UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA
(FINAL PROJECT PAPER)

RAZISKOVANJE ZNOTRAJDOMENE PROŽNOSTI PROTEINSKIH DOMEN Z UPORABO EKSPERIMENTALNIH PODATKOV IZ PODATKOVNE BAZE SCOPe
(EXPLORING INTRA-DOMAIN FLEXIBILITY OF PROTEIN DOMAINS USING EXPERIMENTAL DATA FROM THE SCOPe DATABASE)

MARIJA RAKIĆ

2023
ZAKLJUČNA NALOGA (FINAL PROJECT PAPER)
RAKIĆ

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga
(Final project paper)

Raziskovanje znotrajdomene prožnosti proteinskih domen z uporabo eksperimentalnih podatkov iz podatkovne baze SCOPe

(Exploring intra-domain flexibility of protein domains using experimental data from the SCOPe database)

Ime in priimek: Marija Rakić
Študijski program: Bioinformatika
Mentor: izr. prof. dr. Jure Pražnikar
Somentor: doc. dr. Marko Jukić

Koper, februar 2023

Ključna dokumentacijska informacija

Ime in PRIIMEK: Marija RAKIĆ

Naslov zaključne naloge: Raziskovanje znotrajdomene prožnosti proteinskih domen z uporabo eksperimentalnih podatkov iz podatkovne baze SCOPe

Kraj: Koper

Leto: 2022

Število listov: 37 Število slik: 7 Število tabel: 5

Število referenc: 72

Mentor: izr. prof. dr. Jure Pražnikar

Somentor: doc. dr. Marko Jukić

Ključne besede: SCOPe podatkovna zbirka, koren srednje kvadratne fluktuacije, b-faktor, beljakovinske domene, prožnost proteinov

Izvleček: Konformacijska prožnost proteinov se uporablja na številnih področjih, od bioinformatičnih študij do odkrivanja zdravil. Podatkovna zbirka SCOPe vsebuje več replik proteinskih domen, kar zagotavlja uporabne informacije za preučevanje prožnosti proteinov. Celotno podatkovno zbirko SCOPe smo analizirali z uporabo programskega paketa Bio3D R. Glavni cilj te študije je bil raziskati povezavo med temperaturnim faktorjem in lokalnimi floktuacijami RMSF (root-mean-square-fluctuations). Ugotovili smo, da med številom enakih proteinskih domen in srednjo vrednostjo RMSF ni povezave. Podobno število enakih proteinskih domen ni bilo povezano s korelacijo med RMSF in temperaturnim faktorjem. Med številom preostankov oziroma velikostjo proteinske domene in povprečno RMSF ni bilo korelacije. Poleg tega število ostankov ni bilo v korelaciji s korelacijo med faktorjema RMSF in temperaturnega faktorja. Višja povprečna RMSF je negativno korelirala s korelacijo med povprečno RMSF in temperaturnim faktorjem. Ugotovili smo, da konformacijska prožnost ni povezana z velikostjo domene, in prav tako ni povezan z samim zvitjem proteina. Poleg tega je RMSF v nekaterih primerih zelo dobro povezan s temperaturnim faktorjem. Po drugi strani pa najdemo številne primere, kjer obstaja negativna korelacija med RMSF in temperaturnim faktorjem.

Key document information

Name and SURNAME: Marija RAKIĆ

Title of the final project paper: Exploring intra-domain flexibility of protein domains using experimental data from the SCOPe database

Place: Koper

Year: 2022

Number of pages: 37 Number of figures: 7 Number of tables: 5

Number of references: 72

Mentor: Assoc. Prof. Jure Pražnikar, PhD

Co-Mentor: Assist. Prof. Marko Jukić, PhD

Keywords: SCOPe database, root mean square fluctuations, b-factor, protein domains, protein flexibility

Abstract: Protein flexibility is used in numerous fields, from bioinformatics studies to drug discovery. The SCOPe database contains multiple replicates of protein domains, providing rich information for the study of protein flexibility. We analyzed the entire SCOPe database using Bio3D R packages. The focus of the present study was to investigate the relationship between the temperature factor and root mean square fluctuations (RMSF). We found that there was no correlation between the number of equal/similar entries and the mean RMSF. Similarly, the number of equal/similar entries did not correlate with the correlation between RMSF and temperature factor. There was no correlation between the number of residues and the mean RMSF. In addition, the number of residues did not correlate with the correlation of RMSF and temperature factor. Higher mean RMSF correlated negatively with the correlation of mean RMSF and temperature factor. We found that conformational flexibility was not related to domain size or to the folding of the protein itself. Moreover, RMSF is very well correlated with temperature factor in some cases. On the other hand, we find numerous examples where there is a negative correlation between RMSF and temperature factor.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my mentor and professor Jure Pražnikar, for his support, a lot of patience, encouragement, feedback and guidance during the writing of this final project paper.

Additionally, I would like to thank my family for being my support, especially my brother, who serves as my inspiration and strength for everything that lies ahead.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

3D – three dimensional

CD - circular dichroism

cryo-EM - cryogenic electron microscopy

IDPs - intrinsically disordered proteins

NMR - nuclear magnetic resonance

PCA - principal component analysis

PDB – Protein Data Bank

RMSD – Root Mean Square Deviation

RMSF - Root Mean Square Fluctuations

SCOP - Structural Classification of Proteins

SCOPe - Structural Classification of Proteins extended

SPA - single particle analysis

STA - Sub-tomogram averaging

TEM – transmission electron microscopy

UniProt - Universal Protein Resource

1 INTRODUCTION

In his book How proteins work, Mike Williamson states that „The Last Universal Common Ancestor" probably already had most of the protein folds and functions that we see now." [1] He further states that „once a protein has reached a certain functional level it starts to "fossilize" into an evolutionary dead-end" [1]. Therefore, we can only assume that current enzymes are only the best forms that have been found by evolution so far. Proteins are much more than just macromolecular constituents of cellular structures. They deservingly carry one of the four pillars called "molecules of life", along with lipids, carbohydrates, and nucleic acids. Exploring both their structure and function is crucial to completely understanding its role.

1.1 Protein structure

Four different levels determine protein structure. The primary structure of protein consists of a polypeptide chain formed of linked sequences of amino acids. During protein biosynthesis, peptide bonds link each amino acid to the next amino acid [1]. Residues are amino acids that are linked together to form a polypeptide chain and the main chain consists of carbon, nitrogen, and oxygen atoms that form the protein backbone [2]. An amino group at the start of a peptide chain is known as the N-terminus, and a carboxylic acid group at the end is known as the C-terminus. A primary structure also includes the locations of disulfide bonds that covalently link different polypeptide chains together. Cysteine pairs are connected via thiol groups (-SH) on their side chains, which help stabilize protein structures [3,4].

The secondary structure arises from the arrangement of the primary structure using hydrogen bonds between the C=O and NH group of the polypeptide backbone [5]. There are no amino acid side chains involved in mentioned hydrogen bonds in the secondary structure. As a part of the secondary structure, many proteins are either folded or coiled in distinct patterns. Common secondary elements include alpha-helix and beta-pleated sheets.

The alpha-helix is a coil with hydrogen bonds between every fourth amino acid. In this regular coil shape the R-groups are pointed outwards away from the peptide backbone. A full turn of a helix is completed with approximately 3.6 residues [3].

Alpha-keratin, fingernails, and toenails are the most common examples of proteins comprised almost entirely of alpha-helices [5]. Alpha-helices with specific hydrophobic properties are also present in transmembrane proteins. These hydrophobic properties allow them to traverse membranes and stabilize them within the cell membrane [5].

Beta-sheets are composed of two or more polypeptide chains that lie side by side and are connected by hydrogen bonds [3]. The N-terminus of beta-sheets contains a free amino acid, while the C-terminus contains a free hydroxylic group. There are two types of beta-sheets: parallel and anti-parallel beta-pleated sheets, depending if two beta strands run in the same or the opposite direction [1,5]. The core of many globular proteins is comprised of beta-pleated sheets [3].

Beta-barrel is a form of the secondary structure composed of antiparallel beta-strands, twisted, and coiled into a barrel where the first strand of hydrogen bonds to the last strand. Beta-barrels are common in aquaporins [3].

The three-dimensional (3D) arrangement of the polypeptide chains presents the tertiary protein structure. The tertiary structure is usually divided into domains and contains one or more active sites on the surface [6]. It is important to underline that the primary structure determines the tertiary structure. Certain proteins denatured by heat, extreme pH, or denaturing agents will regain their native structure and function when the conditions return to those in which the native structure was stable [6]. This underlines that protein denaturation is, in many cases, a reversible process and that the primary structure regulates the 3D structure.

Interaction and arrangement of multiple folded protein chains (subunits) and the formation of a larger multisubunit protein complex represent quaternary structure [1,3]. Many proteins are organized into distinct protein domains.

1.2 Protein domains

Short segments of a protein's structure, motifs (or super secondary structure), are arrangements of the secondary forms (i.e. α-helices and β-sheets) into recognizable conformations. Domains are compact independent 3D structures composed of motifs and secondary structure elements in individual proteins [7]. They fold independently from the rest of the polypeptide chain. It is common for larger proteins to contain multiple domains linked together with each domain carrying out a specific function. The arrangement of secondary structure elements that describe a protein domain's shape is called its fold. Interestingly, there are only about 2200 recognizable protein folds despite the number of possible amino acid combinations [3]. The term domain is generally used to describe local, compact units of structure, characterized by hydrophobic interiors and hydrophilic exteriors that do not further subdivide [8]. It is possible to think of domains as semi-independent globular folding units. Consequently, they are capable of successfully combining with other domains and evolving new functions as a result of this property [7,8].

Two of the most commonly used systems for classification of protein domains are the CATH (class (C), architecture (A), topology/fold (T), homologous superfamily (H), which represent the CATH structural hierarchy) and Structural Classification of Proteins (SCOP) systems [9,10]. Both are classification systems in which proteins are organized into different levels based on the structural and sequence similarities. Additionally, the four main protein 'types' which all correlate with characteristic sequence and structural features are: globular (cytoplasmic enzymes), membrane proteins (receptors), fibrous proteins (for example, collagen), and intrinsically disordered proteins [3].

