimilarity to the labeled training data. Consequently, even instances exhibiting substantial divergence from the training data will be assigned labels based on their proximity to the nearest neighbors, potentially resulting in misclassification. Such misclassifications can undermine the reliability of the classification outcomes.

To mitigate this issue, a distance threshold was implemented as a preprocessing step before introducing inputs to the KNN algorithm. By doing so, only inputs exhibiting a minimum degree of similarity to the labeled data are subjected to KNN classification. Instances failing to meet this similarity criterion are designated as "unclassified outliers" and segregated into a separate dataset for subsequent analysis. The determination of the distance threshold was facilitated by the clustering analysis conducted using the DBSCAN algorithm in the previous stage of the study.

Incorporating a distance threshold as a preprocessing step enhances the reliability of the classification results by minimizing the misclassification of dissimilar instances and enabling the identification of unclassified outliers for further examination.

2.3.10 Scavenging the Threshold Outliers

In the concluding stage of the pipeline, an additional round of clustering analysis is conducted on the unclassified outliers obtained from the previous step. The aim is to determine whether it is possible to extract novel groups of protein tandem repeats not yet represented in the existing classification system. By utilizing the insights acquired from the earlier clustering analysis, including the identification of optimal clustering parameters and the degree of structural similarity between the members of a Clan, the same principles can be applied to the current stage.

The clusters generated in this step are subject to further analysis and investigation to discern the presence of common features. Such features typically encompass repeat length, the number of units, and annotations from third-party sources such as Pfam and InterPro. Through this approach, the existence of new groups of repeats can be ascertained, if present. Subsequently, these novel

groups can be incorporated into the classification system, rendering it more comprehensive and inclusive.

2.3.11 Assembling everything into a single pipeline

As the final phase of this section, the previously outlined steps are integrated into a unified pipeline. This integration fosters a more coherent and consistent approach to the study, enhancing its reproducibility. The entire pipeline, encapsulating all the process stages, is graphically represented as a flowchart in the subsequent figure.

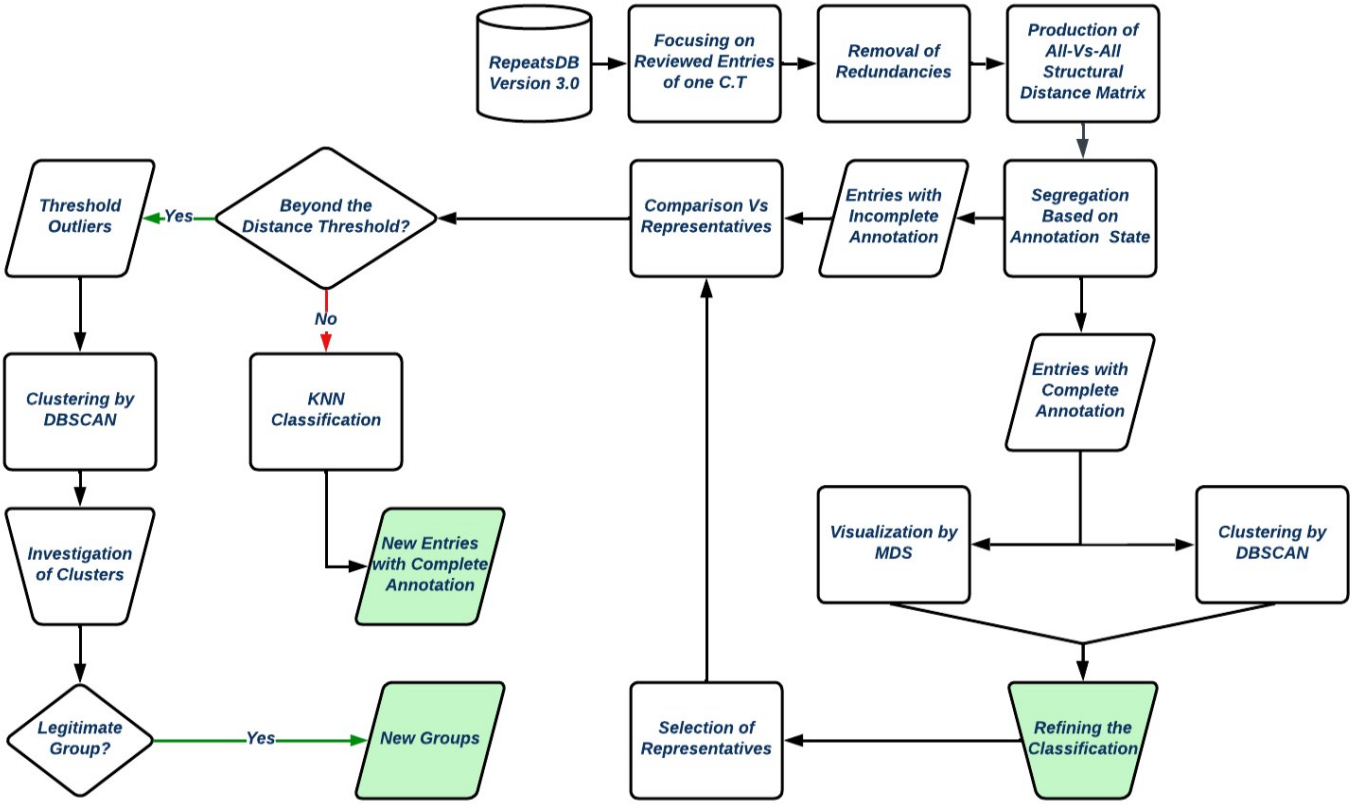


Fig. 2.9: A flowchart depicting the complete pipeline for TRP classification and automatic annotation

3. Results

The forthcoming sections presents the results derived from the statistical analysis and the application of the developed pipeline.

3.1 Statistical Analysis

This section is dedicated to the results of the statistical analysis conducted on the dataset, focusing primarily on three distinct aspects. These aspects encompass the reviewed status and classification diversity of protein repeats, the quantities of different classifications, and the completeness versus incompleteness of annotations of entries. The obtained results play a pivotal role in directing the pipeline towards areas of the dataset of the highest significance, making the process more targeted and efficient.

- **Reviewed status and classification diversity**

The following two pie charts offer valuable insights into the composition of entries within the dataset. The first chart illustrates the proportion of reviewed and unreviewed entries, and the second chart delves into the distribution of only reviewed entries across various classification categories.

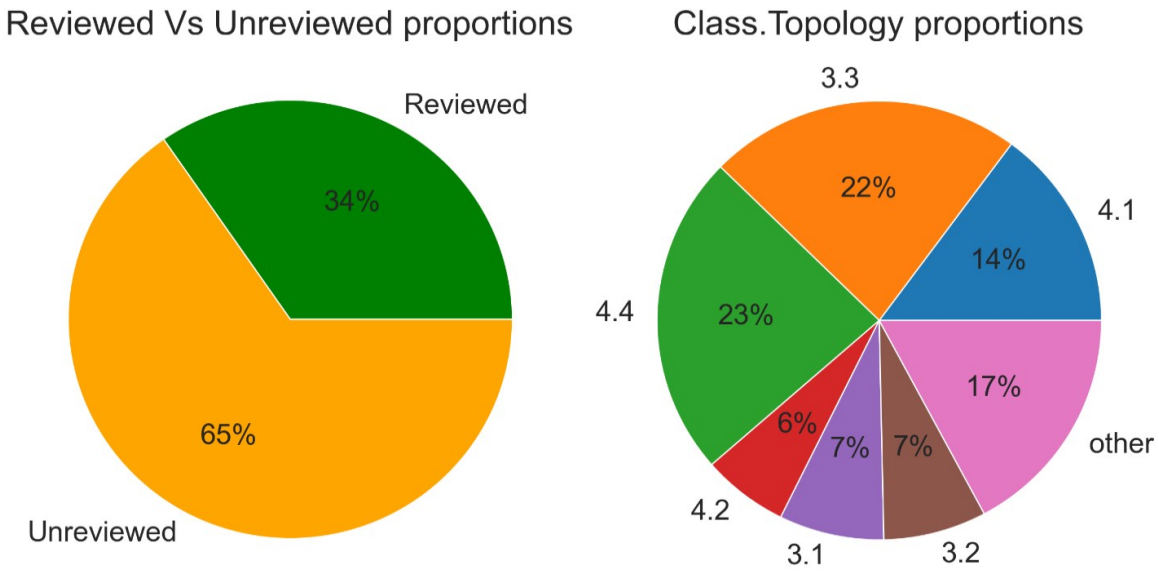


Fig. 3.1: Two pie charts: left, reviewed vs. unreviewed entries; right, distinct class.topology proportions

The pie chart on the left provides an overview of the proportion of reviewed and unreviewed entries within the dataset. The chart shows that 34% of the entries have been reviewed, indicating that these entries have been manually curated and have undergone a quality control process for accuracy and reliability. In contrast, the remaining 66% of the entries are unreviewed, suggesting that these entries are curated automatically and might require further validation or assessment before being utilized for research or analysis purposes.

The second pie chart on the right presents the distribution of reviewed entries across different classification categories. The chart highlights that 23% of the entries belong to the 4.4 class, making it the most common classification in the dataset. The 3.3 class follows closely, representing 22% of the entries. The 4.1 class accounts for 14% of the entries, while classes 4.2, 3.1, and 3.2 each constitute around 6-7% of the dataset. The remaining 17% of the entries are distributed among other classification categories

- **Quantities of different classifications**

The bar graph provides a visual representation of the number of repeat regions in each classification category (Class. Topology) for tandem repeat proteins. On the X-axis, we have 20 distinct classifications, while the Y-axis displays the corresponding count of repeat regions within each class. Also, the number of Reviewed regions in each classification is highlighted on the bars for a more informative presentation.

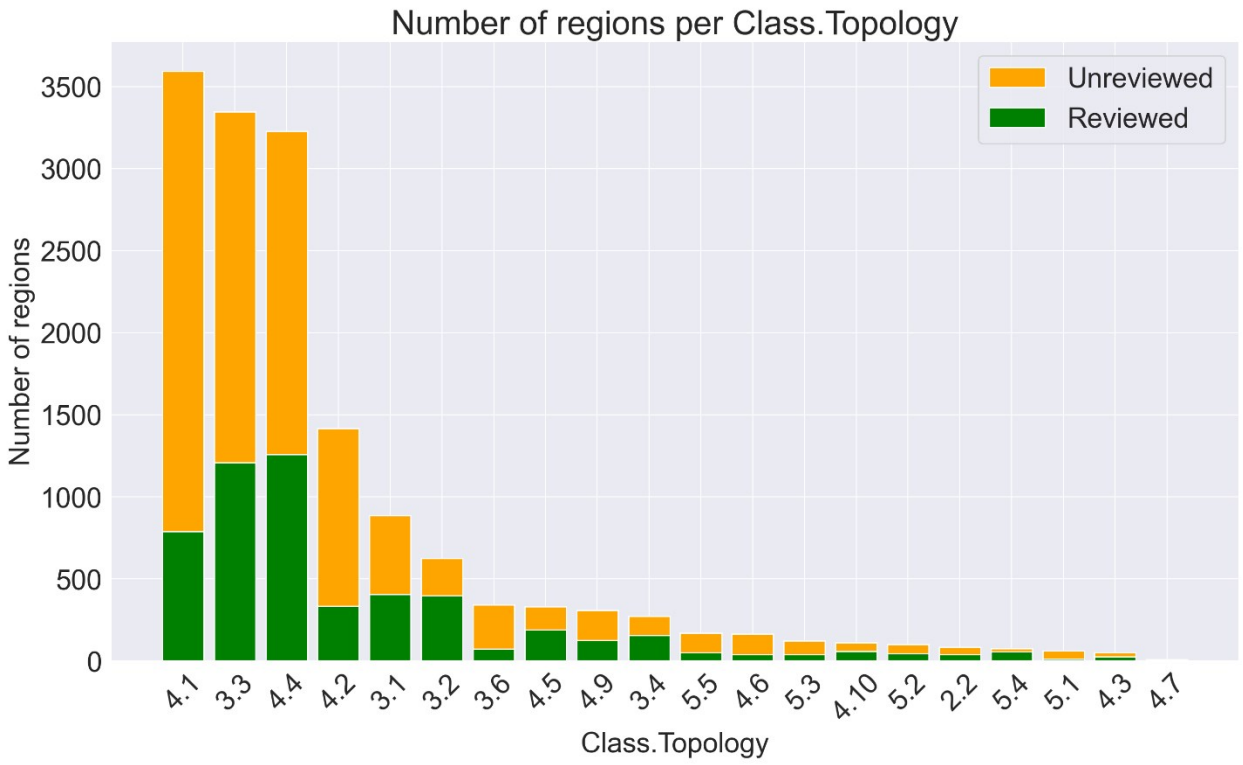


Fig. 3.2: A bar graph displaying the entry count for each class.topology category

The graph reveals that the most populated classes are 4.1, 3.3, and 4.4, each containing more than 3,000 entries. These classes represent the most common types of tandem repeat proteins found within the dataset, highlighting their potential significance in the research domain.

In contrast, the remaining 17 classes exhibit a considerable variation in the number of repeat regions, ranging from approximately 1,500 down to just a few entries. This wide distribution of repeat regions across the different classes demonstrates the diverse nature of tandem repeat proteins and the need for further investigation into these less-represented categories.

- **Complete vs. Incomplete Annotations**

The bar graph represents the classification status of repeats for three categories based on their Reviewed status. The X-axis displays the three categories, while the Y-axis indicates the number of repeats. Each category features double bars, with one bar representing Fully-classified repeats and the other bar representing Not Fully-classified repeats, which lack a clear annotation on at least one of the four levels of annotation in the classification system of RepeatsDB.

