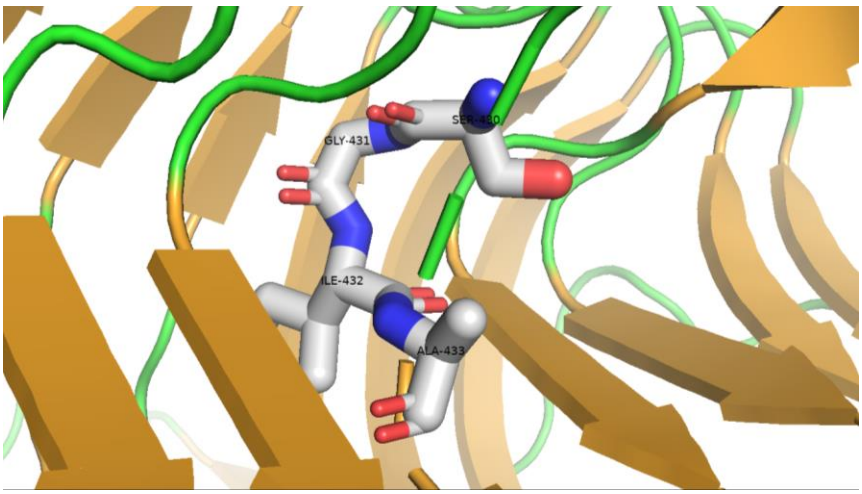
DEMOCRITUS UNIVERSITY OF THRACE
HEALTH SCIENCES SCHOOL
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

Bsc Thesis

"Buried β-turns in Hydrophobic Cores: Structural and Functional Implications"



Author:
Aikaterini - Alexandra Giannaki

Advisor:
Dr. Nicholas M. Glykos
Associate Professor of Structural and Computational Biology

Alexandroupolis, Greece
October 2021

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ

Διπλωματική Εργασία

"Θαμμένες β-στροφές σε Υδρόφοβους Πυρήνες: Δομικές και Λειτουργικές Επιπτώσεις"

Αικατερίνη - Αλεξάνδρα Γιαννάκη
(ΑΕΜ 1950)

Επιβλέπων καθηγητής:
Δρ. Νικόλαος Μ. Γλυκός

Αλεξανδρούπολη
Οκτώβριος 2021

Acknowledgements

I would like to sincerely thank my advisor Dr. Nicholas M. Glykos for his patience and for always pushing my limits and making me work harder and tougher to achieve my goals. His advice has been invaluable and I am deeply grateful that I had the opportunity to work with him.

I would, also, like to thank my friends, who have supported me on this journey. I will cherish our memories and I wish you all the best.

Lastly, I would like to thank my family for their immense support and encouragement throughout my studies.

Table of Contents

Abstract

A turn is a protein secondary structure where the polypeptide chain reverses its direction. In particular, a β–turn consists of only four consecutive residues and a γ– turn, which is the second most commonly found turn after β–turns, involves three consecutive residues. Turns are usually located at the molecular surface. However, they can sporadically be found buried within the protein hydrophobic core. The precise role of such occurrences remains unknown. In this study, we identify such buried turns and we examine putative structural and functional implications of their presence. In order to search for these patterns, we developed a program which incorporates and systematically applies PROMOTIF's and STRIDE's algorithms and identifies buried turns in a large sample of structures obtained from the Protein Data Bank via the PISCES interface. Our current results indicate that type IV for β-turns and inverse type for γ-turns are the most common types found in hydrophobic turns, and, moreover, we show that such motifs seem to occur more frequently in proteins that function as enzymes.

Περίληψη

Στροφή ονομάζεται η πρωτεϊνική δευτεροταγής δομή όπου η πολυπεπτιδική αλυσίδα αναστρέφει την κατεύθυνσή της. Πιο συγκεκριμένα, μια β-στροφή περιλαμβάνει τέσσερα αμινοξικά κατάλοιπα σε σειρά και μια γ-στροφή, η οποία αποτελεί την δεύτερη πιο συχνά παρατηρούμενη στροφή μετά την β-στροφή, αποτελείται από τρία αμινοξικά κατάλοιπα. Οι στροφές συνήθως εντοπίζονται στην επιφάνεια των μορίων. Ωστόσο, σποραδικά έχουν εντοπιστεί θαμμένες στον υδροφοβικό πυρήνα της πρωτεΐνης. Ο ακριβής ρόλος αυτών των ευρημάτων παραμένει άγνωστος. Στην συγκεκριμένη μελέτη, εντοπίζουμε τέτοιες θαμμένες στροφές και ερευνούμε υποθετικές δομικές και λειτουργικές επιπτώσεις της παρουσίας τους. Με σκοπό να αναζητήσουμε αυτά τα μοτίβα, κατασκευάσαμε ένα πρόγραμμα, το οποίο ενσωματώνει και εφαρμόζει συστηματικά τους αλγορίθμους των PROMOTIF και STRIDE και, επιπλέον, αναγνωρίζει θαμμένες στροφές από ένα ευρύ δείγμα δομών, οι οποίες έχουν ληφθεί από την Protein Data Bank μέσω χρήσης του PISCES server. Τα τρέχοντα αποτελέσματά μας υποδεικνύουν πως η στροφή τύπου IV για τις β-στροφές και η inverse στροφή για τις γ-στροφές αποτελούν τους πιο συχνούς τύπους στροφών στις υδρόφοβες στροφές, καθώς και, επίσης, επισημαίνουμε ότι αυτά τα μοτίβα φαίνεται να εμφανίζονται με μεγαλύτερη συχνότητα σε πρωτεΐνες με ενζυμικές λειτουργίες.