The use of spectroscopic methods, such as circular dichroism (CD) and fluorescence, gives basic information on protein structure, whereas high-resolution methods, such as X-ray crystallography, nuclear magnetic resonance (NMR) and cryogenic electron microscopy (cryo-EM), give atomic details about protein structure [3].

1.3 Protein dynamics and fluctuations

The biological function of proteins is highly dependent on their dynamic characteristics. In many cases, proteins transition structurally from one conformation to another in order to engage in biological activity [11]. Biological activity involves adopting a specific conformation, local fluctuations and structural transitions between conformations [11]. These fluctuations vary from small, local fluctuations (induced-fit ligand binding) to slow and global changes in transitions of allosteric proteins [12]. Low-energy modes of fluctuations around equilibrium conformations determine concerted motions in protein structures. Frequently used methods for obtaining these modes are molecular dynamics simulations, with the principal component analysis (PCA) of the covariance matrix and equilibrium structure (using X-ray scattering or NMR) [12]. A complex description of the molecular structural fluctuations is an important part of the analysis in MD simulation [13]. Complex as it is, interpretation of these changes could be hard to interpret from a functional perspective. Root Mean Square Deviation (*RMSD*) and the Root Mean Square Fluctuations (*RMSF*) are two of the most common measures of structural fluctuations [13]. The RMSD is the average displacement of the atoms along a frame of the simulation or in relation to a crystallographic structure at an instant of the simulation [13]. The RMSD can be used to analyze time-dependent motions in a structure. An analysis of the time scale of the simulation is frequently employed to determine whether the structure is stable or if it is diverging from the initial coordinates [13]. The divergence from the initial coordinates is typically interpreted as an indication that the simulation hasn't been equilibrated [13]. In an equilibrated simulation, in which the structure of interest fluctuates around an average conformation, it makes sense to compute the fluctuations of each subset of the structure (each atom, say) relative to the average structure [13]. RMSF measures how far an atom, or group

of atoms, has moved relative to the reference structure, averaged across all atoms [13]. Equations 1 and 2 present the formula for RMSD and RMSF, respectively.

$$\text{RMSD} = \sqrt{\frac{\sum(x_e - x_o)^2}{n}} \quad (1)$$

where x_e presents expected values, x_o observed values, and n total number of values.

$$\text{RMSF} = \left[\frac{1}{T}\sum_{t_j=1}^{T}\left|r_i(t_j) - r_i^{ref}\right|^2\right]^{\frac{1}{2}} \quad (2)$$

where T is the time over which one wants to average and r_i^{ref} is the *i-reference* position of particle *i*.

Oscillation amplitudes of the atoms around their equilibrium positions in crystal structures are monitored by B-factors [14]. B-factor (or Debye−Waller factor) describes the attenuation of X-ray scattering or coherent neutron scattering caused by thermal motion [15]. The following formulas (3 and 4) defines B-factor (also called temperature factor or atomic displacement parameter):

$$B = 8\pi^2(U_i^2) \quad (3)$$

where U_i^2 is the mean square displacement of atom *i*.

$$B = \frac{8\pi^2 RMSF^2}{3} \quad (4)$$

Other factors besides amplitudes of the atomic oscillations around the equilibrium influence the B factor, like static or dynamic conformational disorders, and occupancies, which is in range between zero and one [14]. It is often assumed that oscillation amplitudes are identical in all directions in protein crystallography, but this is rarely the case. In high resolution, when diffraction data are plentiful enough, B-factors are refined anisotropically when three principal components of oscillation are refined independently. This approximation, however, can also deviate considerably from reality [14]. B factors have been used to identify thermal motion paths, correlation of the rotameric state of amino acid side chains, and improvement of the protein superposition algorithms [14,16,17]. Simulations of multiple picosecond molecular dynamics have also been used to predict the B-factor values [18]. There is a certain amount of debate regarding the accuracy of the B factor in protein crystal structures. In a recent paper, Carugo discusses in detail that B factors are not really reproducible [19]. He presented the estimated errors of 9 and 6 Å in an ambient and low temperature structures [19]. Moreover, he noticed that the level of reproducibility remained unchanged for the last two decades. He went further to indicate that the B-factor monitor other features, including crystal defects, diffraction decay, etc. [19] Carugo also stated that normalization of B factors when comparing different crystal structure determinations is mandatory [19]. Referencing the paper by Pearce and Gross, he concluded that decomposition of the B factor into several components to discover the fraction of the B-factor that is really due to local positional fluctuations is a possible solution for the higher accuracy of the B factor [19,20].

1.4 Methods for unveiling protein structure

1.4.1 X-ray crystallography

X-ray crystallography can accurately reveal the biomolecule's structure held within crystals. A wavelength of X-rays approximates the length of covalent bonds, therefore X-rays are widely used since they are ideal for resolving atoms separated by these distances [3]. Since modern crystallography methods can be performed at cryogenic temperatures, we can determine large complex structures (between 2 and 100 nm) [3]. As an X-ray beam is focused on a protein crystal, its electric component interacts with the electron clouds surrounding atom nuclei, causing diffraction. The diffracted X-rays generate reflections on a detector that have an intensity [3]. Based on Bragg's geometric law of constructive interference in crystals, the spots are the result of reflections of the crystal at a certain angle relative to the original beam. The experiment is repeated with the crystal rotated to multiple different orientations which allow the angle of the incoming X-rays to change with respect to the crystal providing new reflections [3]. Upon recording all diffraction patterns, a dataset of spots that corresponds to many of the possible constructive interference diffraction events is compiled for each protein. Crystals produce diffraction patterns based on how many electrons lie on imaginary parallel planes called Bragg planes that cross through them. In the end, it is the electron density around all the protein atoms within a crystal that scatters X-rays. A mathematical operation called a Fourier Transform converts the electron density and diffraction patterns [3]. In order to create an electron density map, we need to know the amplitudes and phases of each reflection. As the square root of the spot intensity, the amplitude is calculated; however, the phase is unknown. A Fourier Transform cannot be applied to the reflections since phase information cannot be measured, so we cannot get the electron density by transforming them. This is known as the phase problem [3]. One way to solve this is called Molecular Replacement [3]. Once the correct orientation is found, we assume the model crystallized in the unknown crystal with this orientation and borrow the phase values generated from this model. Using the molecular replacement results and the experimentally observed amplitudes, a Fourier transformation can now be used to map the unknown structure.

One of the main advantages of crystallography is the structure determination of large biological molecules. X-ray crystallography is nowadays routinely used studies of the structures of biological molecules. X-ray crystallography also provides insight into ligand binding to its target. This allows for the design of new active substances.

1.4.2 Nuclear Magnetic Resonance (NMR)

Initially, solution biomolecular NMR focused on the determination of proteins, nucleic acids, and their complexes' 3D structure [22]. The main advantage of NMR is the ability to investigate dynamic molecular interactions (protein-protein, protein-ligand, or protein-nucleic acid) [23]. But the single most successful application of NMR is in studying intrinsically disordered proteins (IDPs) [23]. In general, protein structural determination by NMR consists of four main stages: Sample preparation of isotope-labeled protein, NMR data collection and analysis, focusing on the chemical shifts of hydrogen, nitrogen, and carbon atoms in the molecule, structural calculation and refinement using distance and/or orientation restraints, and finally, structural quality assessment [24]. From the late 1980s until 2005, NMR was used for the determination of structures ranging from 5 to 82 kDa. Two independent technical developments made larger systems studies possible: the TROSY technique [25], and residual dipolar couplings (RDCs) [22,26].

1.4.3 Cryo Electron Microscopy (Cryo-EM)

Cryo-EM is represents the observation of low-temperature specimens using an electron microscope [21]. This technique uses the same principle as transmission electron microscopy (TEM) but cools the samples to cryogenic temperatures and fixes them in a vitreous ice environment [3]. With improvements and precision of cryo-EM, sub-nanometer atomic resolutions (<4 Å) became a reality[22]. A simplified workflow of Cryo-EM begins with a sample preparation in aqueous solution. What follows is a vitrification – application of the protein sample solution to a grid-mesh and freezing in liquid ethane [3,23]. In this 'vitrified' sample, the 3D structure of the biomolecules is maintained. Proteins are then struck by the electron beam, which produces a faint image on the detector. After the distinction between a protein and the background, and grouping of the similar images, a high signal to noise 2D images is generated by a computer [3]. The models for 3D reconstructions are single particle analysis (SPA) and Sub-tomogram averaging (STA) [23]. This method aims to overcome the challenges of X-ray and NMR presented in structural heterogeneity and large size. On the other hand, high conformational flexibility and long disordered regions represent the main obstacles for the Cryo-EM [23].

1.5 Protein databases

1.5.1 Protein sequence databases

1.5.1.1 Pfam

Pfam is a curated database of protein families founded in 1995 [24]. Each of the entry is represented by multiple sequence alignments and hidden Markov models (HMMs) [25]. HMMs are probabilistic models used for the statistical inference of homology [26]. Entries which are related by similarity of sequence, structure or profile-HMM represent high-level groupings called clans. The newest release, Pfam version 35.0, contains 19,632 families and clans. Each entry is based on the UniProt Reference Proteomes [27]. Each family has a seed alignment that contains a representative set of sequences for the particular entry [22]. HMM is automatically built from the seed alignment and searched against a sequence database called *pfamseq* using the HMMER software [22,28]. All regions that satisfy a gathering threshold are aligned to the profile HMM to create the full alignment.