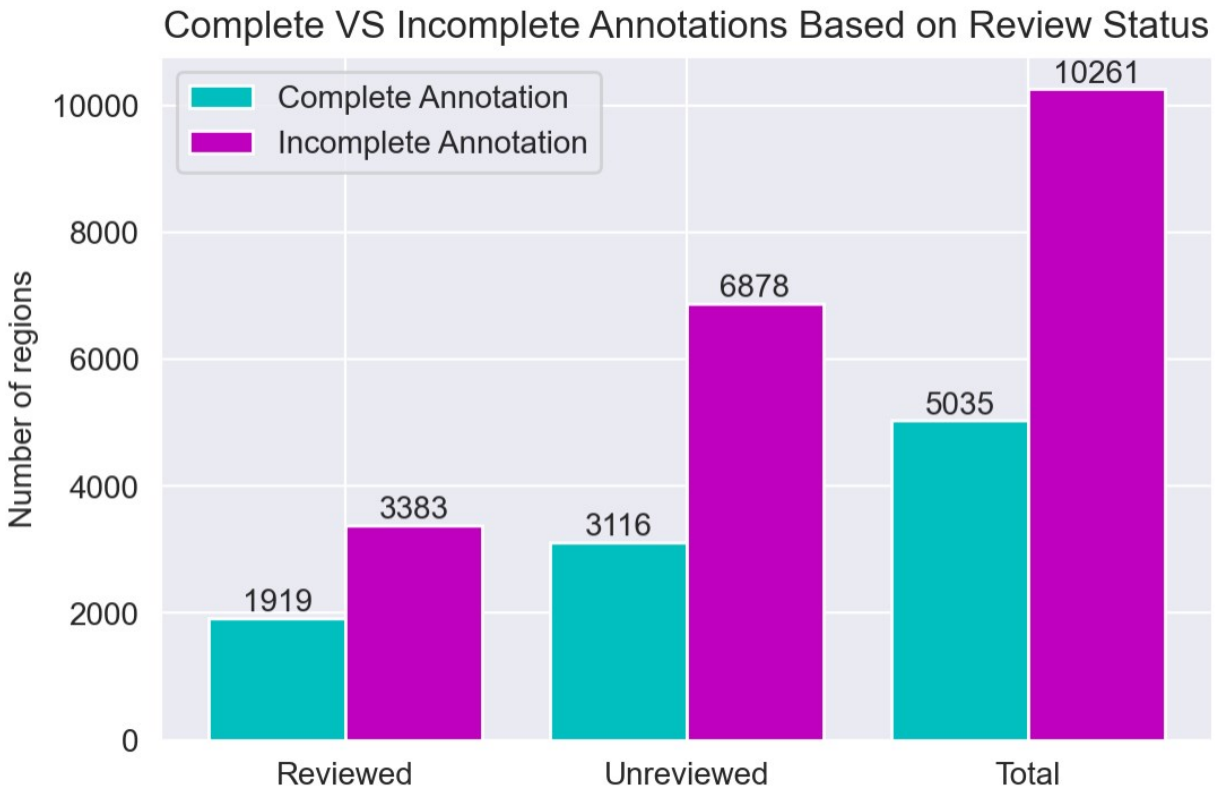


Fig. 3.3: A bar graph comparing the number of entries with complete and incomplete annotation, based on review status

In general, the graph demonstrates that there is a higher number of Not Fully-classified repeats, almost double, compared to Fully-classified ones across both the Reviewed and Unreviewed categories. This suggests that a significant portion of repeats are still not fully classified and have not been validated by curators.

- **The implications of the results**

The statistical analyses have yielded some key insights. In terms of classification quantities, groups 4.1, 3.3, and 4.4 contain the majority of repeats. However, it is worth noting that groups 3.3 and 4.4 stand out with a significantly larger number of reviewed entries compared to group 4.1. As reviewed entries generally provide more reliable data than unreviewed ones, focusing on groups 3.3 and 4.4 would be a more prudent approach.

Additionally, the analysis of the state of entry annotations underscores a pressing need for improvement. The marked disparity between complete and incomplete annotations points towards the potential for enhancing the overall quality and reliability of the dataset.

3.2 Application of the Pipeline on Alpha-Solenoids

The entire pipeline was executed separately for both Alpha-Solenoids (3.3) and Beta-Propellers (4.4). Therefore, the results of each type are examined independently to facilitate focused analysis and mitigate potential confusion. We commence with a detailed examination of the Alpha-Solenoids results, encompassing all relevant information. Subsequently, we present a more concise overview of the Beta-Propellers findings to avert repetition and redundancy.

Through an extensive series of experiments and trials, the decision was ultimately made to employ a distance matrix based on the structural comparison of "3-consecutive units." This choice was primarily influenced by a set of anomalies encountered with the "Representative unit" approach, which will be discussed in further detail at the end of this section.

3.2.1 MDS Results

Initially, to obtain an intuitive understanding of the data, the Multidimensional Scaling (MDS) algorithm was utilized to reduce the data dimensions to two. This approach enabled the visualization of the data via a 2D scatter plot. It is crucial to acknowledge that the ensuing visualizations represent estimations generated by the algorithm to reduce the data's complexity to the desired dimensions. However, these estimations are not guaranteed to be entirely accurate. For instance, in a 2D representation of a distance matrix, some entries may appear distant from one another, but the addition of a third dimension could reveal their proximity. The converse situation may also hold true. The subsequent scatter plots were generated using Plotly Express to create interactive plots, thereby facilitating the analysis process.

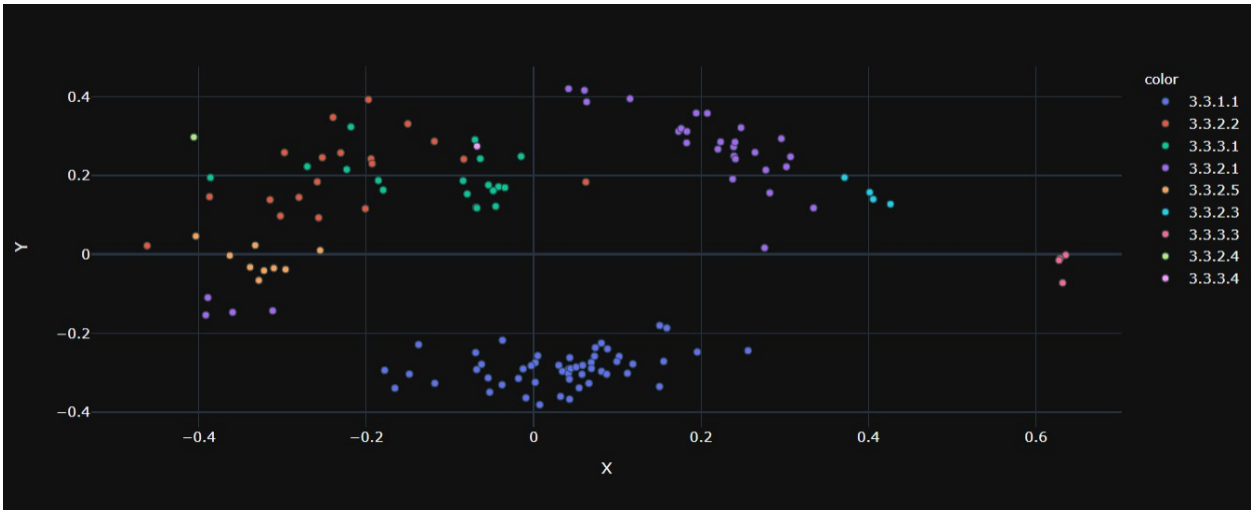


Fig. 3.4: A 2D representation, showcasing the relative structural distances of fully-annotated Alpha-Solenoids by MDS implementation

As observed, the majority of Clans form relatively distinct clusters with reasonably clear boundaries. However, members of the 3.3.2.2 Clan appear to be scattered in a disordered manner across the plot, disrupting the boundaries of other clusters, primarily 3.3.3.1 and 3.3.2.5. This suggests that the 3.3.2.2 Clan may require correction and refinement for a more precise definition.

Another observation of importance is the marked scarcity of data pertaining to Clans 3.3.2.4 and 3.3.3.4, with each having only a single instance in the dataset. This data insufficiency for these Clans impedes a reliable clustering analysis, as a minimum of two samples is required to form a cluster core.

Upon gaining a preliminary comprehension of the data, it is recommended to first remove the members of Clans 3.3.2.2, 3.3.2.4, and 3.3.3.4 from the dataset before advancing to the subsequent clustering step. This measure will diminish data complexity and facilitate the clustering analysis. Following this, we will revisit the aforementioned Clans and attempt to establish a more accurate definition. The figure below illustrates the data after these adjustments, yielding a notable enhancement in data readability.

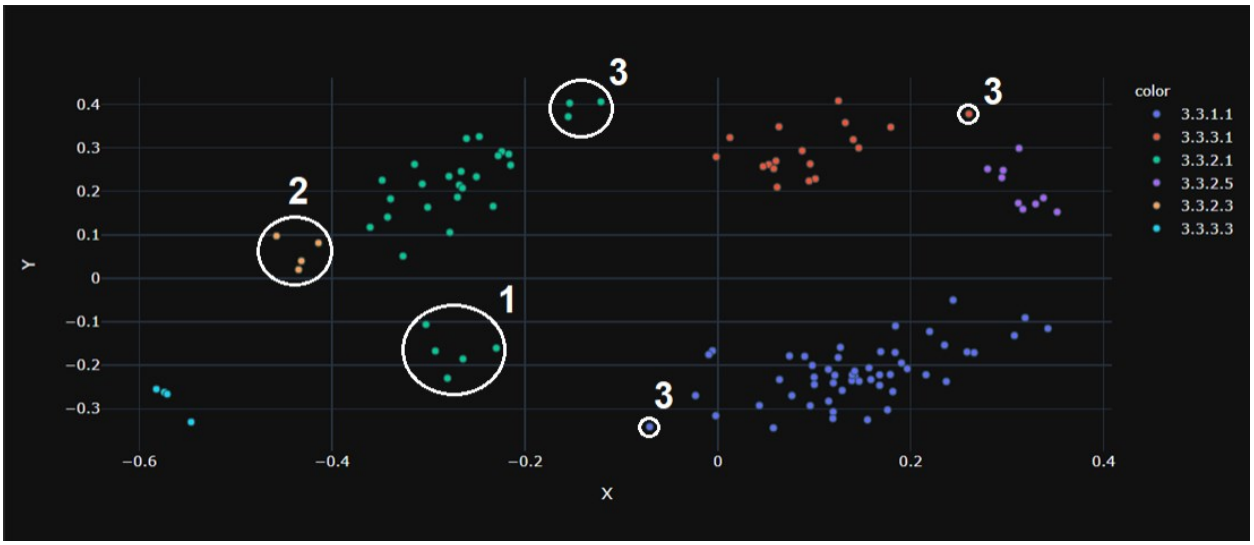


Fig. 3.5: Peculiarities are marked and numbered for further analysis

Several important features are highlighted and numbered in the figure, as described below:

1. As previously noted, these five members belonging to the 3.3.2.1 Clan appear to be situated far from the main body of the Clan. This observation suggests the possibility of two distinct types of repeats erroneously labeled as a single Clan.
2. Members of the 3.3.2.3 Clan are positioned in close proximity to the main body of the 3.3.2.1 Clan. This prompts the question of whether they are genuinely two different Clans or represent the same type of repeats labeled as two separate Clans.
3. A few members belonging to different Clans appear to be notably distanced from the core of their respective Clans. These instances warrant further investigation to determine the underlying cause of their outlying positions.

3.2.2 DBSCAN Results

Subsequently, the data was subjected to the DBSCAN algorithm for clustering. The analysis encompassed a broad range of Eps values to determine the optimal Eps that closely aligns with the existing Clan definitions in the predefined classification system, where each Clan forms an individual cluster. Two critical considerations during this step include avoiding excessive data fractionation and preventing a large number of samples from being classified as outliers.

It was determined that the most effective way to present clustering results was through bar plots, with the cluster index displayed on the X-axis and the quantity of repeats within each cluster represented on the Y-axis. The color composition of each bar correlates with the Clan composition of the corresponding cluster.

A series of clustering plots, each corresponding to a distinct Eps, were meticulously examined to observe the dissociation of different Clans at various distance thresholds. These insights shed light on the relatedness or separation between Clans. Ultimately, it was found that a distance threshold (Eps) within the range of 25% - 30% (75% - 70% similarity) among cluster members closely approximates the existing Clan definitions in the predefined classification system.

The subsequent bar plot represents the outcome of clustering the data at a distance threshold of 28% (72% sim).

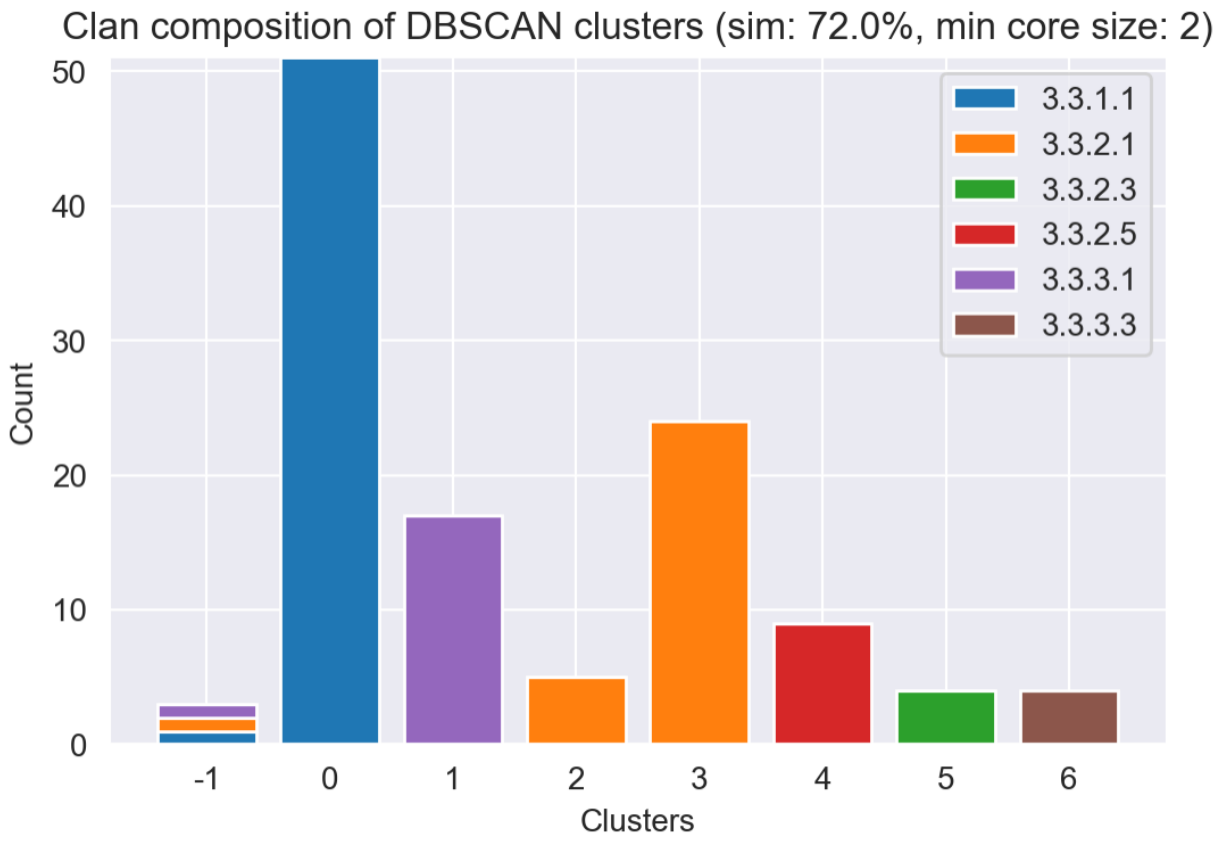


Fig. 3.6: a bar graph displaying the clan composition of the clustering result

It is evident that all clans, except 3.3.2.1, are suitably positioned within an individual cluster. Regarding this particular Clan, 5 members are positioned in a cluster distinct from the main cluster of the Clan. Aligning perfectly with our observation from MDS scatter plots, as these exact members of 3.3.2.1 are positioned so distant from the core of the Clan.

Furthermore, three samples from distinct clans are relegated to the outlier bin, indexed as -1, necessitating a more thorough examination of these repeats to ascertain the underlying cause of this outlying behavior.