1. Introduction

1.1 Proteins

A protein is defined as a polypeptide structure, consisting of combinations of amino acid residues. They are complex molecules that carry out crucial organism functions, such as immune protection, metabolism, structural support. Amino acids are structural units that contain carbon, hydrogen, oxygen, nitrogen and, sometimes, sulfur. Specifically, amino acids consist of a central carbon called the α-carbon connected to an amino group (-NH_2), a hydrogen atom, a carboxyl group (-COOH) and a side chain (R). Proteins are polymers of these structural units, they can be "built" using only a total of 20 amino acids, each of which has a distinct side chain.

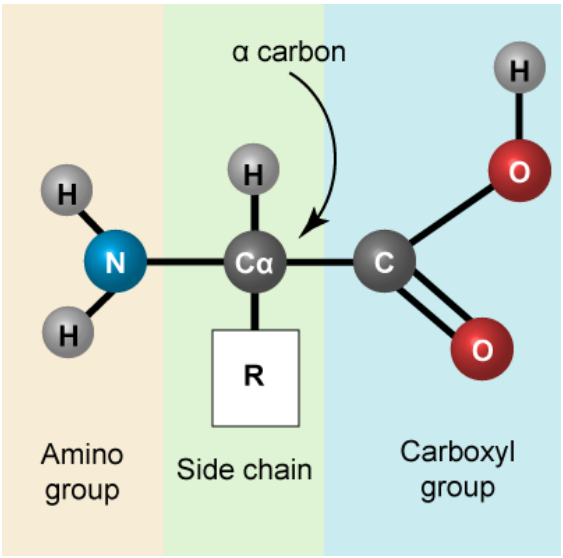


Figure 1. Amino acid structure (reproduced without permission from https://bio.libretexts.org/)

Due to the side chains of the amino acids having different chemical properties, amino acids can be classified into four different groups based on the polarity of the R group. Group I consists of non-polar amino acids, glycine, alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine and tryptophan. The side chains of these amino acids contain either aliphatic or aromatic groups. These amino acids are hydrophobic, which means that they lack affinity for water and are usually buried in the protein interior. Group II consists of polar, uncharged amino acids, serine, cysteine, threonine, tyrosine, asparagine and glutamine. Most of the side chains from this group have at least one atom that can participate in hydrogen bonding. Group III, which is the acidic (negatively charged) amino acids, includes only two, aspartic acid and glutamic acid, called aspartate and glutamate in the ionic forms. These amino acids have a carboxylic acid group on their side chain, which offers them proton-donating properties. Lastly, group IV contains three basic (positively charged) amino acids, arginine, histidine, lysine. The side chains in this group are basic, which means they can accept a proton. Most of the amino acids in the three last groups show hydrophilic properties, i.e. strong affinity for water[1].

Amino acids can be linked by a peptide bond. A peptide bond is a condensation reaction, in which the carbonyl carbon of the first amino acid and the α-nitrogen atom of the second form a covalent bond by removal of water. Peptide bonds are mainly found in trans conformation with a torsion angle $\Omega \approx 180^{o}$, with proline being the only amino acid that can be found in a cis conformation ($\Omega \approx 0^{o}$). The linear sequence of amino acids joined by a series of peptide bonds constitute the primary structure of the protein. Protein structure can be classified into four major categories, primary, secondary, tertiary and quaternary structure. The secondary structure involves local interactions between parts of the polypeptide chain. The two most common elements that can alter the secondary structure are α-helices and β-sheets. The tertiary structure, known as the native state of the protein, refers to the folding of the secondary structural elements into a distinct three-

dimensional shape. Many proteins consist of more than one polypeptide chain or subunit. The arrangement of multiple protein chains or subunits into a single functional molecule establish the quaternary structure. The interactions stabilizing this structure can be covalent or noncovalent[2].