1.5.1.2 UniProt

Universal Protein Resource (UniProt) consortium was founded in 2002 by the joint forces of the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI), and the Protein Information Resource (PIR) [29]. The UniProt Knowledgebase is the highlight of UniProt and contains two sections: UniProtKB/Swiss-Prot, which consists of manually annotated entries, and UniProtKB/TrEMBL, which contains computer translations and annotations of CoDing Sequences (CDS) retrieved from the European Molecular Biology Laboratory (EMBL) [29]. There are several distinctive features of UniProtKB/Swiss-Prot. They include substantial manual annotation, minimal redundancy (separate protein entries for a species are merged into one entry), and many cross-links with other databases (such as sequence-related databases as well as specialized data collections) [29–31]. This computer-annotated database, UniProtKB/TrEMBL, contains translations of all CDS submitted to EMBL/GenBank/DNA Databank of Japan (DDBJ), which have not yet been integrated into UniProtKB/Swiss-Prot [29]. In addition to UniProtKB, the UniProt consists of UniProt Archive (UniParc), which contains all publicly available sequences from UniProtKB, RefSeq, Patent offices, and others. UniProt also includes UniProt Reference Clusters (UniRef), which consists of clusters of sequences sharing 100% identity for UniRef100, 90% for UniRef90, and 50% for UniRef50 [29].

1.5.2 Protein structure databases

1.5.2.1 Protein Data Bank (PDB)

Established in 1971, the Protein Data Bank (PDB) brought an evolution into the biological research world [32]. It was the first open-access digital source in this field. PDB is now an essential tool for bioengineers, biotechnologists, and researchers in biomedicine and biology. More than 175,000 experimentally determined structures of proteins, nucleic acids, and their complexes with one another and small molecules and drugs are stored in PDB. As of July 2022, PDB contains 167716 different proteins, mostly obtained using X-ray (147750), NMR (11977), and EM (7698) [33]. The management of PDB is in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles [34]. It is managed by the Worldwide Protein Data Bank organization (wwPDB), which includes the RCSB Protein Data Bank (RCSB PDB), the Protein Data Bank Japan (PDBj), the Protein Data Bank in Europe (PDBe), and BioMagResBank (BMRB) [35–37]. These four partners operate a global software system that supports data Deposition, Biocuration, and Validation of each new PDB entry [32]. In a recent review, Westbrook et al. reported that more than 90% of the new FDA-approved antineoplastic agents were facilitated by PDB [38]. More specifically, the publicly-available 3D structure helped understanding target biology and structure-guided lead optimization of the new agents [38].

1.5.2.2 CATH

CATH is a publicly accessible resource and provides a classification of protein domains based on structural data deposited into the wwPDB [39]. This classification was established in 1993. CATH clusters protein domain structures into sequence and structure similarity-dependent evolutionary families and structural groupings [40]. CATH uses sensitive structure-comparison and sequence-comparison tools (SSAP (3), HMMER3, PRC) to manually curate these remote evolutionary relationships [28,39,41,42]. All domains are classified into four hierarchical levels: Class (C), Architecture (A), Topology (T) and Homologous superfamilies (H). If proteins are having either significant sequence similarity (≥35% identity) or high structural similarity and some sequence similarity (≥20% identity), they are grouped into Homologous familes [40]. SSAP is used for assessment of the structural similarity. Moreover, to investigate features, such as the evolution of functional sites, superfamilies sub-classified into Functional Families (FunFams) using FunFHMMer protocol [10,39]. The scores of more distantly related folds are usually higher (Topology or fold level), though this may simply represent convergent evolution, further supporting the contention that nature has a limited number of folds [39,40]. If the protein folds 3D

arrangements of the secondary structures are similar, then these proteins are grouped in the Architecture level [40]. The Class, as the top level, displays the proportion of α-helix or β-strand secondary structures. Three major classes are recognized: mainly-α, mainly-β and α−β (since considerable overlap between the α+β and alternating α/β classes has been revealed by analyses) [40].

1.5.2.3 Structural Classification of Proteins (SCOP)

In 1994 MRC LMB and CPE in Cambridge established the Structural Classification of Proteins (SCOP) database, where know protein domains were organized according to their evolutionary and structural relationships [43]. SCOP's classification is built on two evolutionary levels: family and superfamily [44]. They are further classified into structural fold, although not necessarily based on evolutionary origin [44]. Protein domains that belong to the same family are related to their evolutionary origin, while domains belonging to the same superfamily are more distantly related [44]. Families and superfamilies both have domain boundaries to accommodate the fact that these relationships can span structural regions of different sizes [44]. Shared features of secondary structure composition in the domain core, architecture, and topology by the majority of the members, group superfamilies into folds. Even though fold is an attribute of a superfamily, some families belonging to exact superfamilies can belong to a different fold. Superfamilie*s* of proteins or protein regions that do not adopt globular folded structure are grouped in IUPRs (Intrinsically Unstructured Protein Region) [44]. The five structural classes are based on the secondary structure content of IUPRs and folds. As well as all-alpha and all-beta proteins that contain predominantly alpha-helices and beta-strands, there are also mixed alpha/beta proteins (a/b) and (a+b) that have alternate alpha-helices and beta-strands, and small proteins that contain little or no secondary structure [44]. IUPRs and folds are also categorized based on their type of protein, into four categories: soluble, membrane, fibrous, and intrinsically disordered [44]. SCOP is built as a classification of non-redundant protein domains. A prototype is selected based on sequence and structure and used for the classification. The sequence and the structure are obtained from UniProtKB and PDB, respectively [30,37]. Therefore, there are two boundaries assigned to the entry. This manual classification is then automatically extended using SIFTS [45]. SIFTS provides cross-references between protein sequences from UniProtKB on a residue-level, along with a three-dimensional models from PDB [45]. Families usually have unique sequence fingerprints defined by residues that are highly conserved at the interface between these domains. Domains within the protein can be structurally and functionally distant from one another, or distantly related to each other from functionally distinct proteins [44]. In contrast to the domain boundaries of the family, the domain boundaries of the superfamily span over individual domain boundaries. There is also

the case of a highly elaborated protein domain containing additional secondary structures [44]. However, this substructure has not been observed in other proteins and does not define an evolutionary conserved domain [44]. Again, the family domain identifies the entire region, including the substructure, whereas the superfamily domain identifies the smaller evolutionary conserved core. Further classification is purely structural: similar superfamilies without convincing evidence of a common evolutionary origin are grouped into *Folds*, which are then arranged into *Classes* based largely on the secondary structure content and organization [46]. To provide the ongoing updates and classification of new protein structures, after SCOP version 1 concluded in 2009, the authors have made the extended version of SCOP, SCOPe, which incorporates and updates the ASTRAL compendium [47,48]. For every SCOPe domain, as well as all PDB chains categorized in SCOPe, ASTRAL provides sequences and coordinates. Amino acids that have been chemically modified are translated back into their original sequences [49]. Crystallographically determined structures are assigned AEROSPACI scores, which offer a numeric estimation of their quality and precision [48]. In this manner, the highest quality representative is chosen in each subset.

SCOPe is fully backwards compatible with SCOP version 1.75 with considerably greater automation of the same hierarchical system. Nevertheless, in order to maintain the precise structure assignment, manual curation was reintroduced in 2014 [49].

2 BODY

2.1 Data mining and extraction

Starting with SCOPe version 2.01, data from all versions of SCOPe, SCOP, and Astral since 1.55 are stored in a relational database (https://scop.berkeley.edu). First, we downloaded the „dir.des.scope.txt“ file under Parseable Files & Software. Then, from the ASTRAL sequences and Subsets section, we downloaded and extracted all eight parts of the PDB-style files archive. We placed files in the "SCOPe" folder in the main directory, along with a bash script. The script reads the lines from the dir.des.scope.txt file.

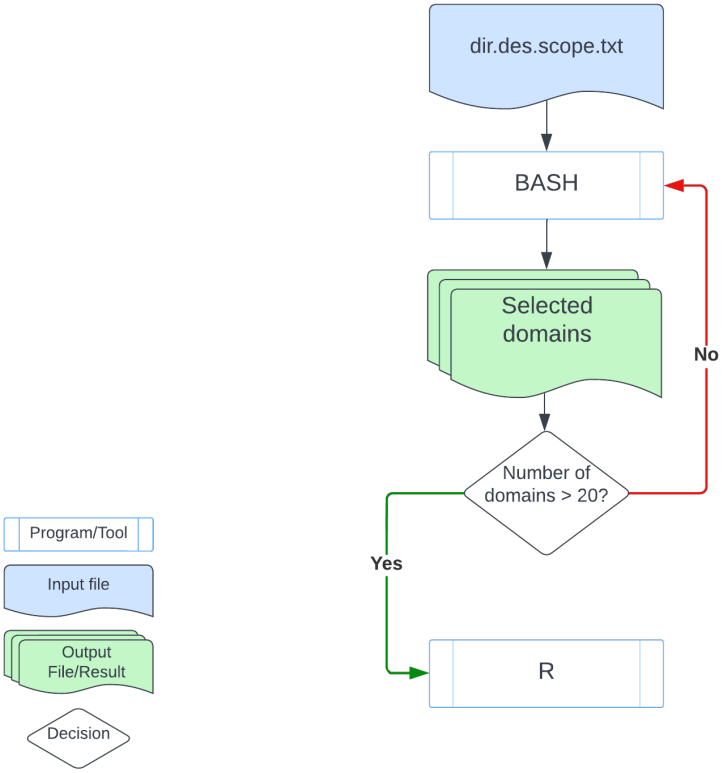

Figure 1: Workflow of the analysis which shows steps for choosing domains for the analysis in R

Following the reading of all protein families from the file, the script is finding all domains in SCOPe that are within the same family. To perform the analysis, it is necessary that over twenty domains are found present within one family, with the number being chosen arbitrarily as the threshold. Upon meeting this criterion, all domains are placed in a separate directory, in order to facilitate the use of the R script to perform the analysis. In cases where there are fewer than the set threshold of similar domains, the analysis of that ID is skipped.