3.2.3 Identification of Outliers

By looking through the outlier bin (-1), the following list of region IDs and their associated Uniprot IDs was retrieved:

Region ID	**Uniprot ID**
1sw6A_243_502	P09959
3grlA_18_631	P41541
4hotA_48_467	Q13325

Upon conducting a meticulous examination of each entry individually, which involved comparing their 3D structures with other repeat structures from their corresponding Clans and reviewing their annotations in other databases, primarily Pfam and InterPro, we were able to validate the likelihood of these repeats genuinely belonging to their respective Clans. The results of this analysis are as follows:

1sw6A_243_502 (P09959): An Ankyrin repeat containing a number of insertions, which compromises its TM-score relative to other Ankyrin repeats.

3grlA_18_631 (P41541): An ARM-like repeat with a number of insertions, resulting in a compromised TM-score compared to other Armadillo repeats.

4hotA_48_467 (Q13325): A genuine TRP exhibiting a substantial insertion in the trimmed structure, consequently compromising its TM-score in relation to other Tetratricopeptide repeats.

By and large, the presence of insertions in the aforementioned outlier repeats has led to a reduction in their TM-scores, rendering them distinct from their respective repeat groups.

3.2.4 Fractionation of Clans

This Clan is supposed to contain only Tetratrichopeptide (TPR) repeats. TPR repeats are structural motifs found in a wide range of proteins, playing a crucial role in mediating protein-protein interactions. Typically, TPR repeats on average consist of 34 amino acid residues, characterized by a helix-turn-helix structure, forming a pair of antiparallel alpha-helices connected by a short loop. TPR-containing proteins usually exhibit multiple tandem repeats, ranging from 3 to 16 repeats, which assemble into a superhelical structure to generate a binding groove for target proteins [54].

- **First Fractionation of 3.3.2.1**

During the clustering analysis, it was observed that at relatively very high distance thresholds, starting from an Eps of 0.4 (60% similarity), a cluster of 5 repeats dissociates from the main body of the Clan (Cluster 2 in Fig. 3.7).

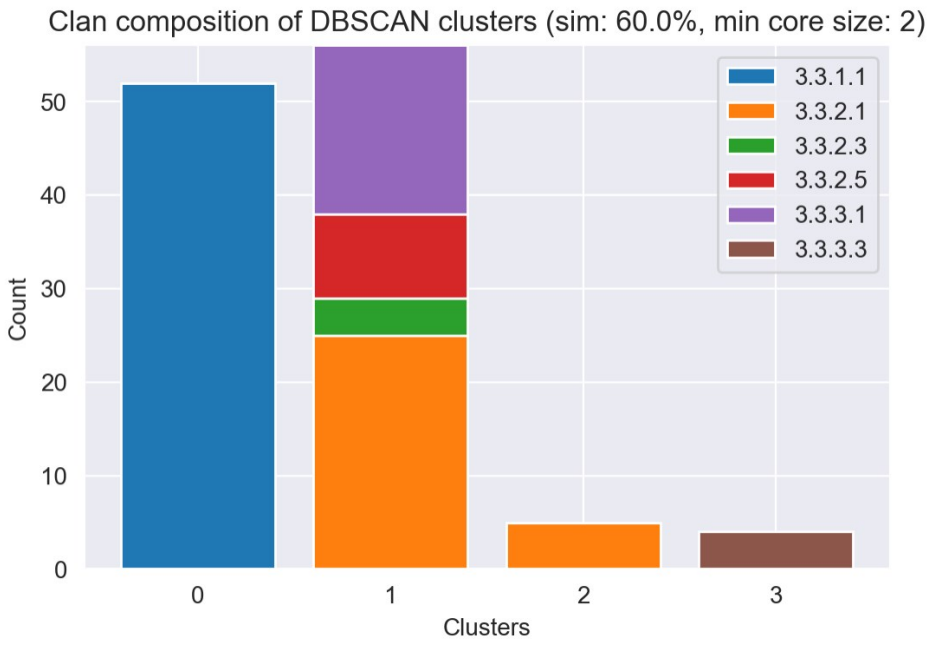


Fig. 3.7: Clan composition of clusters at the first fractionation of 3.3.2.1 clan

Looking through the members that form this cluster, the following repeats were obtained:

Region ID	**Uniprot ID**
4gtrA_113_368	Q04631
1ltxA_49_442	Q08602
2f0yA_114_366	P49354
3draA_48_288	Q9Y765
3q78A_56_334	Q55S71

Further investigation of the members of this independent cluster revealed the truth that even though they resemble the shape of Tetratricopeptide repeats, none of them is annotated as genuine TPRs. They are not even included in the "Tetratricopeptide-like helical domain superfamily" (IPR011990) in InterPro database. Instead, they are annotated as "Protein prenyltransferase, alpha subunit" (PPTA) repeats both in Interpro (IPR002088) and in Pfam (PF01239).

The studies regarding these particular repeats show that they are structural motifs found in the alpha subunit of protein prenyltransferases, a group of enzymes that play a vital role in protein modification through the prenylation process [55].

Next, the structures of PPTA repeats were compared to canonical TPRs to elucidate structural differences. Upon pairwise structural alignment of representative units, it was observed that both the connecting loops between helices within a unit and the loops connecting adjacent units exhibit greater length in PPTA repeats relative to TPRs.

At the fold level, TPRs demonstrate a more compact arrangement, with each unit exhibiting a higher rotational angle relative to its preceding unit, resulting in a more pronounced curvature in comparison to PPTAs. The results of the structural alignment is shown in Fig. 3.8.

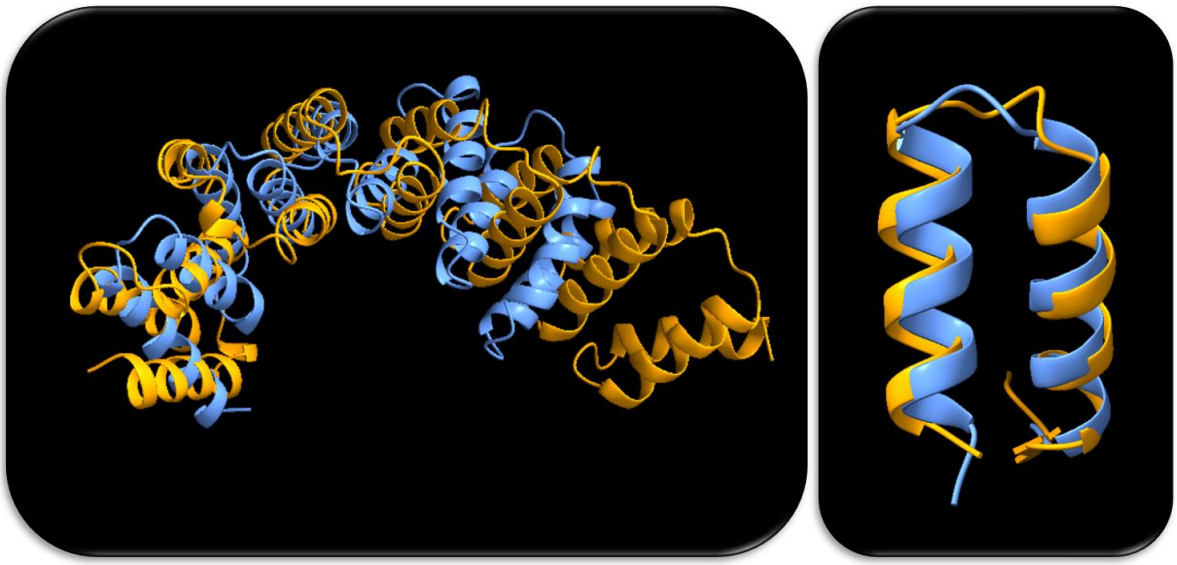


Fig. 3.8: Structural alignment of PPTA (orange) vs. TPR (blue) repeats; left: region level, right: unit level

- **Second fractionation of 3.3.2.1**

If the distance threshold is brought down to around 25%, we will witness another round of fractionation of 3.3.2.1's. At this level, another 2 members of this Clan fractionate into a third individual cluster (Cluster 6 in Fig. 3.9).

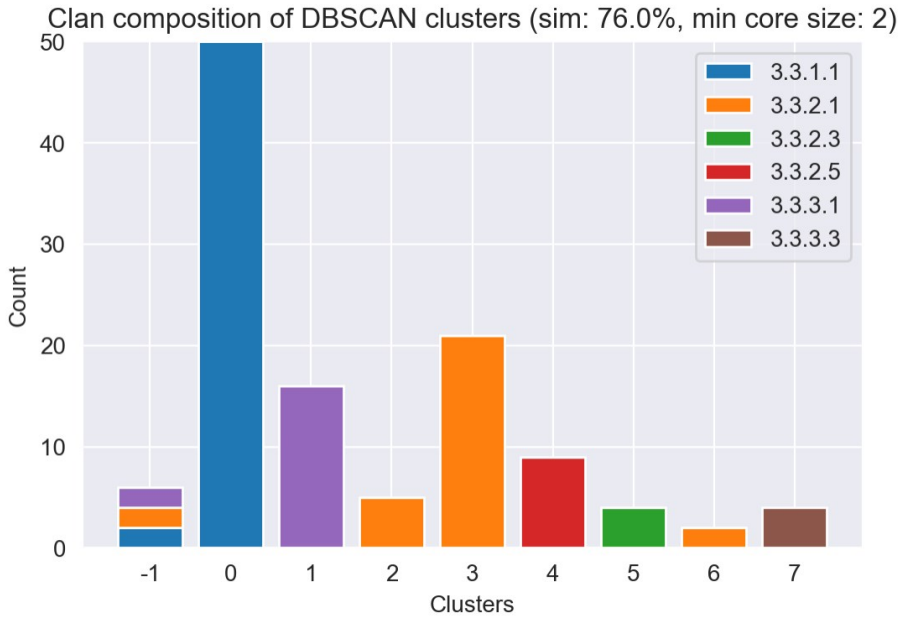


Fig. 3.9: Clan composition of clusters at the second fractionation of 3.3.2.1's

Looking through the members of this newly formed cluster, the following repeats were obtained:

Region ID	**Uniprot ID**
2ifuD_6_244	Q5BJK3
1qqeA_29_268	P32602

The same approach for investigation of these repeats was adopted and the results are as follows:

They both encompass 6 units of an average length of 39 residues and even though they are not annotated as "Tetratricopeptide repeat" either in InterPro (IPR019734) or Pfam (PF13181), they are at least included in the "Tetratricopeptide-like helical domain superfamily" (IPR011990). On

the other hand, Pfam has annotated them as “Soluble NSF attachment protein (SNAP)” repeats (PF14938).

Pertained studies suggest that Soluble NSF Attachment Protein repeats, also known as SNAP repeats, are a class of protein repeats found in the SNAP family of proteins. These repeats are involved in intracellular membrane fusion events and are essential for vesicle trafficking. The SNAP proteins play a critical role in the SNARE (Soluble NSF Attachment Protein Receptor) complex, which facilitates the fusion of vesicles with target membranes in eukaryotic cells.

These repeats are characterized by their α-helical structure with each repeat unit being composed of two α-helices connected by a short loop. These tandem repeat units are organized in a right-handed superhelical manner, creating a larger curved structure that enables protein-protein interactions and contributes to the stability of the SNAP proteins in their functional complexes [56].

Aside from publications and 3rd party annotations, our structural examination of these repeats revealed interesting points which are as follows:

At the unit level, the constituent helices of SNAP repeats are 1 turn longer than those of TPRs . On the fold level, not much difference in curvature and positioning of units compared to each other can be seen, except that, as mentioned earlier, the helices are longer in SNAPs compared to TRPs. (Fig. 3.10)

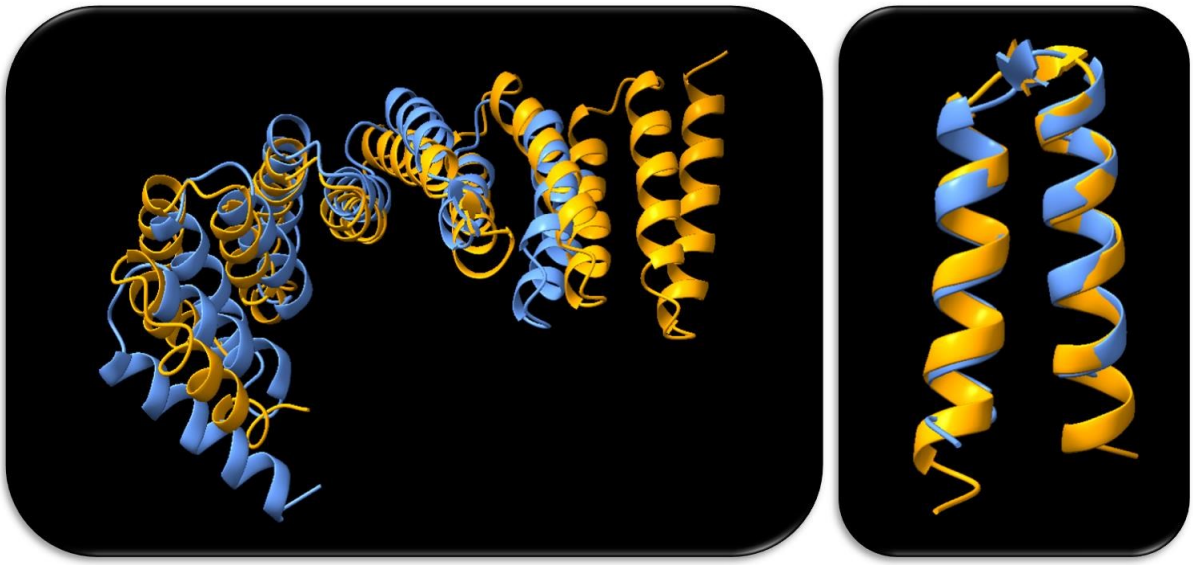


Fig. 3.10: Structural alignment of SNAP (orange) vs. TPR (blue) repeats
left: region level, right: unit level

These observations explain why the unit lengths of these repeats are above the average length of TPR units, as their constituent helices are almost one complete turn longer. Also, the proximity of the structures both at the unit and fold level explains why they are positioned fairly close to the main body of the 3.3.2.1 Clan in the MDS scatter plots.