2.2 Bio3D analysis

The R script includes the use of Bio3D [50] - a group of R packages that enables processing, organization, and analysis of biomolecular structures [51]. It offers search interfaces for major bioinformatics databases, sequence and structure conservation analysis, as well as popular computational methods for analyzing and predicting protein structure dynamics, like PCA, correlation network analysis (CNA), normal mode analysis (NMA), and new ensemble difference distance matrix (eDDM) analysis [51–53]. The Bio3D core package provides functions for data input and output, format conversion and data manipulation.

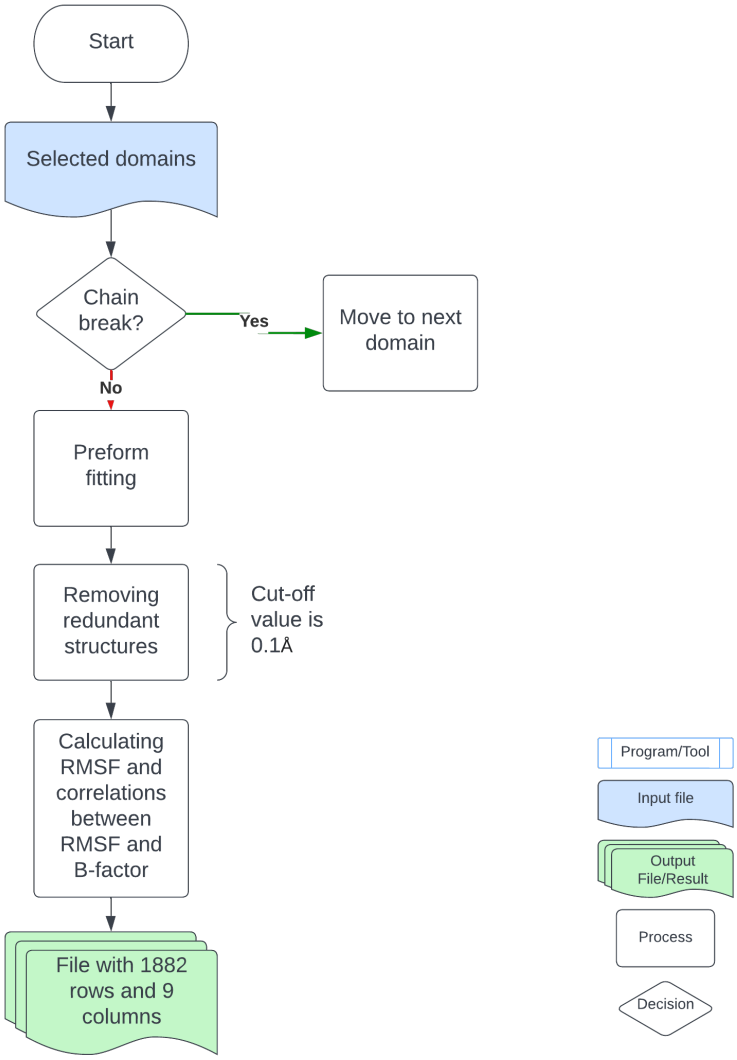

Figure 2: Workflow of the analysis in R with the Bio3D package

The package provides the following steps which were used for the analysis, which can be seen in Figure 2:

1. Finding possible chain breaks based on connective Calpha atoms with the use of chain.break() function [51], where we chose to keep only the protein domains with no chain breaks.
2. Multiple alignment: There are numerous options for performing multiple alignment, including the calling of external programs, such as MUSCLE [54], using the MUSCLE algorithm internally as implemented in the Bioconductor "msa" package [55]. These methods can be implemented using the pdbaln() and seqaln() functions [51], of which we chose pdbaln().
3. Structure fitting and analysis: In this step, all aligned structures are fitted using the pdbfit() function [51]. The fitting is based on their invariant structural core as identified by the core.find() function [51]. At this stage, individual residue fluctuations (RMSF) and structural deviations like RMSD can be performed.
After that, conformational redundant structures were removed.
Finally, in the output file will be nine columns, eight of which are of interest, containing the following: name of the file which contains SCOP ID, number of "equal/similar" domains in SCOPe, domain size and the number of residues, mean RMSF, standard deviation of RMSF, mean correlation between RMSF and temperature factors, minimal correlation between RMSF and temperature factors, and maximal correlation between RMSF and temperature factors.
Superimposed molecules that were selected for visualization were visualized and presented using Visual Molecular Dynamics (VMD) software [56].
Statistical analysis was performed using R statistical software. The normality of the data distribution was determined using Kolmogorov – Smirnov test. Pearson correlation coefficient was used for the investigation of correlations between each of the following features of the entries: number of equal/similar entries, number of residues, mean RMSF, and mean RMSF, and temperature factor correlation. In all analyses, the significant level of difference was considered to be $p \leq 0.05$.

3 RESULTS AND DISCUSSION

The ability to adapt to changes, more specifically flexibility, enables processes like enzyme catalysis, antigen recognition, and protein transport. On the other hand, protein flexibility has an important place in drug discovery [57,58]. In their thorough review of protein flexibility, Teilum et al noted that protein flexibility is a consequence of their dynamics, yet the result of the dynamics is not always flexibility [59].

In our study, using the data from the SCOPe database, we wanted to explore the aspects of protein domain flexibility using RMSF and temperature factors as measures. Moreover, we wanted to explore the SCOPe entries themselves, investigating the number of similar domain entries and the domain size (expressed in the number of residues).

Our analysis revealed that 120 domains had 100 or more replicates in the SCOPe database. Three domains with the most equal/similar entries in the SCOPe database were d1a30a_, d1t2ja1, and d1x0ja_ with 757, 747, and 745 entries, respectively, shown in Table 1. Besides the number of equal/similar entries in the SCOPe database, Table 1 contains the number of residues, mean RMSF, a standard deviation of RMSF, and the correlation between RMSF and temperature factors for the three domains with the highest number of equal/similar entries. Even though we used a Pearson correlation coefficient to investigate the potential relations between mentioned features, there were no significant correlations between a higher number of entries and the mentioned features. Finally, we could not observe any logical template of changes in mean RMSF, a standard deviation of RMSF, and correlation between RMSF and temperature factors, as the number of equal/similar entries changed.

Table 1: Domains with the highest number of equal/similar entries

SCOPe ID	No. Of similar/equal entries	No. Of residues	Mean RMSF	Standard Deviation of RMSF	Mean RMSF and temperature factors correlation
d1a30a_	757	99	0.448	0.293	0.365
d1t2ja1	747	103	2.006	1.509	0.033
d1x0ja_	745	106	1.906	1.615	0.274

The domain d1a30a_ belongs to all beta proteins class, acid proteases superfamily, and retroviral protease family. This is the domain of Human immunodeficiency virus type 1 protease. This enzyme has a vital role in viral replication, but more importantly, this is one of the key drug targets for the treatment of HIV infection [60]. This is an aspartic protease that exists as a homodimer. In each monomer, there are 99 amino acids and a catalytic center composed of D25, T26, and G27, where both the D25 aspartic acids are involved in catalysis [60]. This protein is visually presented in Figure 3. Some domains were investigated more

often since their documented role in various physiological and pathophysiological processes. Moreover, as we have mentioned in the Introduction, some domains were used as a base for new drug development, especially in oncology, infectious diseases, and metabolic diseases. Even though the domain has a big number of similar/equal domains, a noticeable characteristic is that the mean RMSF value is lower. In Figure 3, we can see a stable core, with flexible side chains of the protein domains (left). On the right figure, the flexibility of the side chain is confirmed with a higher temperature factor, colored in red, while the more stable region of the structure is colored in blue.

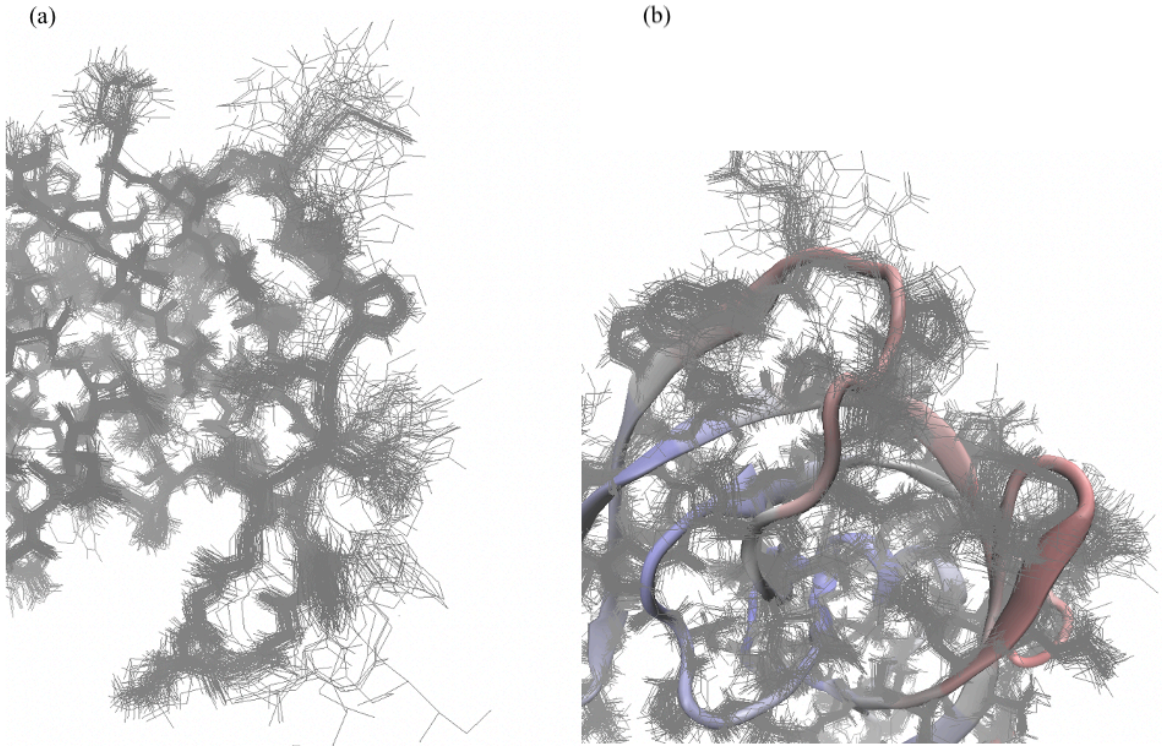

Figure 3: (a) Visual representation of the superimposed domain d1a30a_; (b) B-factor presented on the superimposed structure, where red color is representing a higher value of the B-factor and blue is representing smaller values of the B-factor

d1t2ja1 and d1x0ja_ with 747 and 745 similar entries respectively are domains of molecules that are marked in the SCOPe database as “not true proteins”. d1t2ja1 belongs to the all-beta proteins class with immunoglobulin-like beta-sandwich fold, with 7 strands in 2 sheets. This molecule is a part of the immunoglobulin superfamily and V set domains (antibody variable domain-like) family. On the other hand, d1x0ja_ belongs to all alpha proteins with bromodomain-like fold, with four helices, and the bromodomain superfamily.

Five domains with the highest mean RMSF were d3mrwa_, d3h1ha1, d2lnia1, d1x1ea, and d5hgxa1 (Table 2). Table 2 contains the number of similar entries, number of residues, mean RMSF, a standard deviation of RMSF, and the correlation between RMSF and temperature factors. Pearson correlation coefficient showed no significant correlations between a higher RMSF and the mentioned features.

Four of the mentioned domains (d3mrwa_, d3h1ha1, d1x1ea, and d5hgxa1) belong to the alpha and beta proteins (the first being a+b while the latter two a/b).

Table 2: Domains with the highest mean RMSF

SCOPe ID	No. Of similar/equal entries	No. Of residues	Mean RMSF	Standard Deviation of RMSF	Mean RMSF and temperature factors correlation
d3mrwa_	103	240	9.987	5.397	0.114
d3h1ha1	23	106	9.936	5.079	0.257
d2lnia1	14	111	9.741	7.048	0.393
d1x1ea	12	155	9.732	5.443	0.130
d5hgxa1	10	100	9.709	5.784	0.294

d3mrwa_ belongs to the Ribosome inactivation proteins (RIP) superfamily and Plant cytotoxins family. These molecules irreversibly inhibit the synthesis of protein by the removal of adenine residues from ribosomal RNA (rRNA) [61]. Single-chain RIPs of type I have an approximate molecular weight of 30 kDa, while RIPs of type II have an approximate molecular weight of 56–65 kDa and contain an enzymatically active A-chain and a slightly larger B chain (lectin subunit) that is specifically reactive to sugars exhibiting galactose-like structures [61]. This protein is presented in Figure 4.

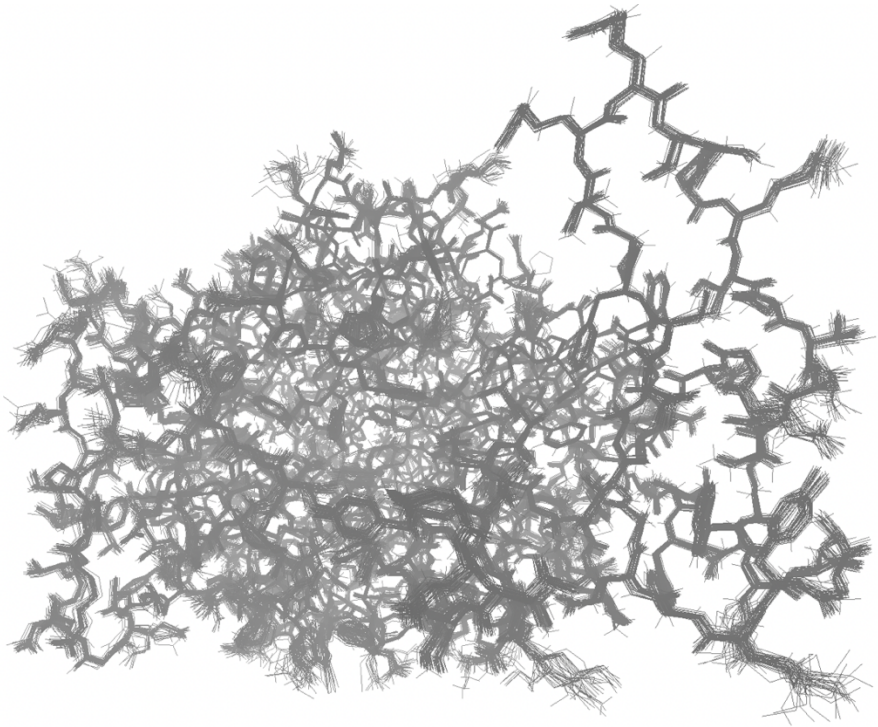

Figure 4: Visual representation of the superimposed domain d3mrwa_

d3h1ha1 is a domain of cytochrome bc1 complex from chicken, that belong to the LuxS/MMP-like metallohydrolase superfamily.
d1x1ea is a domain found in the Crystal Structure of TT0495 protein from Thermus thermophilus HB8, which belongs to the NAD(P)-binding Rossman-fold domains.
d5hgxa1 is a domain found in the crystal structure of transketolase mutant - h261f from *Pichia stipitis*. This molecule is a part of the thiamin diphosphate-binding fold with three layers: a/b/a, with a parallel beta-sheet of 6 strands.
d2lnia1 domain belong to all alpha proteins class and alpha-alpha superhelix fold of TPR-like superfamily. The molecule is a solution NMR Structure of Stress-induced-phosphoprotein 1 (STI1) from Homo sapiens.
d5bsea2 is a domain of the molecule with the lowest RMSF (0.090, Figure 5). This is domain of a molecule that is not a true protein, nor does it belong to a true family. The superfamily of this domain is 6-phosphogluconate dehydrogenase C-terminal domain-like.

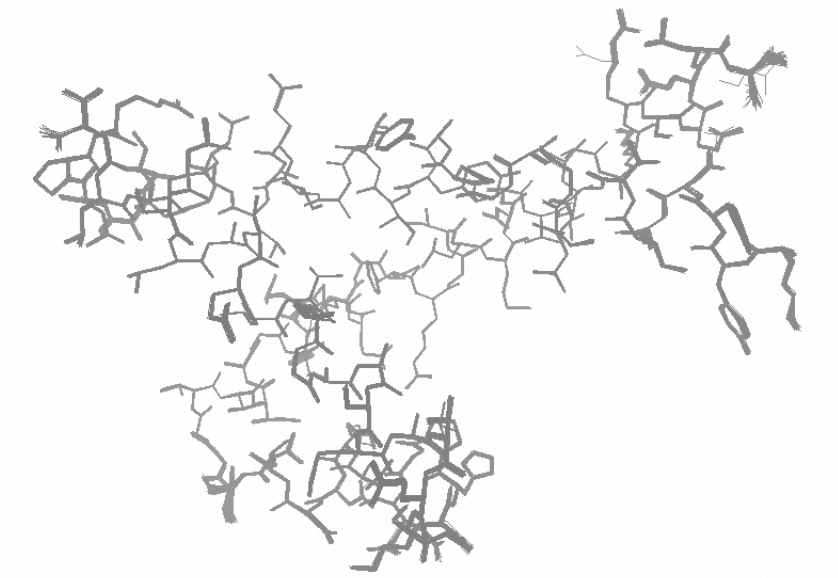

Figure 5: Visual representation of the superimposed domain with the lowest RMSF - d5bsea2

As it is presented in Table 2, the mean correlation between RMSF and temperature factors for these domains ranged from very weak to weak (0,114 for d3mrwa and 0,393 for d2lnia1). Table 3 contains the number of similar entries, number of residues, mean RMSF, a standard deviation of RMSF, and the correlation between RMSF and temperature factors for the five entries with the highest number of amino acid residues. Mean correlation between RMSF was moderate for d5faua_ and d1ti2a2. For the rest of the entries in Table 3, correlation ranged from negligible to low. Pearson correlation coefficient showed no significant correlations between a higher number of residues and the mentioned features. We could not observe any template of features change and the higher/lower number of residues.

d1axra_ is the domain of the glycogen phosphorylase, an a/b protein that belongs to the UDP-Glycosyltransferase/glycogen phosphorylase superfamily and oligosaccharide phosphorylase family. Our analysis marked this protein as the one with the highest number of residues (830, Table 3, Figure 6). Glycogen phosphorylase consists of two non-similar domains with three layers (a/b/a) each: Domain 1 has parallel beta-sheet of7 strands, while domain 2 has a parallel beta-sheet of 6 strands. Three separate genes, PYGM, PYGL, and PYGB encode three isoforms of this protein, which are found in muscle, liver, and brain, respectively.

The visual representation of the superimposed structures is represented in Figure 6. We can observe a smaller amount of side chains and, at the same time, a lower mean RMSF value, regardless of the number of residues. The stable region of the structure implies low fluctuations, which are supported by the low B-factor, colored in blue.