The figure below highlights the discoveries obtained through the conducted analyses up to this point:

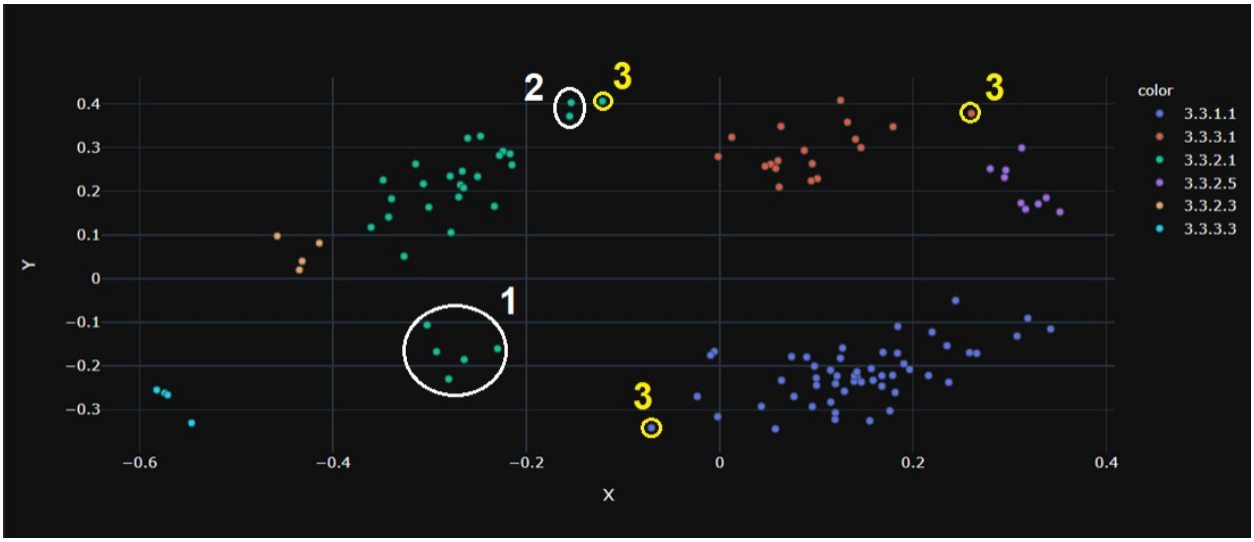


Fig. 3.11: Pecularities that were previously identified, devided and maked by two different tones

1. This group of five data points represents the PPTA repeats identified by the initial fractionation of 3.3.2.1 Clan at an Eps of 0.4 (60% similarity) through DBSCAN clustering.
2. This small group of two data points represents the SNAP repeats that dissociated from the main body of 3.3.2.1 Clan at a low Eps of 0.24 (76% similarity) through the clustering analysis.
3. These three data points, highlighted in yellow, represent the three repeats that fall into the outlier bin at an Eps of 0.28 (72% similarity), which yields an optimal result in the clustering analysis, where the majority of Clans form individual clusters.

3.2.5 Pre-Classification

In order to improve the classification process, it is necessary to further refine the dataset by addressing outliers, defining a new Clan corresponding to the PPTA repeat clusters discovered during the previous step, and addressing Clans 3.3.2.2, 3.3.2.4, and 3.3.3.4, which were eliminated from the dataset at the onset of the analysis.

Initially, the three outlying repeats of 1sw6A_243_502, 3grlA_18_631, and 4hotA_48_467 must be removed to ensure these exceptional repeats do not influence the analytical steps when selecting representative repeats for each group.

Subsequently, the primary 3.3.2.1 Clan and its dissociated clusters of PPTA and SNAP repeats will be labeled as 3.3.2.1_PPTA and 3.3.2.1_SNAP, respectively. This significantly improves the labeling of existing groups within the dataset, which is crucial for supervised classification procedures.

To address the eliminated clans, a smaller-scale classification analysis must be performed. Two representative repeats are selected for each defined Clan (including the modified ones), and the members of the eliminated Clans are passed through a distance filter. The repeats that pass the filter are then subjected to KNN classification to determine if they belong to any of the existing Clans in the dataset.

First, representative repeats are chosen by calculating the average distance of each repeat within a Clan to other members of that clan and selecting the two repeats with the lowest average distance, indicating their higher similarity to the rest of the Clan members.

Next, it is determined whether any members of the eliminated Clans have a distance lower than 28% from any of the representatives. This distance threshold is based on observations from the clustering steps, as clustering the data with a similar Eps reveals the current state of Clans within the dataset. The filter's purpose is to eliminate repeats less likely to be related to existing Clans. Subsequently, the repeats that pass the filter are subjected to KNN classification, with the training data comprising representative repeats and their associated Clan labels.

Upon completion of the designated procedures, the ensuing results were as follows:

Among 19 repeats associated with Clan 3.3.2.2 and one repeat each for Clans 3.3.2.4 and 3.3.3.4 (comprising a total of 21 repeats), 18 did not meet the filtering criteria, whereas the three, represented in the following table, that surpassed the threshold were classified as 3.3.3.1 by the algorithm.

Region ID	**Designated Clan**	**Predicted Clan**
2h4mA_60_847	3.3.2.2	3.3.3.1
3nd2A_50_839	3.3.2.2	3.3.3.1
3w3xA_21_1077	3.3.3.4	3.3.3.1

To assess the validity of these predictions, a subsequent investigation was conducted. By examining the constituents of Clan 3.3.3.1 in third-party databases, it was discerned that the majority of members are annotated as Armadillo repeats linked to Importin proteins. This observation is consistent with the annotation of these three repeats, as they are implicated in both the "Armadillo-like helical" superfamily (IPR011989) and the "Importin beta" family (IPR040122). The final determination regarding the potential reclassification of these repeats will ultimately reside with the lead curator of RepeatsDB. However, for enhanced data representation, the aforementioned repeats were henceforth designated as 3.3.3.1_ADDED within the scope of this study.

The other 18 repeats were clustered using an Eps of 0.3, a more relaxed distance threshold compared to 0.28, to observe their clustering patterns. As expected, and shown in the figure below, many entries were found in the outlier bin (Cluster -1). This confirms that the initial organization and labeling of these repeat data were not optimal.

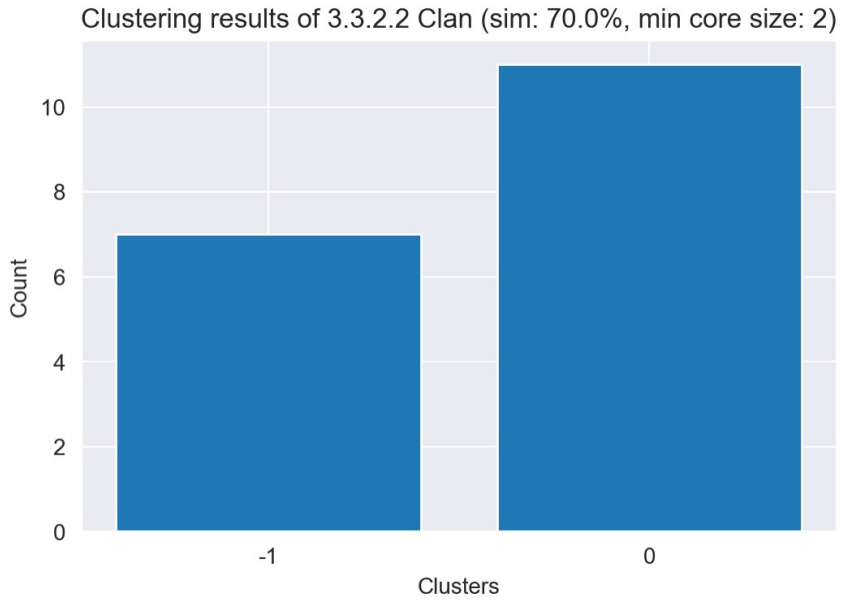


Fig. 3.12: Clustering result shows a main cluster (0) and everything else as outliers (-1)

By examining the only formed cluster (Cluster 0) and investigating the repeats within to identify a common feature, it was found that most, but not all, were annotated as "Adaptin N terminal region (Adaptin_N)" (PF01602) in Pfam. Using a conservative approach, only the repeats with this specific annotation were extracted and labeled as 3.3.2.2_New. The improvements can be seen in the MDS scatter plot (Fig. 3.13).

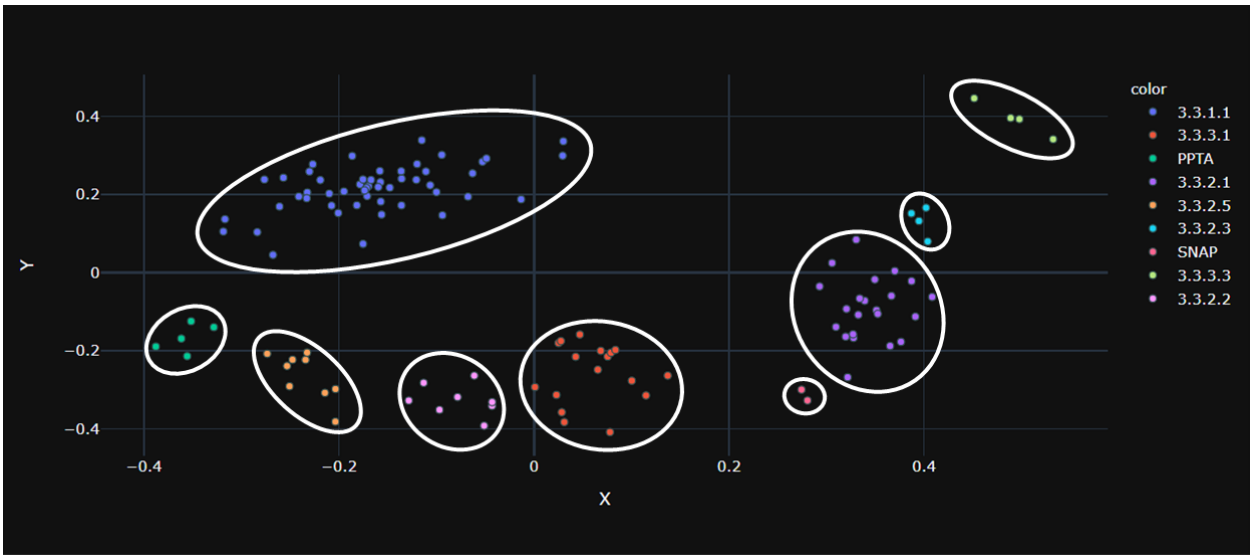


Fig. 3.13: MDS scatterplot of the refined classification with each clan having a distinct boundary

In this updated dataset, each Clan has clearer boundaries, and the distinctions between different Clans are more easily discernible than before.

3.2.6 Clan-based Classification

With a refined dataset and well-defined classification, we are now prepared to proceed with the classification procedure for the not-fully classified repeats. First, we will select two representative repeats per each Clan from the most updated and refined dataset. Subsequently, we will implement a maximum distance filter of 25%, serving as a conservative approach to exclude samples that are irrelevant or only marginally relevant to the training dataset. It should be noted that a more lenient filter would result in an increased number of classified data; however, this would compromise the confidence and accuracy of the predictions.

Following the establishment of the distance filter, we trained the k-nearest neighbors (KNN) algorithm on the representative repeats and their corresponding labels. All the reviewed repeats that were not fully classified and the repeats that were previously eliminated from the dataset during the refinement and cleaning procedures were then introduced to the algorithm for labeling.

The classification results for a total of 231 repeats are presented below:

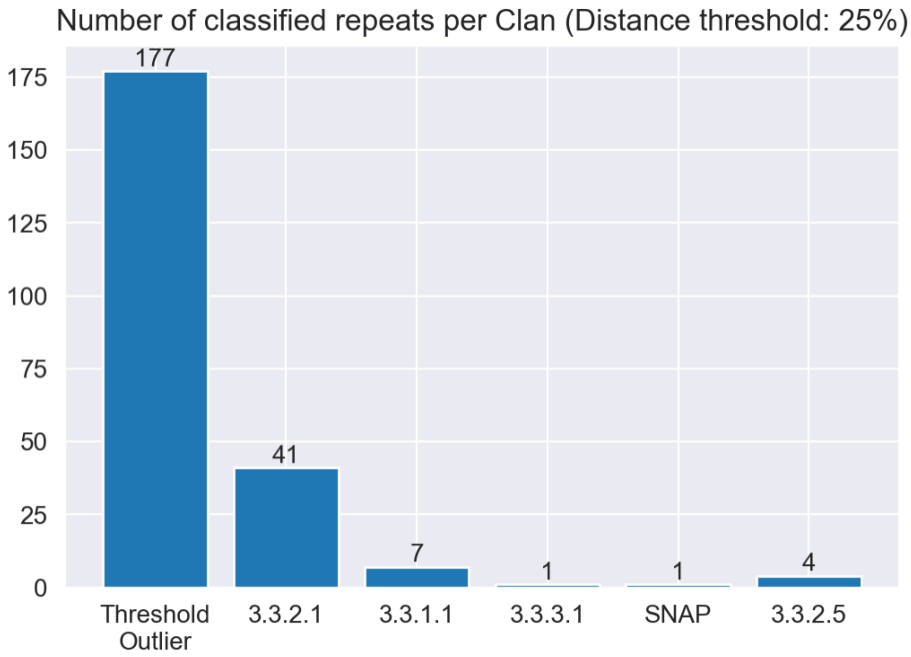


Fig. 3.14: KNN classification results of classifying the unclassified repeats

Upon examination of the results, only 23% (54 out of 231) of the not fully-classified entries were accurately categorized into five distinct clans, with clan 3.3.2.1 exhibiting the highest frequency of predictions. Although a comprehensive evaluation by RepeatsDB curators is pending, a heuristic investigation of the newly labeled data suggests a high degree of accuracy. For instance, the specialized and recently defined clan of SNAP repeats, which comprises a mere two members in the dataset, acquired an additional entry bearing a similar label. A thorough examination of the annotations corroborates the accuracy of this prediction.

Nevertheless, by moderately relaxing the filtering constraint by 5% (equating to a maximum distance of 30%), a notable 10% increase in the number of classified entries is observed, resulting

in an overall classification rate of 33% (77 out of 231), which is commensurate with one-third of the total data.

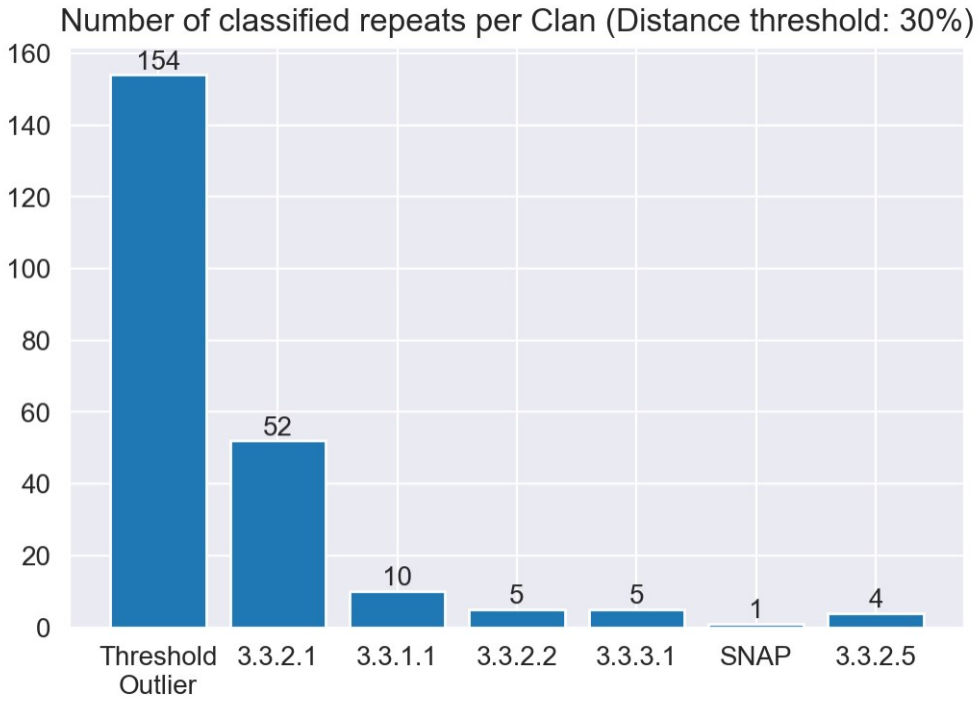


Fig. 3.15: KNN classification results of classifying the unclassified repeats (loosen threshold)

3.2.7 Family-based Classification

During the experiments conducted with DBSCAN, a notable observation emerged. In a relatively wide Eps range of 0.38 to 0.33 (corresponding to 62% - 67% similarity), the resulting clusters remain consistent and, interestingly, align with the Family annotations of the repeats within.