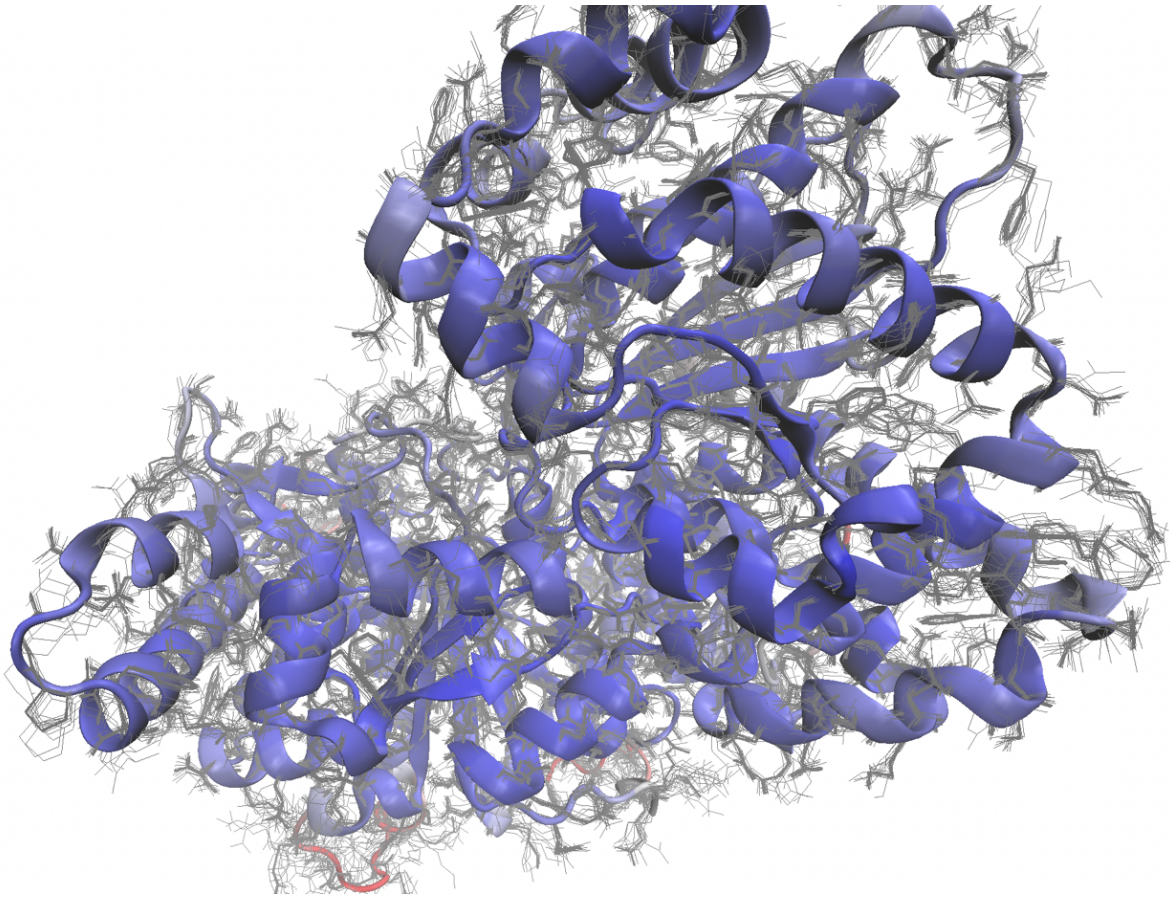

Figure 6: Visual representation of the superimposed domain d1axra_

Among domains belonging to proteins with the highest number of amino-acid residues, d1axra_ is followed by d5faua_, d1ti2a2, d1f8na1, and d1fiqc2 with 794, 728,690, and 625 residues, respectively.

Table 3: Domains with the highest number of amino-acid residues

SCOPe ID	No. Of similar/equal entries	No. Of residues	Mean RMSF	Standard Deviation of RMSF	Mean RMSF and temperature factors correlation
d1axra_	16	830	0.517	1.00	0.246
d5faua_	20	794	0.166	0.08	0.588
d1ti2a2	30	728	0.117	0.04	0.516
d1f8na1	12	690	0.155	0.102	0.394
d1fiqc2	12	625	0.196	0.428	0.264

d5faua_ belongs to alpha and beta proteins with PFL-like glycyl radical enzymes fold and Superfamily of the same name.
d1tu2a2 is a domain of alpha and beta protein with fumarate dehydrogenase/DMSO reductase domains 1-3-fold. This fold contains two similar intertwined domains related by a pseudo dyad. The domain is a part of the transhydroxylase alpha subunit, AthL protein.

d1f8na1 is a C-terminal domain of lipoxygenase, an all-alpha protein from the lipoxygenase (LOX) superfamily and family. This family catalyzes the oxidation of polyunsaturated fatty acids (PUFA) to produce hydroperoxides [62]. Members of this protein family can be found in animals and plants, as well as cyanobacteria [63]. Human lipoxygenases stimulate inflammatory reactions by stimulating lipid oxidation [64]. Reactive oxygen species can trigger inflammation by stimulating the release of cytokines and the activation of LOXs [64]. Several diseases are linked to inflammation, including cancer, strokes, cardiovascular disease, and neurodegenerative disease. The LOX enzymes are involved in the synthesis of prostaglandins and leukotrienes [64]. Their inhibition may play a crucial role in preventing diseases due to their association with disease development.
Finally, d1fiqc2 is a C-terminal domain of xanthine oxidase, an alpha and beta protein with molybdenum cofactor-binding domain fold, and a member of family and superfamily of the same name. XO catalyses the oxidation of hypoxanthine to xanthine and subsequently to uric acid [65]. Higher concentrations of uric acid are connected to the higher levels of xanthine oxidase activity and to the oxidative stress, an essential feature for many vascular diseases [65,66]. The overactivity of this enzyme also results in a condition known as gout, an acute inflammatory arthritis [65,67].
Table 4 presents the entries with the highest mean correlation between RMSF and temperature factors, as well as number of similar/equal entries, number of residues, mean RMSF, mean RMSF, Mean RMSF and temperature factors correlation, minimum and maximum RMSF and temperature factors correlation. For all these entries, the minimum and maximum correlation between RMSF and temperature factors ranged from moderate, high to very high. Only d4gefa_ had minimum correlation that was low (0.368).

Pearson correlation coefficient showed no significant correlations between a higher mean correlation between RMSF and temperature factors and number of similar entries, number of residues and the minimum and maximum RMSF and temperature factors correlation. We did not observe any template of features change and the higher/lower mean correlation between RMSF and temperature factors besides mean RMSF

Table 4: Domains with the highest mean RMSF and temperature factor correlation

SCOPe ID	No. Of similar/equal entries	No. Of residues	Mean RMSF	Mean RMSF and temperature factors correlation	Minimum RMSF and temperature factors correlation	Maximum RMSF and temperature factors correlation
d5mkna_	19	71	0.292	0.839	0.771	0.909
d3kjja1	24	116	0.154	0.792	0.526	0.917
d4xdca2	32	83	0.140	0.776	0.590	0.826
d4gefa_	10	157	0.129	0.773	0.368	0.863
d1ch6e1	22	203	0.570	0.773	0.665	0.810

Among the five domains of proteins with the highest correlation between RMSF and temperature factors, four are not a part of true proteins (d5mkna_, d3kjja1, d4xdca2, and d4gefa_). Moreover, the first three mentioned domains do not belong to a true family.
d5mkna_ is a domain of the molecule from all beta protein classes with an Sm-like fold and a part of the Sm-like ribonucleoproteins superfamily (Figure 7).

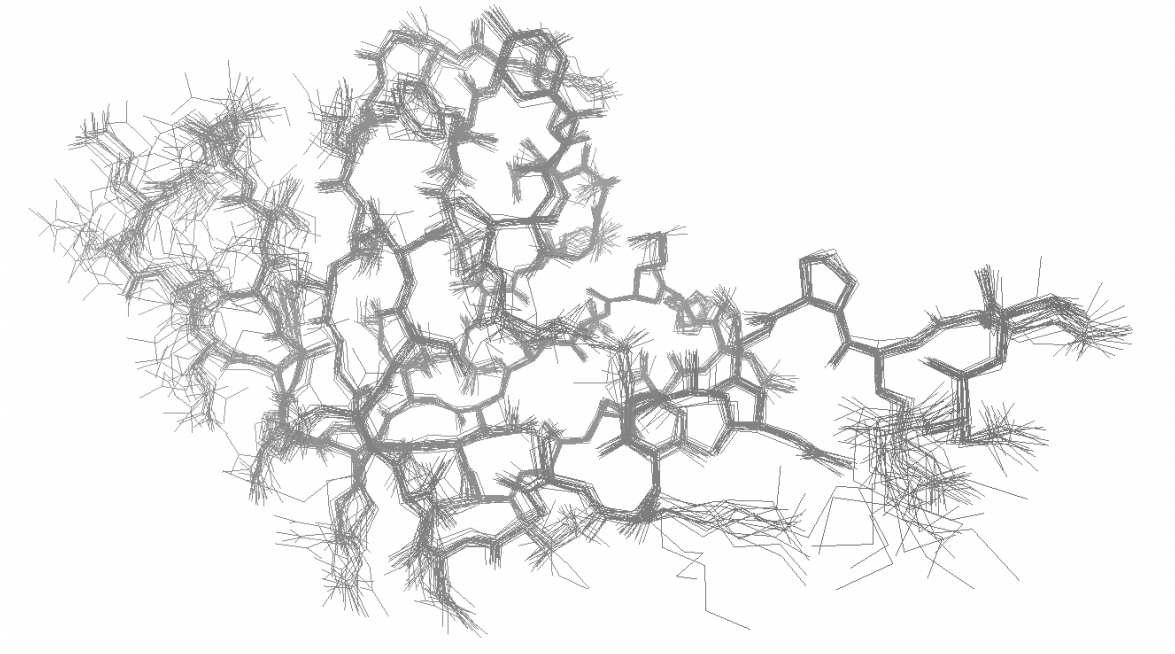

Figure 7: Visual representation of the superimposed domain d5mkna_

d1ch6e1 is a domain of glutamate dehydrogenase, an alpha and beta protein from the amino acid dehydrogenase family. Glutamate dehydrogenase catalyzes the oxidative deamination of L-glutamate to α-KG using NAD(P)+ as a coenzyme [68]. GTP and ATP are inhibitors of the reaction, increasing the binding affinity for the product, and therefore decreasing enzymatic activity [68]. The activity of this enzyme was proven to be important in insulin metabolism [68]. Moreover, GDH has an important role in neural development and tumor growth [68,69]. Namely, up-regulation of human glutamate dehydrogenase expression permits neoplastic cells to utilize glutamine and/or glutamate for their growth [69].

Our results showed that there was no correlation between the number of equal/similar entries per domain and the mean RMSF of the same domain (Table 5). Likewise, the number of equal/similar entries did not correlate with the correlation between RMSF and temperature factors. This could highlight the general evolutionary conservation of protein flexibility. Analysis of the large datasets of homologous proteins showed that backbone flexibility (measured by B-factor) diverges slowly, and it is highly conserved at the family and superfamily levels [70]. Studies focused on local flexibility involving small-scale conformational changes showed that homologous proteins share a similar pattern of flexibility [12,71]. A recent study has shown that homologous proteins display highly conserved conformational changes across a broad range of evolutionary distances [72]. Moreover, there was no correlation between the number of residues and the mean RMSF. Likewise, the number of residues did not correlate with the correlation between RMSF and temperature factors.

Higher mean RMSF negatively correlated with mean RMSF and temperature factor correlation (Pearson correlation coefficient -0.390; 95% CI -0.429 - -0.351; p-value = 2.2×10^{-16}, Table 5). This finding could highlight that temperature (B) factors may not be the most useful and exact measure of the dynamics of proteins/domains with higher fluctuation (higher RMSF). This needs further investigation.

Table 5: Correlation between different entries features with the correlation coefficient and p value in the bracket.

Features of the entries	**Number of amino-acid residues**	**Mean RMSF**	**Mean RMSF and temperature correlation**
Number of equal/similar entries	-0.044 (0.051)	0.030 (0.192)	-0.031 (0.169)
Number of amino-acid residues		0.030 (0.192)	0.030 (0.192)
Mean RMSF			-0.390 (2.2×10^{-16})*

* - indicates statistically significant correlation (Pearson correlation coefficient)

4 CONCLUSION

The Protein Data Bank and the SCOPe database are updated daily. The SCOPe database contains valuable information about protein domains. Domains are classified based on their 3D shape. Usually, a single protein domain contains multiple entries. These entries are the result of multiple experiments performed under different conditions and for different purposes. As shown in this study, multiple entries can be used to analyse the differences between entries. In other words, we can extract protein flexibility data. We should be aware that protein flexibility can be studied in many ways, such as molecular dynamics, normal mode analysis, or the study of temperature factors listed in the PDB. Here we present an alternative method to study protein dynamics by comparing multiple entries in the SCOPe database. Because proteins are flexible macromolecules, models from different experiments do not match exactly. The differences have many causes, such as experimental errors, the use of different software, and also because proteins are flexible. When the SCOPe database contains multiple entries for a particular domain, direct comparison of the domains provides information about the flexibility of the proteins. We must be aware that when a model is solved by X-ray crystallography, the data are recorded in the crystals at low temperature. Therefore, the flexibility calculated from the SCOPe database is more of a lower limit, and we might expect the protein to have greater flexibility in aqueous solution at room temperature than observed in the protein crystals. None of the above methods for studying protein flexibility is very accurate, so the development of new methods and approaches is necessary. If different methods give similar results, we can be more confident that our conclusions are closer to the true solution. In the near future, we can expect even more rapid growth of the PDB and SCOPe database, which will provide new opportunities to develop and improve methods for studying protein flexibility.

5 DALJŠI POVZETEK V SLOVENSKEM JEZIKU

Podatkovna zbirka proteinov (Protein Data Bank) in podatkovna zbirka proteinskih domen (SCOPe) se dnevno posodabljata. Podatkovna zbirka SCOPe vsebuje uporabne informacije o beljakovinskih domenah. Domene so razvrščene na podlagi njihove 3D-oblike. Običajno ena beljakovinska domena vsebuje več vnosov. Ti vnosi so rezultat več poskusov, izvedenih v različnih pogojih in za različne namene. Kot je prikazano v tej študiji, se lahko več vnosov uporabi za analizo razlik med vnosi. Z drugimi besedami, pridobimo lahko podatke o prožnosti oziroma konformacijski fleksibilnosti beljakovin. Zavedati se moramo, da lahko prožnost proteinov preučujemo na različne načine, na primer z molekulsko dinamiko, analizo normalnih modov ali preučevanjem temperaturnih faktorjev, ki so navedenih v podatkovni zbirki proteinov. V tej študiji predstavljamo alternativno metodo za preučevanje dinamike proteinov s primerjavo več vnosov v podatkovni zbirki SCOPe. Ker so proteini prožne makromolekule, se modeli iz različnih poskusov ne ujemajo popolnoma. Za prisotne razlike lahko najdemo številne vzroke, kot so eksperimentalne napake, uporaba različne programske opreme in tudi zato, ker so proteini prožni. Kadar podatkovna zbirka SCOPe vsebuje več vnosov za določeno domeno, potem je z neposredno primerjavo možno dobiti informacije o prožnosti proteinov. Zavedati se moramo, da se pri reševanju modela z rentgensko kristalografijo podatki beležijo v kristalih pri nizki temperaturi. Zato je prožnost, izračunana iz podatkovne zbirke SCOPe, spodnja meja prožnosti in lahko pričakujemo, da ima beljakovina v vodni raztopini pri sobni temperaturi večjo prožnost, kot je bila opažena v samih kristalih. Nobena od zgornjih metod za preučevanje prožnosti proteinov ni zelo natančna, zato je potreben razvoj novih metod in pristopov. Če različne metode dajejo podobne rezultate, smo lahko bolj prepričani, da so naše ugotovitve bližje pravi rešitvi. V bližnji prihodnosti lahko pričakujemo še hitrejšo rast podatkovnih zbirk PDB in SCOPe, kar bo zagotovilo nove priložnosti za razvoj in izboljšanje metod za preučevanje prožnosti proteinov.

Naša analiza je pokazala, da beljakovinska domena, d1a30a_, z največ vnosi pripada vsem razredom beta proteinov, naddružini kislih proteaz in družini retrovirusnih proteaz. To je področje proteaze virusa humane imunske pomanjkljivosti tipa 1. Nekatera področja so bila preiskana pogosteje, ker je bila dokumentirana njihova vloga v različnih fizioloških in patofizioloških procesih. Poleg tega so bila nekatera področja uporabljena kot osnova za razvoj novih zdravil, zlasti v onkologiji, nalezljivih boleznih in presnovnih boleznih. Čeprav ima ta domena veliko število podobnih/enakih domen, je opazna značilnost, da je srednja vrednost RMSF nižja. Proteinska domena, ki ima najvišjo srednjo vrednost RMSF, spada v superdružino ribosomskih inaktivacijskih proteinov (RIP) in družino rastlinskih citotoksinov. d1axra_ je domena glikogen fosforilaze, a/b proteina, ki spada v superdružino UDP-glikoziltransferaze/glikogen fosforilaze in družino oligosaharidnih fosforilaz. Naša analiza je ta protein označila kot tistega z največjim številom ostankov. Domena proteinov z najvišjo korelacijo med RMSF in temperaturnimi faktorji, d5mkna_, je razvrščena kot neprava beljakovina. Ugotovili smo, da konformacijska prožnost ni povezana z velikostjo domene, in prav tako ni povezan z samim zvitjem proteina. Poleg tega je RMSF v nekaterih primerih zelo dobro povezan s temperaturnim faktorjem. Po drugi strani pa najdemo številne primere, kjer obstaja negativna korelacija med RMSF in temperaturnim faktorjem.

6 REFERENCES

[1] Williamson M. How Proteins Work. 1st ed. Garland Science; 2012. https://doi.org/10.1201/9781136665493.
[2] Sanvictores T, Farci F. Biochemistry, Primary Protein Structure. StatPearls, Treasure Island (FL): StatPearls Publishing; 2022.
[3] Stollar EJ, Smith DP. Uncovering protein structure. Essays in Biochemistry 2020;64:649–80. https://doi.org/10.1042/EBC20190042.
[4] Bayse CA, Pollard DB. Conformation dynamics of cyclic disulfides an d selenosulfides in CXXC(U) (X = Gly, Ala) tetrapeptide redox motifs. Journal of Peptide Science 2019;25:e3160. https://doi.org/10.1002/psc.3160.
[5] Rehman I, Farooq M, Botelho S. Biochemistry, Secondary Protein Structure. StatPearls, Treasure Island (FL): StatPearls Publishing; 2022.
[6] Rehman I, Kerndt CC, Botelho S. Biochemistry, Tertiary Protein Structure. StatPearls, Treasure Island (FL): StatPearls Publishing; 2022.
[7] Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Briefings in Bioinformatics 2009;10:205–16. https://doi.org/10.1093/bib/bbn057.
[8] Wang Y, Zhang H, Zhong H, Xue Z. Protein domain identification methods and online resources. Computational and Structural Biotechnology Journal 2021;19:1145–53. https://doi.org/10.1016/j.csbj.2021.01.041.
[9] Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Research 2000;28:257–9. https://doi.org/10.1093/nar/28.1.257.
[10] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Research 2021;49:D266–73. https://doi.org/10.1093/nar/gkaa1079.
[11] Kmiecik S, Kouza M, Badaczewska-Dawid AE, Kloczkowski A, Kolinski A. Modeling of Protein Structural Flexibility and Large-Scale Dynamics: Coarse-Grained Simulations and Elastic Network Models. International Journal of Molecular Sciences 2018;19:3496. https://doi.org/10.3390/ijms19113496.
[12] Fuglebakk E, Echave J, Reuter N. Measuring and comparing structural fluctuation patterns in large protein datasets. Bioinformatics 2012;28:2431–40. https://doi.org/10.1093/bioinformatics/bts445.
[13] Martínez L. Automatic Identification of Mobile and Rigid Substructures in Molecular Dynamics Simulations and Fractional Structural Fluctuation Analysis. PLOS ONE 2015;10:e0119264. https://doi.org/10.1371/journal.pone.0119264.
[14] Carugo O. How large B-factors can be in protein crystal structures. BMC Bioinformatics 2018;19:61. https://doi.org/10.1186/s12859-018-2083-8.
[15] Sun Z, Liu Q, Qu G, Feng Y, Reetz MT. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. Chem Rev 2019;119:1626–65. https://doi.org/10.1021/acs.chemrev.8b00290.
[16] Carugo O, Argos P. Correlation between side chain mobility and conformation in protein structures. Protein Engineering, Design and Selection 1997;10:777–87. https://doi.org/10.1093/protein/10.7.777.