The dataset primarily consists of three major families: Ankyrin, Tetratrichopeptide, and Armadillo repeats. The subsequent figure illustrates the correspondence between the generated clusters and these families. Additionally, it reveals that when the repeats do not belong to these families, they segregate into distinct individual clusters.

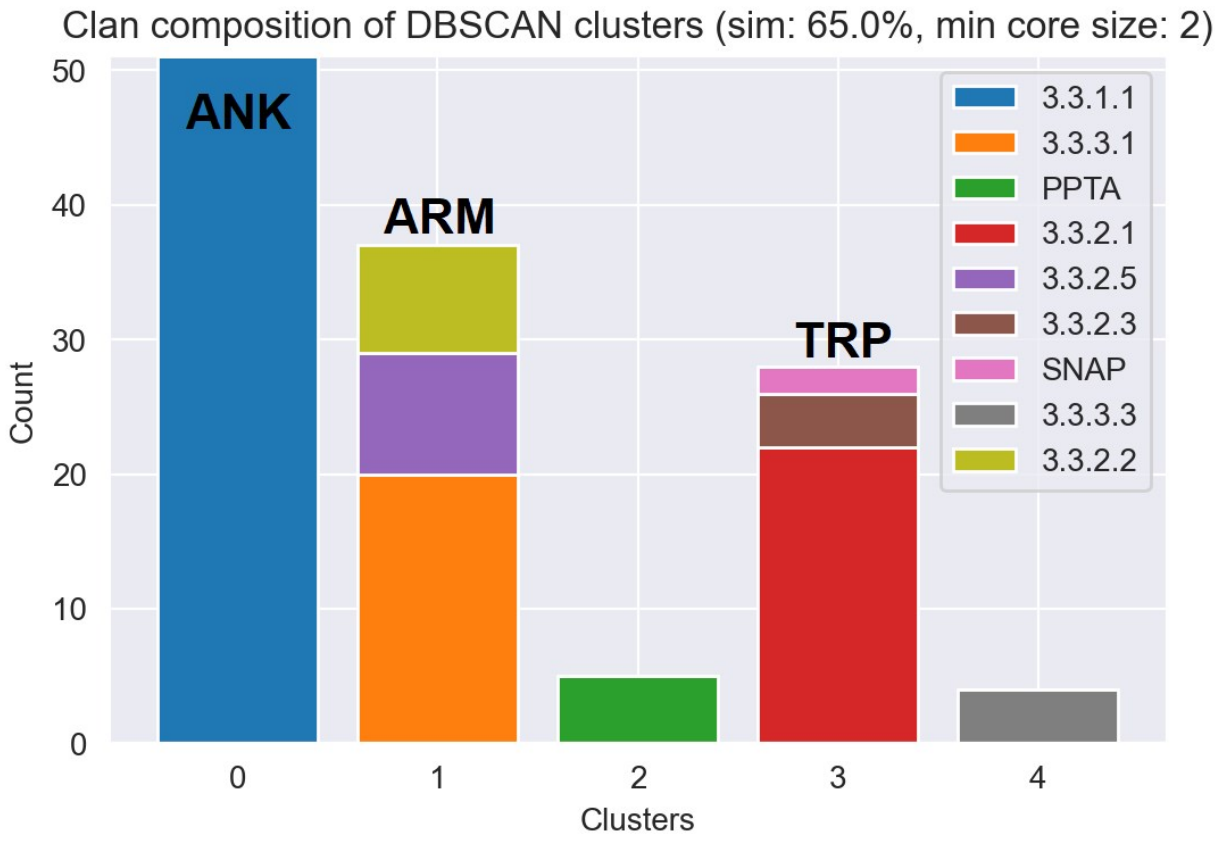


Fig. 3.16: Clustering at lower similarities can yield clusters that represent repeat families

With this newly gained knowledge, an additional experiment can be performed to automatically assign Family annotations to entries with missing annotations. The methodology for this task is not significantly divergent from the clan-based classification, with the exception that if the predicted clan labels pertain to any of the previously mentioned families, the entry will be assigned the corresponding family label instead.

However, a crucial distinction is that the distance filter can be loosened to 0.35% (65% similarity), based on the observed clustering behavior. This adjustment will substantially augment the quantity of entries that can be subsequently subjected to KNN classification.

The results of this experiment are presented below:

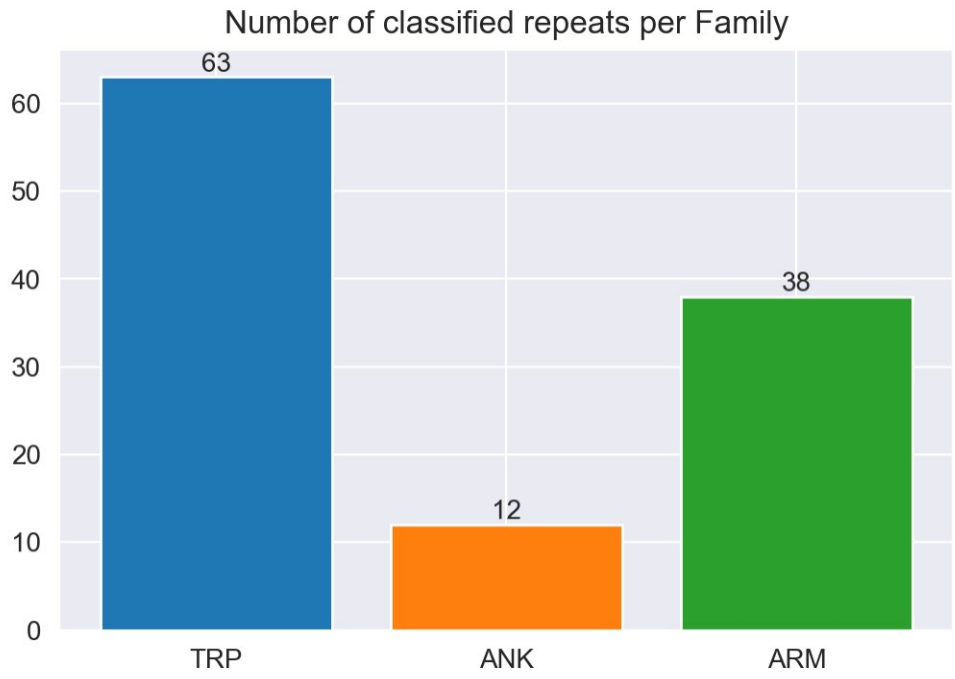


Fig. 3.17: KNN classification based on family labels

As demonstrated, a total of 113 entries were successfully auto-annotated through this approach, accounting for nearly half of the initial 231 entries. Another notable aspect evident in the plot is the considerably larger proportion of repeats from the Armadillo and Tetratrichopeptide families in comparison to the Ankyrin family. This discrepancy may be attributed to the ease of distinguishing and annotating Ankyrin repeats relative to repeats from the other families.

3.2.8 Scavenging the Outliers

As was shown in the last experiment, we were unable to give any sort of annotation either on the clan level or family level to an approximately 40% of the, suggesting that these entries are unlikely to be even marginally relevant to the existing groups within the classification system. Nonetheless, this does not imply that they cannot be classified or annotated, as the classification system itself may be inadequate.

To ensure a comprehensive attempt to classify the repeats, these outliers can be examined by clustering them using a conservative Eps of 0.25 and a minimum core size of 3 to avoid excessive data fragmentation. The outcome of this experiment is illustrated in the following figure:

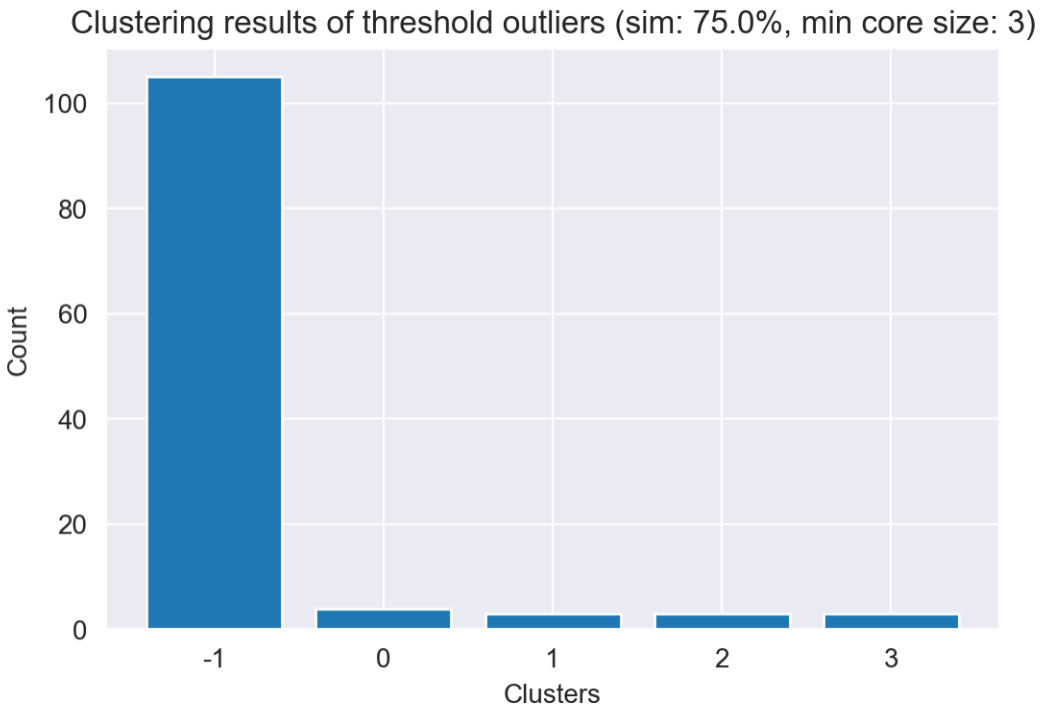


Fig. 3.18: Clustering of threshold outlier from classification step

It is obvious from the results, that in fact, a very large proportion of the data has such a divergent nature, that cannot be even clustered with a normal set of parameters. Aside from that, 4 other clusters were formed and a comprehensive investigation through every single of them was carried out with the hope of identifying new groups of repeats to be added to the classification system. The results of this endeavor are listed below:

(Whether the listed groups are eligible to be recognized as new Clans or not is out of the scoop of this research and relies in the hand of the curators of RepeatsDB)

Cluster 0:

4 members

Region ID	Uniprot ID
1a38A_2_201	P63103
4zq0D_6_213	E1F4V5
5nasB_2_208	P63104
6fbyC_1_209	P31947

Annotations: 14-3-3 protein (PF00244), 14-3-3 domain (IPR023410)

Description: Structurally, 14-3-3 proteins are typically around 200-300 amino acids in length and form homo- or heterodimers. Each monomer consists of nine antiparallel alpha-helices, with the overall structure resembling a cupped, elongated hemisphere. The dimeric structure creates a central, negatively charged channel that can accommodate target peptide sequences containing a phosphorylated serine or threonine residue. The binding of target proteins to the 14-3-3 domain is primarily facilitated through specific phosphorylation-dependent motifs. This interaction can lead to changes in the target protein's conformation, stability, activity, or subcellular localization, thus modulating its function in the cell [57].

Structure in RepeatsDB:

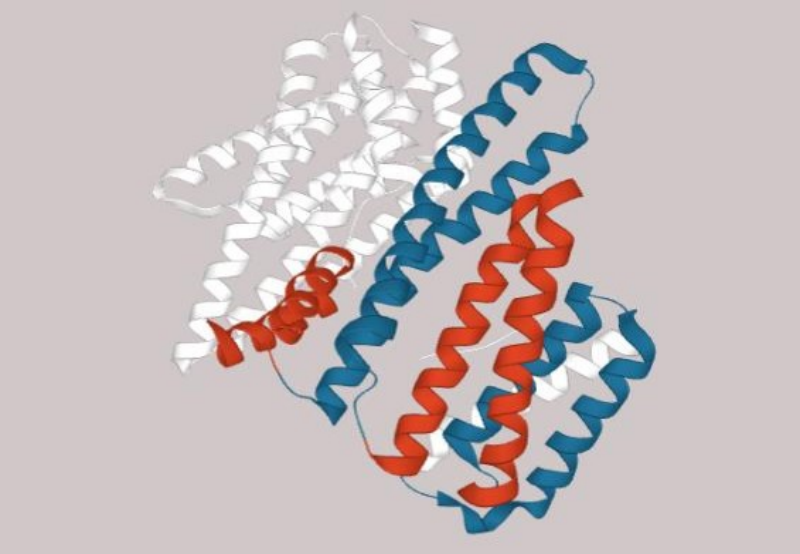


Fig. 3.19: Structure of 4zq0D in RepeatsDB

Cluster 1:

3 members

Region ID	**Uniprot ID**
2of3A_649_867	G5EEM5
6mzeE_310_541	A0A493R6X7
6mzgE_310_543	A0A493R6X8

Annotations: CLASP N terminal (PF12348)

Description: The CLASP N-terminal domain plays a crucial role in modulating microtubule dynamics by specifically recognizing and binding to the growing plus ends of microtubules, thereby affecting their growth, shrinkage, or stabilization. This domain is characterized by the presence of multiple repeats [58].