[17] Carugo O, Argos P. Reliability of atomic displacement parameters in protein crystal structures. Acta Cryst D 1999;55:473–8. https://doi.org/10.1107/S0907444998011688.
[18] Pang Y-P. Use of multiple picosecond high-mass molecular dynamics simulations to predict crystallographic B-factors of folded globular proteins. Heliyon 2016;2. https://doi.org/10.1016/j.heliyon.2016.e00161.
[19] Carugo O. B-factor accuracy in protein crystal structures. Acta Cryst D 2022;78:69–74. https://doi.org/10.1107/S2059798321011736.
[20] Pearce NM, Gros P. A method for intuitively extracting macromolecular dynamics from structural disorder. Nat Commun 2021;12:5493. https://doi.org/10.1038/s41467-021-25814-x.
[21] Savva C. A beginner's guide to cryogenic electron microscopy. The Biochemist 2019;41:46–52. https://doi.org/10.1042/BIO04102046.
[22] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research 2021;49:D412–9. https://doi.org/10.1093/nar/gkaa913.
[23] Nwanochie E, Uversky VN. Structure Determination by Single-Particle Cryo-Electron Microscopy: Only the Sky (and Intrinsic Disorder) is the Limit. International Journal of Molecular Sciences 2019;20:4186. https://doi.org/10.3390/ijms20174186.
[24] Sonnhammer ELL, Eddy SR, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins: Structure, Function, and Bioinformatics 1997;28:405–20. https://doi.org/10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L.
[25] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Research 2014;42:D222–30. https://doi.org/10.1093/nar/gkt1223.
[26] Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. Journal of Molecular Biology 1994;235:1501–31. https://doi.org/10.1006/jmbi.1994.1104.
[27] What are reference proteomes? | UniProt help | UniProt n.d. https://www.uniprot.org/help/reference_proteome (accessed July 10, 2022).
[28] HMMER n.d. http://hmmer.org/ (accessed July 10, 2022).
[29] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols, Totowa, NJ: Humana Press; 2007, p. 89–112. https://doi.org/10.1007/978-1-59745-535-0_4.
[30] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
[31] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 2021;49:D480–9. https://doi.org/10.1093/nar/gkaa1100.
[32] Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. In: Wlodawer A, Dauter Z, Jaskolski M, editors. Protein Crystallography: Methods and Protocols, New York, NY: Springer; 2017, p. 627–41. https://doi.org/10.1007/978-1-4939-7000-1_26.
[33] Bank RPD. PDB Statistics: PDB Data Distribution by Experimental Method and Molecular Type n.d. https://www.rcsb.org/stats/summary (accessed July 9, 2022).
[34] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[35] Burley SK. Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. Journal of Biological Chemistry 2021;296. https://doi.org/10.1016/j.jbc.2021.100559.
[36] Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Mol Biol 2003;10:980–980. https://doi.org/10.1038/nsb1203-980.
[37] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research 2019;47:D520–8. https://doi.org/10.1093/nar/gky949.
[38] Westbrook JD, Soskind R, Hudson BP, Burley SK. Impact of the Protein Data Bank on antineoplastic approvals. Drug Discovery Today 2020;25:837–50. https://doi.org/10.1016/j.drudis.2020.02.002.
[39] Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Research 2019;47:D280–4. https://doi.org/10.1093/nar/gky1097.
[40] Orengo CA, Pearl FMG, Bray JE, Todd AE, Martin AC, Lo Conte L, et al. The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Research 1999;27:275–9. https://doi.org/10.1093/nar/27.1.275.
[41] Orengo CA, Taylor WR. [36] SSAP: Sequential structure alignment program for protein structure comparison. Methods in Enzymology, vol. 266, Academic Press; 1996, p. 617–35. https://doi.org/10.1016/S0076-6879(96)66038-8.
[42] Brandt BW, Heringa J. webPRC: the Profile Comparer for alignment-based searching of public domain databases. Nucleic Acids Research 2009;37:W48–52. https://doi.org/10.1093/nar/gkp279.
[43] Hubbard TJP, Murzin AG, Brenner SE, Chothia C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Research 1997;25:236–9. https://doi.org/10.1093/nar/25.1.236.
[44] Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Research 2020;48:D376–82. https://doi.org/10.1093/nar/gkz1064.
[45] Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Research 2019;47:D482–9. https://doi.org/10.1093/nar/gky1114.
[46] Chandonia J-M, Guan L, Lin S, Yu C, Fox NK, Brenner SE. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. Nucleic Acids Research 2022;50:D553–9. https://doi.org/10.1093/nar/gkab1054.
[47] Chandonia J-M, Fox NK, Brenner SE. SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins – extended Database. Journal of Molecular Biology 2017;429:348–55. https://doi.org/10.1016/j.jmb.2016.11.023.
[48] Chandonia J, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, et al. The ASTRAL Compendium in 2004. Nucleic Acids Research 2004;32:D189–92. https://doi.org/10.1093/nar/gkh034.
[49] Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research 2014;42:D304–9. https://doi.org/10.1093/nar/gkt1240.

[50] Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 2006;22:2695–6. https://doi.org/10.1093/bioinformatics/btl461.
[51] Grant BJ, Skjærven L, Yao X-Q. The Bio3D packages for structural bioinformatics. Protein Science 2021;30:20–30. https://doi.org/10.1002/pro.3923.
[52] Sethi A, Eargle J, Black AA, Luthey-Schulten Z. Dynamical networks in tRNA:protein complexes. Proceedings of the National Academy of Sciences 2009;106:6620–5. https://doi.org/10.1073/pnas.0810961106.
[53] Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. Biophysical Journal 2001;80:505–15. https://doi.org/10.1016/S0006-3495(01)76033-X.
[54] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004;32:1792–7. https://doi.org/10.1093/nar/gkh340.
[55] Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. Bioinformatics 2015;31:3997–9. https://doi.org/10.1093/bioinformatics/btv494.
[56] Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. Journal of Molecular Graphics 1996;14:33–8. https://doi.org/10.1016/0263-7855(96)00018-5.
[57] Antunes DA, Devaurs D, Kavraki LE. Understanding the challenges of protein flexibility in drug design. Expert Opinion on Drug Discovery 2015;10:1301–13. https://doi.org/10.1517/17460441.2015.1094458.
[58] Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov 2003;2:527–41. https://doi.org/10.1038/nrd1129.
[59] Teilum K, Olsen JG, Kragelund BB. Functional aspects of protein flexibility. Cell Mol Life Sci 2009;66:2231. https://doi.org/10.1007/s00018-009-0014-6.
[60] Maximova K, Wojtczak J, Trylska J. Enzymatic activity of human immunodeficiency virus type 1 protease in crowded solutions. Eur Biophys J 2019;48:685–9. https://doi.org/10.1007/s00249-019-01392-1.
[61] Puri M, Kaur I, Perugini MA, Gupta RC. Ribosome-inactivating proteins: current status and biomedical applications. Drug Discovery Today 2012;17:774–83. https://doi.org/10.1016/j.drudis.2012.03.007.
[62] Newcomer ME, Brash AR. The structural basis for specificity in lipoxygenase catalysis. Protein Science 2015;24:298–309. https://doi.org/10.1002/pro.2626.
[63] Chedea VS, Jisaka M. Lipoxygenase and carotenoids: A co-oxidation story. African Journal of Biotechnology 2013;12. https://doi.org/10.4314/ajb.v12i20.
[64] Lončarić M, Strelec I, Moslavac T, Šubarić D, Pavić V, Molnar M. Lipoxygenase Inhibition by Plant Extracts. Biomolecules 2021;11:152. https://doi.org/10.3390/biom11020152.
[65] Kostić DA, Dimitrijević DS, Stojanović GS, Palić IR, Đorđević AS, Ickovski JD. Xanthine Oxidase: Isolation, Assays of Activity, and Inhibition. Journal of Chemistry 2015;2015:e294858. https://doi.org/10.1155/2015/294858.
[66] Nieto FJ, Iribarren C, Gross MD, Comstock GW, Cutler RG. Uric acid and serum antioxidant capacity: a reaction to atherosclerosis? Atherosclerosis 2000;148:131–9. https://doi.org/10.1016/S0021-9150(99)00214-2.
[67] Choi HK, Mount DB, Reginato AM. Pathogenesis of Gout. Ann Intern Med 2005;143:499–516. https://doi.org/10.7326/0003-4819-143-7-200510040-00009.

[68] Smith HQ, Li C, Stanley CA, Smith TJ. Glutamate Dehydrogenase, a Complex Enzyme at a Crucial Metabolic Branch Point. Neurochem Res 2019;44:117–32. https://doi.org/10.1007/s11064-017-2428-0.
[69] Plaitakis A, Kalef-Ezra E, Kotzamani D, Zaganas I, Spanaki C. The Glutamate Dehydrogenase Pathway and Its Roles in Cell and Tissue Biology in Health and Disease. Biology 2017;6:11. https://doi.org/10.3390/biology6010011.
[70] Maguid S, Fernández-Alberti S, Parisi G, Echave J. Evolutionary Conservation of Protein Backbone Flexibility. J Mol Evol 2006;63:448–57. https://doi.org/10.1007/s00239-005-0209-x.
[71] Ramanathan A, Agarwal PK. Evolutionarily Conserved Linkage between Enzyme Fold, Flexibility, and Catalysis. PLOS Biology 2011;9:e1001193. https://doi.org/10.1371/journal.pbio.1001193.
[72] Iyer M, Jaroszewski L, Sedova M, Godzik A. What the protein data bank tells us about the evolutionary conservation of protein conformational diversity. Protein Science 2022;31:e4325. https://doi.org/10.1002/pro.4325.