Structure in RepeatsDB:

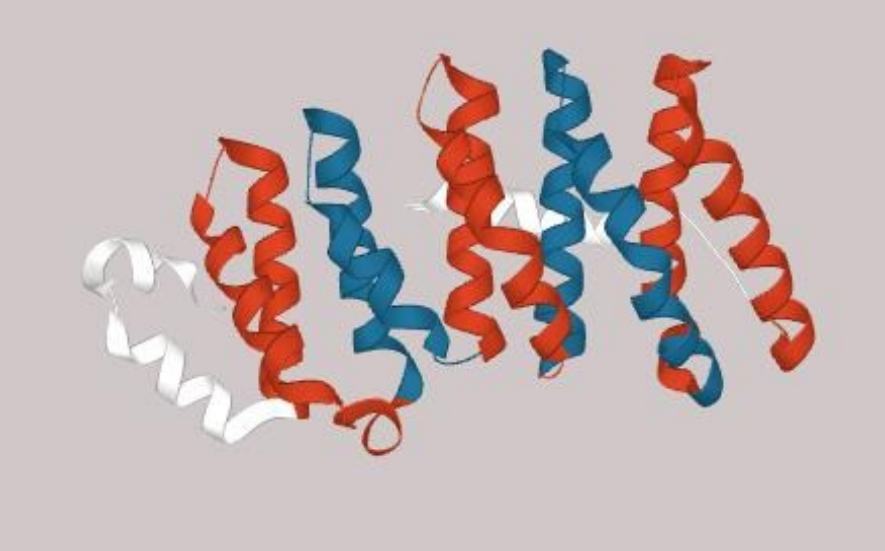


Fig. 3.20: Structure of 2of3A in RepeatsDB

Cluster 2:

3 members

Region ID	**Uniprot ID**
2wviA_57_220	O60566
3eslA_31_194	A6ZUJ9
4a1gB_3_148	O43683

Annotations: Mad3/BUB1 homology region 1 (PF08311)

Description: The Mad3/Bub1 homology region 1 (also known as the Bub1 and BubR1 N-terminal (BBN) domain) is a conserved domain found in the mitotic checkpoint proteins Mad3, Bub1, and BubR1. These proteins are crucial components of the spindle assembly checkpoint (SAC), a surveillance mechanism that ensures the proper segregation of chromosomes during cell division by monitoring kinetochore-microtubule attachment [59].

Structure in RepeatsDB:

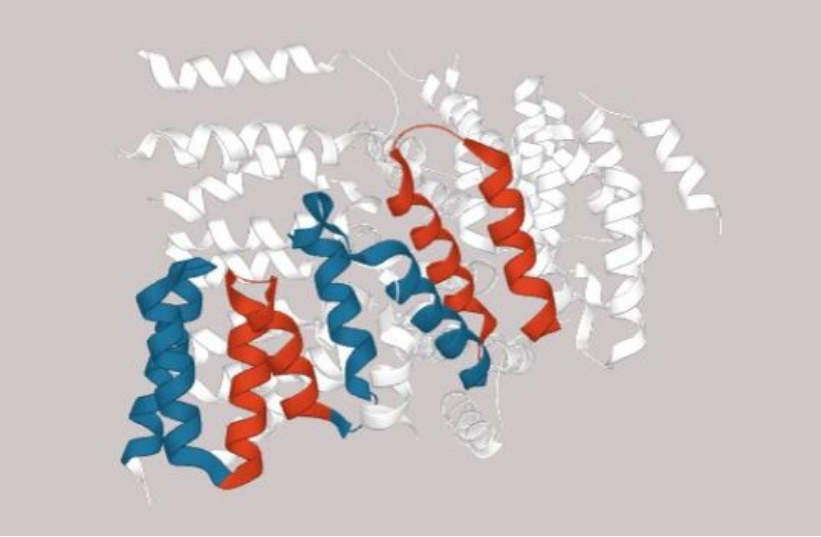


Fig. 3.21: Structure of 3eslA in RepeatsDB

Cluster 3:

3 members

Region ID	**Uniprot ID**
4xtlA_21_235	N0DKS8
4y9hA_9_199	B0R5N9
5h2pA_9_194	P02945

Annotations: Bacteriorhodopsin-like protein (PF01036)

Description: Bacteriorhodopsin-like proteins are a family of transmembrane proteins that share structural similarities with bacteriorhodopsin, a light-driven proton pump found in the plasma membrane of certain halophilic archaea. These proteins are characterized by their seven-transmembrane alpha-helical structure and their ability to bind a retinal chromophore.

The structure consists of seven transmembrane alpha-helices arranged in a barrel-like shape, with the retinal chromophore covalently bound to a lysine residue in one of the helices. Light-induced

isomerization of the retinal chromophore triggers a series of conformational changes in the protein, resulting in the translocation of protons [60].

Structure in RepeatsDB:

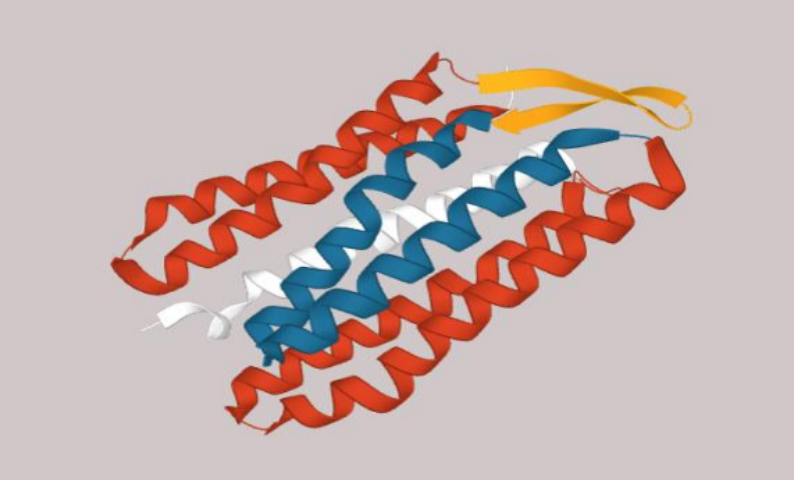


Fig. 3.22: Structure of 4xtlA in RepeatsDB

3.2.9 Curation Inconsistency

During the investigation of Alpha-Solenoids, a primary challenge encountered was the inconsistency in the curation of specific entries. Initially, the "Representative Unit" matrix was employed as a basis, but as the clustering and classification processes progressed, issues arose with certain entries. Members within the same clan formed two distinct clusters, despite being the same type of repeats. This conundrum persisted until a comparison of their curation order in RepeatsDB revealed inconsistencies in their curation.

The affected clans were 3.3.3.1 and 3.3.2.5, both of which will be elaborated upon in the subsequent sections, detailing the nature of the issue and the temporary solution implemented to maintain the research trajectory.

Case 1: 3.3.3.1 Clan

As depicted in the following figure, clustering the Representative Units (1-unit samples) resulted in two separate clusters at relatively high distance thresholds (Eps 0.35), even though evidence suggested they were the same type of repeats.

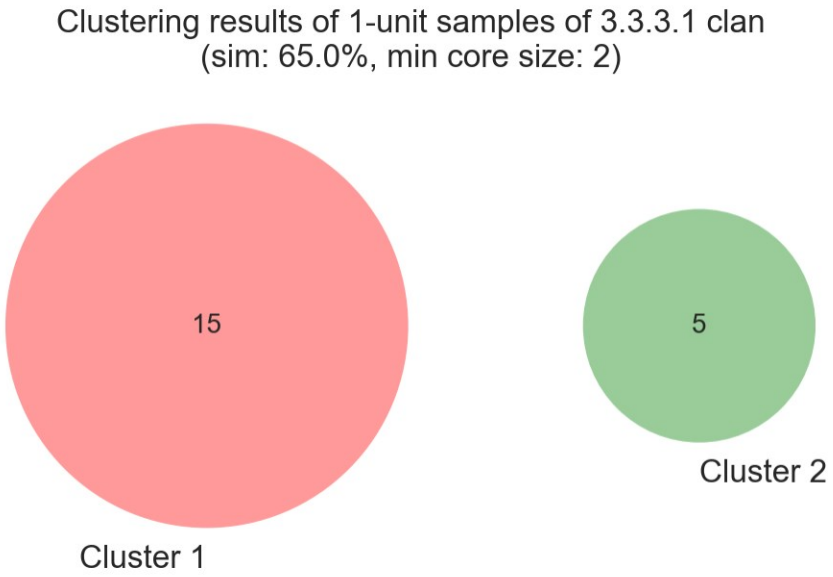


Fig. 3.23: 1-unit samples of 3.3.3.1 clan forming two distinct clusters at low similarities

Examination of their structures in RepeatsDB demonstrated two distinct curation approaches. Of the 20 clan members, 15 were curated similarly to A, and 5 were curated similarly to B (Fig. 3.24).

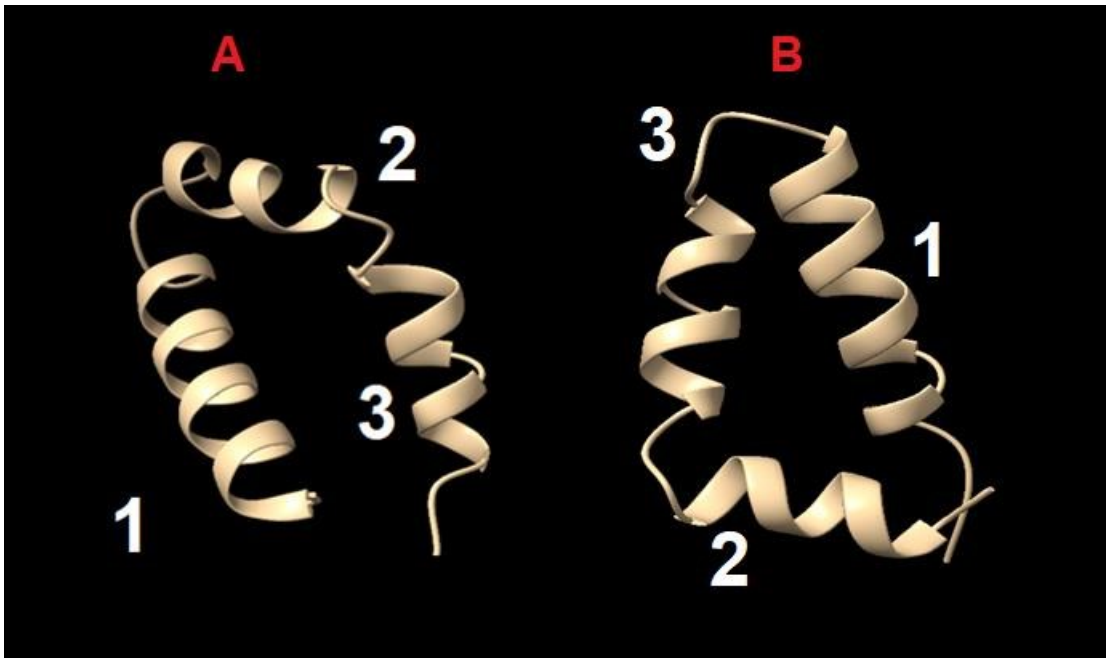


Fig. 3.24: Different order of structural motifs in two different annotations of the same repeat

Further inquiries in third-party databases and associated literature confirmed that the appropriate curation order for these repeats is, indeed, form A. To bypass waiting for curation corrections, a distance matrix based on "3-Consecutive Units" was utilized, as previously described in the Materials and Methods section. This approach resolved the issue, allowing clan members to remain in a single cluster even at more stringent distance thresholds (Eps 0.25) (Fig. 3.25).

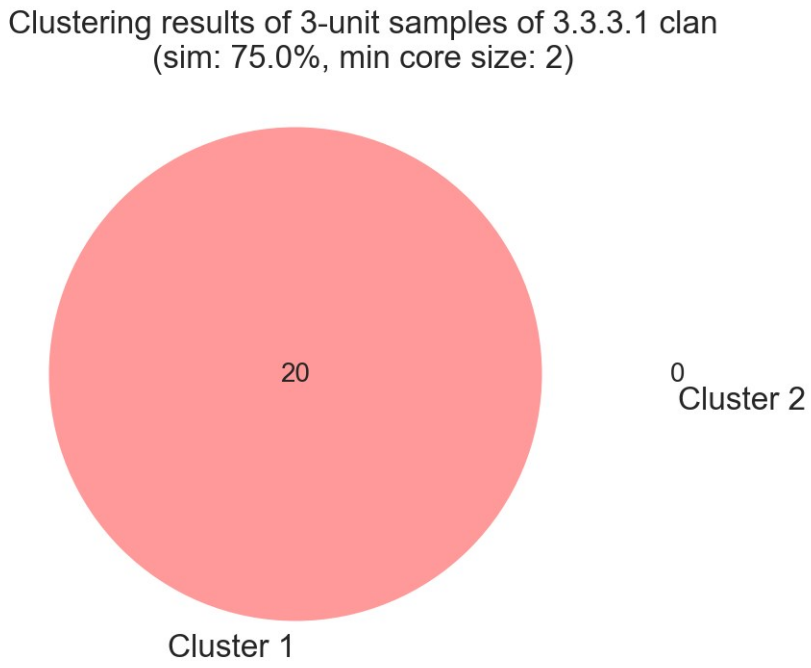


Fig. 3.25: Clustering of 3-unit samples of 3.3.3.1 at high similarity levels results in a single unified cluster

Case 2: 3.3.2.5 Clan

An identical issue arose with this clan, and it was addressed in the same manner (Fig. 3.26).

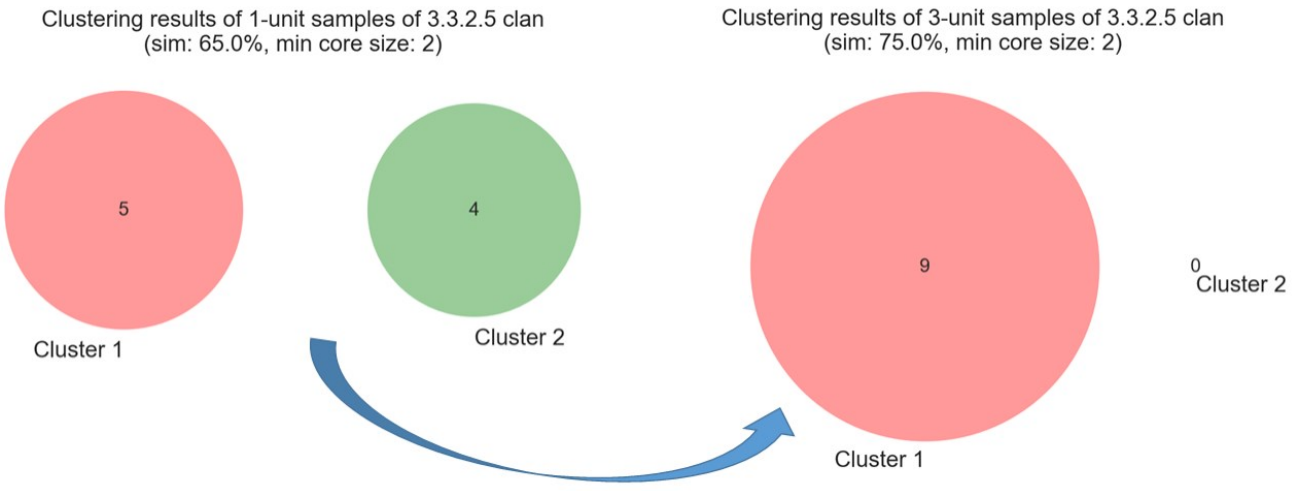


Fig. 3.26: How the change of samples from 1-unit to 3-unit resolves the fractionation

Five clan members were curated similarly to A, and four were curated similarly to B, as shown in the following figure.

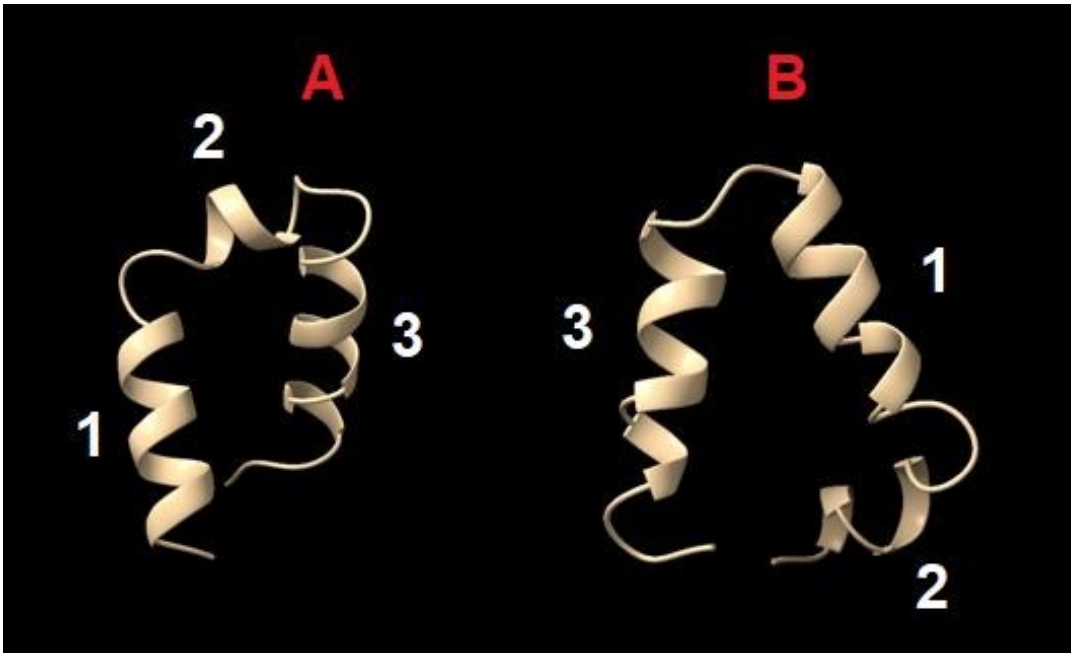


Fig. 3.27: Different order of structural motifs in two different annotations of the same type of repeat

Employing a similar approach to the previous instance, it was determined that form A is the correct curation method for these repeats. Typically, in the Tetratrichopeptides family, the shorter helix (A. 2) serves as the connecting segment between the two longer anti-parallel helices (A. 1&3).

The Permanent Solution

Ultimately, the complete resolution of this issue necessitated correcting the curation for those curated in the incorrect order. However, this occurred long after the research had concluded. The subsequent charts illustrate how the corrected curation impacted the clustering of these repeats using a "Representative Unit" (1-unit sample) matrix.

Fig. 3.28: Unified clusters are formed for both clans after correcting and unifying the annotations

In the new clustering experiment with refined curations, 1-unit samples from both groups formed a unified cluster at 70% similarity, a 5% increase compared to the previous experiment, where they separated into two clusters. This substantiates the claim that they do genuinely belong to the same clan.

3.3 Summary

As delineated in the preceding section, the pipeline was implemented for both sets of Alpha-Solenoids and Beta-Propellers. However, in order to mitigate superfluous repetition, the comprehensive outcomes obtained through the successive stages of analysis were exclusively presented for Alpha-Solenoids. This section aims to provide a concise summary of the significant findings obtained from the application of the pipeline on both groups of TRPs.

3.3.1 Alpha-Solenoids

Key findings and actions taken during this research include the following:

1. A distance matrix of "3-consecutive units" appeared to be the most reliable choice
2. Members within a clan demonstrated a structural similarity range of 70-75%.
3. Two Clans of 3.3.2.2 and 3.3.2.1 underwent modifications:

a. Clan 3.3.2.2 was refined to exclusively include Adaptin_N repeats.
b. Clan 3.3.2.1 was subdivided into three clans encompassing: 1. Genuine TPR 2. SNAP (TPR-like) 3. PPTA

4. Three outlier entries in the dataset were detected and subsequently identified
5. A total of 54 entries with missing annotations were completely annotated with high confidence.
6. Members of the three major families (Ankyrin, Tetratrichopeptides, and Armadillo repeats) exhibited a structural similarity range of 62-67%.
7. To annotate repeats with their associated family, 113 entries with missing annotations (Almost half of the total) were annotated with high confidence.
8. Through the examination of entries failing to gain any sort of annotations from the experiments, four groups of repeats were proposed via an additional clustering round.
9. The curation inconsistency in Clans 3.3.3.1 and 3.3.2.5 was identified and subsequently addressed and resolved.

3.3.2 Beta-Propellers

The comprehensive pipeline employed for studying the Alpha-Solenoids was replicated with minor modifications to accommodate the distinct properties of Beta-Propellers (4.4.x.x), which belong to a separate class of protein repeats. To minimize redundancy, only key findings and critical steps taken during the investigation are summarized below:

1. Selection of Distance Matrix: A "Whole Region" distance matrix was adopted for the basis of the investigation, as the number of units within each region could be an important factor for the classification of Beta-Propellers.

2. Analogous to Alpha-Solenoids, structural similarities of 70-75% were observed within Beta-Propeller clan members (Fig. 3.29).

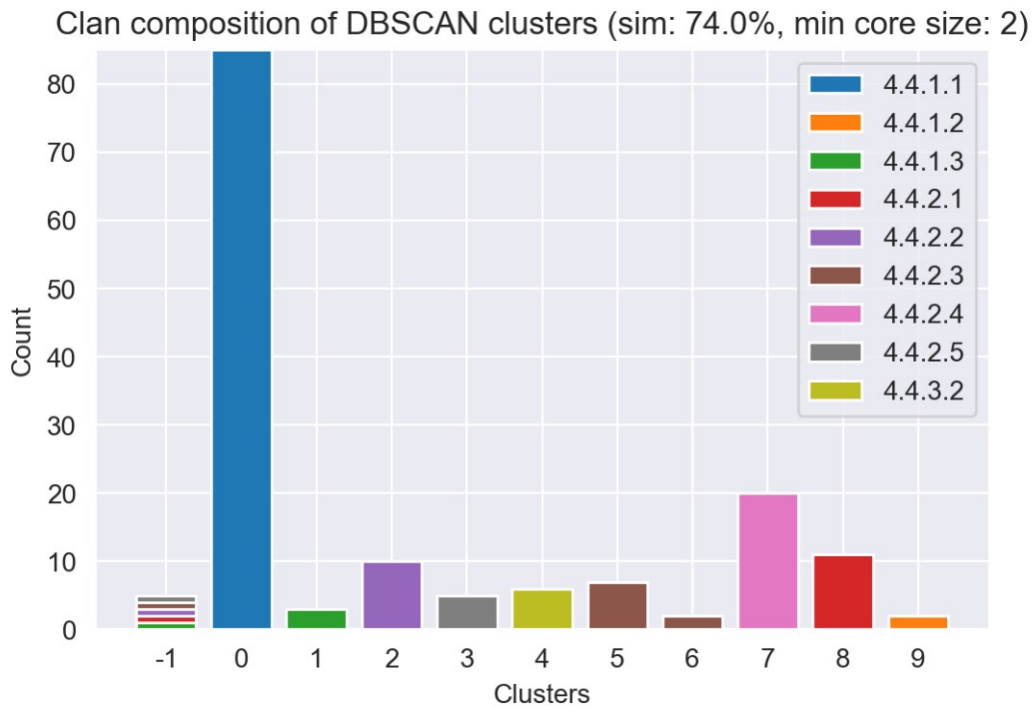


Fig. 3.29: Clan composition of clusters formed by clustering of Beta-Propeller

3. The Clan 4.4.2.3 was subdivided into two clans:
    a. 4.4.2.3_A comprising Sialidase BNR repeats (Cluster 5 in Fig. 3.29)
    b. 4.4.2.3_B comprising Xyloglucanase BNR repeats (Cluster 6 in Fig. 3.29)

    As depicted in the figure below, the structural comparison of the two groups indicated that 4.4.2.3_B (orange) features seven blades, as opposed to the six-bladed propellers of 4.4.2.3_A (blue), and lacks the characteristic insertion found in Sialidases (4.4.2.3_A).

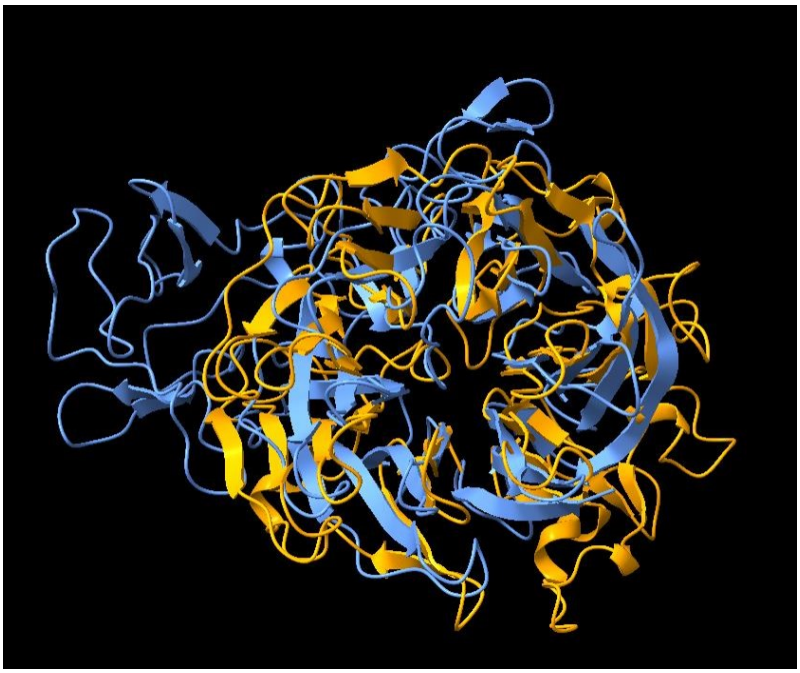


Fig. 3.30: Structural alignment of the fractionated clusters of 4.4.2.3 clan (4.4.2.3_A vs 4.4.2.3_B)

4. Five outlier entries within the dataset were detected and subsequently identified

5. A total of 76 entries with missing annotations were accurately and confidently annotated.

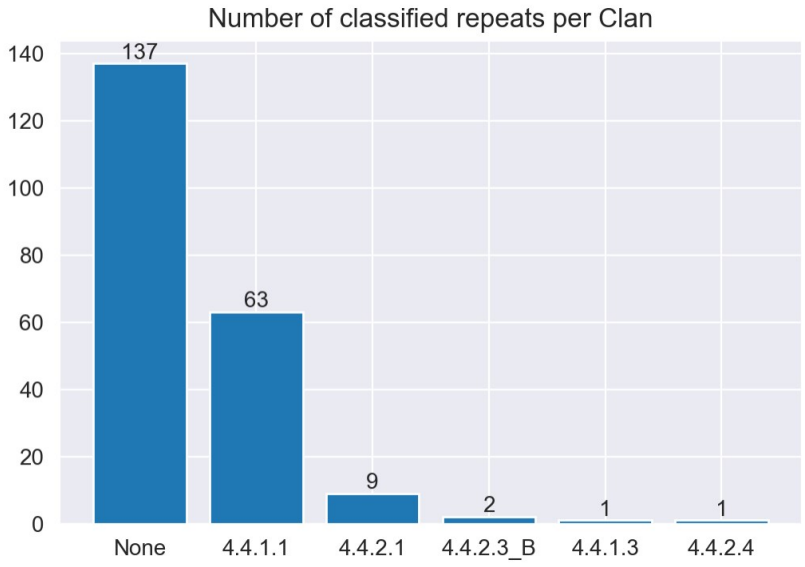


Fig. 3.31: KNN classification results of classifying the unclassified Beta-Propeller repeats

6. Upon evaluating entries that failed to receive annotations from the experiments, an additional clustering round proposed 12 groups of repeats, including PQQ, Hemopexin, and FG-GAP repeats.

4. Conclusion

In this study, we conducted an exhaustive and systematic analysis of protein tandem repeats in RepeatsDB, with a primary focus on Alpha-Solenoids and Beta-Propellers. By employing a variety of advanced computational techniques, algorithms, and statistical analyses, we made substantial contributions to the enhancement of the existing classification system in RepeatsDB, ultimately providing a deeper and more comprehensive understanding of protein tandem repeats.

Our initial statistical analysis served as an essential step in identifying the key tandem repeat protein groups for further investigation. This analysis revealed that Alpha-Solenoids and Beta-Propellers were of the utmost importance due to their high number of Reviewed entries. Additionally, we observed a significant number of entries with missing annotations, underscoring the need for more rigorous efforts towards annotating these protein repeats. Furthermore, despite RepeatsDB's existing classification system being sufficiently accurate and effective, we identified opportunities for improvement and refinement.

As the study was centered on the structural analysis of protein tandem repeats, to establish a solid foundation for our analysis, we employed TM-align for pairwise structural alignment, with the resultant TM-scores being used to construct a variety of distance matrices.

Multidimensional Scaling (MDS) was utilized for dimensionality reduction and visualization, enabling us to generate scatter plots and efficiently analyze the relationships between different protein repeat structures. This approach allowed us to uncover previously unknown structural patterns, providing valuable insights into the complex nature of protein structural comparison.

We further refined our understanding of the dataset by employing the density-based clustering algorithm, DBSCAN. One of the most significant findings concerning the clustering of protein repeat structures was that members within a Clan exhibit a structural similarity range of 70-75%. This discovery is crucial, as prior to this study, there was no computational support for establishing Clan boundaries. Additionally, this method proved particularly effective in detecting and identifying outlier entries, refining existing clans, and proposing new repeat groups.

Upon establishing a more reliable classification system through clustering analyses, we turned our attention to supervised classification methods. We carefully selected representative entries for each group of protein tandem repeats and conducted a supervised classification experiment using the

K-Nearest Neighbors (KNN) algorithm. This approach facilitated the automatic annotation and labeling of entries with missing annotations in the database, leading to the successful annotation of a total of 130 entries with high confidence.

The automatic annotation methodology developed in this study can considerably enhance the performance of curators of RepeatsDB in the sensitive and time-consuming process of curation and annotation of protein repeats. As it is of great value to double-validate biological findings through machine learning techniques, this methodology enables curators to focus on high-confidence annotations while reducing the manual annotation workload. Furthermore, the approach can be applied to other software and programs for a more accurate prediction of tandem repeat proteins, contributing to a wider range of applications in the field of bioinformatics.

In conclusion, our research has made significant strides in advancing the understanding of protein tandem repeats in RepeatsDB, with particular emphasis on Alpha-Solenoids and Beta-Propellers. The improvements in the classification system and database enhancements can greatly benefit future research in the fields of structural biology and bioinformatics. Furthermore, the methodologies, algorithms, and statistical analyses employed in this study can be extended to other classes and subclasses of protein repeats, opening up new avenues for a more comprehensive understanding of protein structure and function in the future.

Acknowledgments

I extend my heartfelt gratitude to Prof. Silvio Tosatto, my supervisor at Biocomputing UP, whose guidance and support were invaluable throughout my research.

I am deeply thankful to Dr. Alexander Monzon, my co-supervisor, for his knowledge and assistance in completing this study.

I would like to express my appreciation to Dr. Paula Arrías, the lead biocurator of RepeatsDB, for her valuable insights into protein structures.

Lastly, my sincere appreciation to all individuals who have supported and assisted me on this academic journey.

References

[1] "Crick, F. H. C. (1958). On Protein Synthesis. Symposia of the Society for Experimental Biology, 12, 138-163."

[2] "Goeddel DV, Kleid DG, Bolivar F, Heyneker HL, Yansura DG, Crea R, Hirose T, Kraszewski A, Itakura K, Riggs AD. Expression in Escherichia Coli of Chemically Synthesized Genes for Human Insulin. Proc Natl Acad Sci U S A. 1979 Jan;76(1):106-10. Doi: 10.1073/Pnas.76.1.106. PMID: 85300; PMCID: PMC382885.," n.d.

[3] "Shelton AM, Zhao JZ, Roush RT. Economic, Ecological, Food Safety, and Social Consequences of the Deployment of Bt Transgenic Plants. Annu Rev Entomol. 2002;47:845-81. Doi: 10.1146/Annurev.Ento.47.091201.145309. PMID: 11729093.," n.d.

[4] *Branden, C., & Tooze, J. (1999). Introduction to Protein Structure. Garland Science.*

[5] *Voet, D., Voet, J. G., & Pratt, C. W. (2016). Fundamentals of Biochemistry: Life at the Molecular Level. Wiley.*

[6] "Perutz MF. Hemoglobin Structure and Respiratory Transport. Sci Am. 1978 Dec;239(6):92-125. Doi: 10.1038/Scientificamerican1278-92. PMID: 734439."

[7] "Koshland Jr, D. E., Némethy, G., & Filmer, D. 'Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits.' Biochemistry, 1966, 5(1), 365-385."

[8] "Dill KA, MacCallum JL. The Protein-Folding Problem, 50 Years on. Science. 2012 Nov 23;338(6110):1042-6. Doi: 10.1126/Science.1219021. PMID: 23180855."

[9] "Perutz, Max F., 'Myoglobin and the Structure of Proteins.' Nobel Lecture, December 11, 1962."

[10] "Ellegren, H. (2004). Microsatellites: Simple Sequences with Complex Evolution. Nature Reviews Genetics, 5(6), 435-445. Https://Doi.Org/10.1038/Nrg1348."

[11] "Gemayel, R., Cho, J., Boeynaems, S., & Verstrepen, K. J. (2012). Beyond Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory and Coding Sequences. Genes, 3(3), 461-480. Https://Doi.Org/10.3390/Genes3030461."

[12] "Kajava AV. Tandem Repeats in Proteins: From Sequence to Structure. J Struct Biol. 2012 Sep;179(3):279-88. Doi: 10.1016/j.Jsb.2011.08.009. Epub 2011 Aug 24. PMID: 21884799."

[13] "Usdin, K. and Woodford, K. J. 'CGG Repeats Associated with DNA Instability and Chromosome Fragility Form Structures That Block DNA Synthesis in Vitro.' Nucleic Acids Research, 1995, 23(20), 4202-4209."

[14] "Huang, H., Baxa, U., Sachs, G., and Sosnick, T. R. 'The Protein Structure and Structural Pathway of the Molten Globule State of Apomyoglobin.' Journal of Molecular Biology, 1999, 287(4), 755-773. Doi: 10.1006/Jmbi.1999.2632."

[15] "Mecham RP. Elastin Synthesis and Fiber Assembly. Ann N Y Acad Sci. 1991;624:137-46. Doi: 10.1111/j.1749-6632.1991.Tb17013.x. PMID: 1648323."

[16] "Prockop DJ, Kivirikko KI. Collagens: Molecular Biology, Diseases, and Potentials for Therapy. Annu Rev Biochem. 1995;64:403-34. Doi: 10.1146/Annurev.Bi.64.070195.002155. PMID: 7574488."

[17] "Cadigan KM, Nusse R. Wnt Signaling: A Common Theme in Animal Development. Genes Dev. 1997 Dec 15;11(24):3286-305. Doi: 10.1101/Gad.11.24.3286. PMID: 9407023."

[18] "Blom AM, Kask L, Dahlbäck B. Structural Requirements for the Complement Regulatory Activities of C4BP. J Biol Chem. 2001 Jul 20;276(29):27136-44. Doi: 10.1074/Jbc.M102445200. Epub 2001 May 21. PMID: 11369776."

[19] "Winokur, S. T., et al. (2003). Expression Profiling of FSHD Muscle Supports a Defect in Specific Stages of Myogenic Differentiation. Human Molecular Genetics, 12(22), 2895-2907."

[20] "Ter Haar E, Musacchio A, Harrison SC, Kirchhausen T. Atomic Structure of Clathrin: A Beta Propeller Terminal Domain Joins an Alpha Zigzag Linker. Cell. 1998 Nov 13;95(4):563-73. Doi: 10.1016/S0092-8674(00)81623-2. PMID: 9827808; PMCID: PMC4428171."

[21] "Ford, M. G. J., et al. (2001). Simultaneous Binding of PtdIns(4,5)P2 and Clathrin by AP180 in the Nucleation of Clathrin Lattices on Membranes. Science, 291(5506), 1051-1055."

[22] "Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P., 2001. Protein Repeats: Structures, Functions, and Evolution. Journal of Structural Biology, 134(2-3), Pp.117-131.," n.d.

[23] "Walsh I, Sirocco FG, Minervini G, Di Domenico T, Ferrari C, Tosatto SC. RAPHAEL: Recognition, Periodicity and Insertion Assignment of Solenoid Protein Structures. Bioinformatics. 2012;28:3257–3264.," n.d.

[24] “Kajava, A.V., 2001. Proteins with Repeated Sequence—Structural Prediction and Modeling. Journal of Structural Biology, 134(2-3), Pp.132-144.,” n.d.

[25] “Orr HT, Zoghbi HY. Trinucleotide Repeat Disorders. Annu Rev Neurosci. 2007;30:575-621. Doi: 10.1146/Annurev.Neuro.29.051605.113042. PMID: 17417937.”

[26] “Bella J, Eaton M, Brodsky B, Berman HM. Crystal and Molecular Structure of a Collagen-like Peptide at 1.9 A Resolution. Science. 1994 Oct 7;266(5182):75-81. Doi: 10.1126/Science.7695699. PMID: 7695699.”

[27] “Lupas AN, Gruber M. The Structure of Alpha-Helical Coiled Coils. Adv Protein Chem. 2005;70:37-78. Doi: 10.1016/S0065-3233(05)70003-6. PMID: 15837513.”

[28] “Kobe B, Kajava AV. When Protein Folding Is Simplified to Protein Coiling: The Continuum of Solenoid Protein Structures. Trends Biochem Sci. 2000 Oct;25(10):509-15. Doi: 10.1016/S0968-0004(00)01667-4. PMID: 11050437.”

[29] “Van Raaij MJ, Mitraki A, Lavigne G, Cusack S. A Triple Beta-Spiral in the Adenovirus Fibre Shaft Reveals a New Structural Motif for a Fibrous Protein. Nature. 1999 Oct;401(6756):935-938. DOI: 10.1038/44880. PMID: 10553913.”

[30] “Neer EJ, Schmidt CJ, Nambudripad R, Smith TF. The Ancient Regulatory-Protein Family of WD-Repeat Proteins. Nature. 1994 Sep 22;371(6495):297-300. Doi: 10.1038/371297a0. Erratum in: Nature 1994 Oct 27;371(6500):812. PMID: 8090199.”

[31] “Lee MS, Gippert GP, Soman KV, Case DA, Wright PE. Three-Dimensional Solution Structure of a Single Zinc Finger DNA-Binding Domain. Science. 1989 Aug 11;245(4918):635-7. Doi: 10.1126/Science.2503871. PMID: 2503871.”

[32] “Coward, E. and DrablvТ s, F., 1998. Detecting Periodic Patterns in Biological Sequences. Bioinformatics (Oxford, England), 14(6), Pp.498-507.”

[33] “Newman, A.M. and Cooper, J.B., 2007. XSTREAM: A Practical Algorithm for Identification and Architecture Modeling of Tandem Repeats in Protein Sequences. BMC Bioinformatics, 8(1), Pp.1-19.”

[34] “Jorda, J. and Kajava, A.V., 2009. T-REKS: Identification of Tandem REpeats in Sequences with a K-MeanS Based Algorithm. Bioinformatics, 25(20), Pp.2632-2638.”

[35] “Heger, A. and Holm, L., 2000. Rapid Automatic Detection and Alignment of Repeats in Protein Sequences. Proteins: Structure, Function, and Bioinformatics, 41(2), Pp.224-237.”

[36] “Szklarczyk, R. and Heringa, J., 2004. Tracking Repeats Using Significance and Transitivity. Bioinformatics, 20(Suppl_1), Pp.I311-I317.”

[37] “Bucher, P., Karplus, K., Moeri, N. and Hofmann, K., 1996. A Flexible Motif Search Technique Based on Generalized Profiles. Computers & Chemistry, 20(1), Pp.3-23.”

[38] “Biegert, A. and Söding, J., 2008. De Novo Identification of Highly Diverged Protein Repeats by Probabilistic Consistency. Bioinformatics, 24(6), Pp.807-814.”

[39] “Hrabe, T. and Godzik, A., 2014. ConSole: Using Modularity of Contact Maps to Locate Solenoid Domains in Protein Structures. BMC Bioinformatics, 15(1), Pp.1-12.”

[40] “Do Viet, P., Roche, D.B. and Kajava, A.V., 2015. TAPO: A Combined Method for the Identification of Tandem Repeats in Protein Structures. FEBS Letters, 589(19), Pp.2611-2619.”

[41] “Hirsh, L., Piovesan, D., Paladin, L. and Tosatto, S.C., 2016. Identification of Repetitive Units in Protein Structures with ReUPred. Amino Acids, 48, Pp.1391-1400.”

[42] “Di Domenico T, Potenza E, Walsh I, Parra RG, Giollo M, Minervini G, Piovesan D, Ihsan A, Ferrari C, Kajava AV, Tosatto SC. RepeatsDB: A Database of Tandem Repeat Protein Structures. Nucleic Acids Res. 2014 Jan;42(Database Issue):D352-7. Doi: 10.1093/Nar/Gkt1175. Epub 2013 Dec 5. PMID: 24311564; PMCID: PMC3964956.”

[43] “Chakrabarty B, Parekh N. DbStRiPs: Database of Structural Repeats in Proteins. Protein Sci. 2022 Jan;31(1):23-36. Doi: 10.1002/pro.4052. Epub 2021 Mar 6. PMID: 33641184; PMCID: PMC8740836.,” n.d.

[44] “Libbrecht MW, Noble WS. Machine Learning Applications in Genetics and Genomics. Nat Rev Genet. 2015 Jun;16(6):321-32. Doi: 10.1038/Nrg3920. Epub 2015 May 7. PMID: 25948244; PMCID: PMC5204302.”

[45] “Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine Learning Applications in Cancer Prognosis and Prediction. Comput Struct Biotechnol J. 2014 Nov 15;13:8-17. Doi: 10.1016/j.Csbj.2014.11.005. PMID: 25750696; PMCID: PMC4348437.”

[46] "Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The Rise of Deep Learning in Drug Discovery. Drug Discov Today. 2018 Jun;23(6):1241-1250. Doi: 10.1016/j.Drudis.2018.01.039. Epub 2018 Jan 31. PMID: 29366762."
[47] "Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13(1), 21-27."
[48] "Schubert, E., Sander, J., Ester, M., Kriegel, H.P. and Xu, X., 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), Pp.1-21."
[49] "Kruskal, J.B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika 29, 1–27 (1964). Https://Doi.Org/10.1007/BF02289565."
[50] "Yang Zhang, Jeffrey Skolnick, TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score, Nucleic Acids Research, Volume 33, Issue 7, 1 April 2005, Pages 2302–2309."
[51] "Dong R, Peng Z, Zhang Y, Yang J. MTM-Align: An Algorithm for Fast and Accurate Multiple Protein Structure Alignment. Bioinformatics. 2018 May 15;34(10):1719-1725. Doi: 10.1093/Bioinformatics/Btx828. PMID: 29281009; PMCID: PMC5946935."
[52] "Kruskal, J.B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika 29, 1–27 (1964). Https://Doi.Org/10.1007/BF02289565."
[53] "Jin, X., Han, J. (2011). K-Means Clustering. In: Sammut, C., Webb, G.I. (Eds) Encyclopedia of Machine Learning. Springer, Boston, MA. Https://Doi.Org/10.1007/978-0-387-30164-8_425."
[54] "D'Andrea LD, Regan L. TPR Proteins: The Versatile Helix. Trends Biochem Sci. 2003 Dec;28(12):655-62. Doi: 10.1016/j.Tibs.2003.10.007. PMID: 14659697."
[55] "Maurer-Stroh S, Washietl S, Eisenhaber F. Protein Prenyltransferases. Genome Biology. 2003 ;4(4):212. DOI: 10.1186/Gb-2003-4-4-212. PMID: 12702202; PMCID: PMC154572."
[56] "Rice LM, Brunger AT. Crystal Structure of the Vesicular Transport Protein Sec17: Implications for SNAP Function in SNARE Complex Disassembly. Mol Cell. 1999 Jul;4(1):85-95. Doi: 10.1016/S1097-2765(00)80190-2. PMID: 10445030."
[57] "Yaffe MB. How Do 14-3-3 Proteins Work?-- Gatekeeper Phosphorylation and the Molecular Anvil Hypothesis. FEBS Letters. 2002 Feb;513(1):53-57. DOI: 10.1016/S0014-5793(01)03288-4. PMID: 11911880."
[58] "Akhmanova A, Steinmetz MO. Control of Microtubule Organization and Dynamics: Two Ends in the Limelight. Nat Rev Mol Cell Biol. 2015 Dec;16(12):711-26. Doi: 10.1038/Nrm4084. Epub 2015 Nov 12. PMID: 26562752."
[59] "Kevin G. Hardwick, Raymond C. Johnston, Dana L. Smith, Andrew W. Murray; MAD3 Encodes a Novel Component of the Spindle Checkpoint Which Interacts with Bub3p, Cdc20p, and Mad2p. J Cell Biol 6 March 2000; 148 (5): 871–882. Doi: Https://Doi.Org/10.1083/Jcb.148.5.871."
[60] "Ernst OP, Lodowski DT, Elstner M, Hegemann P, Brown LS, Kandori H. Microbial and Animal Rhodopsins: Structures, Functions, and Molecular Mechanisms. Chem Rev. 2014 Jan 8;114(1):126-63. Doi: 10.1021/Cr4003769. Epub 2013 Dec 23. PMID: 24364740; PMCID: PMC3979449."