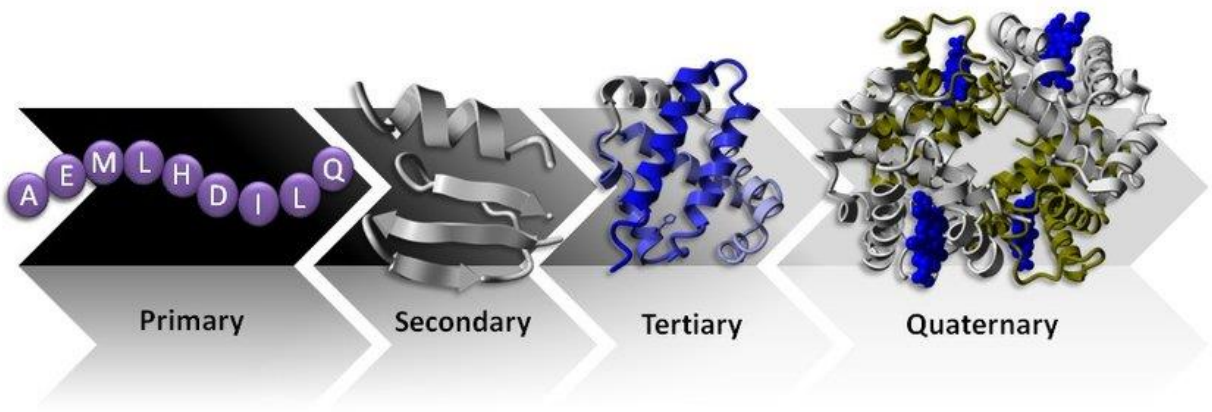


Figure 2. Protein structure (reproduced without permission from Joosten, R. P. https://hdl.handle.net/2066/76549)

1.2 Secondary structure elements

As mentioned before, protein secondary structure can be described as the conformation of the polypeptide backbone, due to steric constraints applied to peptide bonds and amino acid residues. Secondary structures arise as hydrogen bonds form between backbone atoms, and include the partially negatively charged oxygen atom and the partially positively charged nitrogen atom. Notably, the hydrogen bonds involved in secondary structure do not comprise of any amino acid R groups. The reason for this phenomenon can be explained considering the existence of imine groups (NH) and carbonyl groups (C'=O) in each amino acid, which act as proton donors and proton receptors respectively, resulting in high backbone hydrophilicity. As the peptide chain folds, hydrophobic side chains are packed towards the center of the protein molecule, forming a hydrophobic core and a hydrophilic outer surface. In such hydrophobic environment, polar groups need to be neutralized, and this can be achieved by the formation of hydrogen bonds between these groups[3]. Henceforth, stable conformational patterns, known as secondary structure elements, are formed.

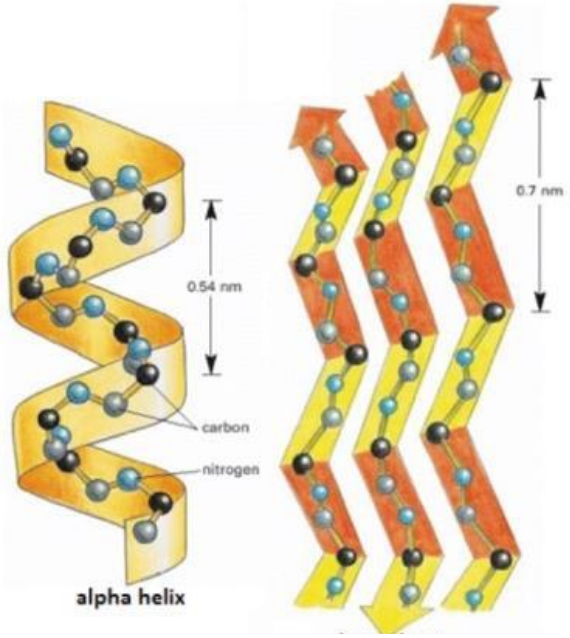


Figure 3.α-helix and β-sheet (reproduced without permission from Alberts et al., https://www.researchgate.net/publication/297316514_Techniques_and_applications_of_proteomics_in_plant_ecophysiology)

The two most common folds are α-helices and β-sheets. The α-helix is a right-handed helical coil, in which the backbone NH group form a hydrogen bond with the backbone C'=O every fourth amino acid. In every helical turn, 3.6 amino acid residues are involved. In the β-sheet, two different short regions of the polypeptide chain are aligned side by side and linked by hydrogen bonds. There are two types of β-sheets, parallel and antiparallel β-sheets. If the

adjacent polypeptide chains have the same N to C direction it is a parallel β-sheet, and if they have the opposite direction the β-sheet is antiparallel. However, these two structural elements are not the only ones found in protein structures. Some examples are α_L-helices, 3_{10}-helices with 3 residues per turn, π-helices with 4.1 residues per turn, β- or γ- turns and random coils.

1.3 φ and ψ dihedral angles

It is not possible to strictly define a secondary structure element only by identifying the hydrogen bonding pattern. Comprehension of the basic parameters that affect the conformation of a peptide is essential. Since peptide bonds have a rigid planar configuration due to their partial double bond character, each amino acid residue backbone has two degrees of freedom that correspond to the torsion angles of the N-Cα and Cα-C' bond. These dihedral torsion angles are called φ and ψ respectively. In theory, the value of the φ and ψ dihedral torsion angles can range from -180° to +180°. However, in practice, many values are restricted by the steric hindrance between main chain and side chain atoms.

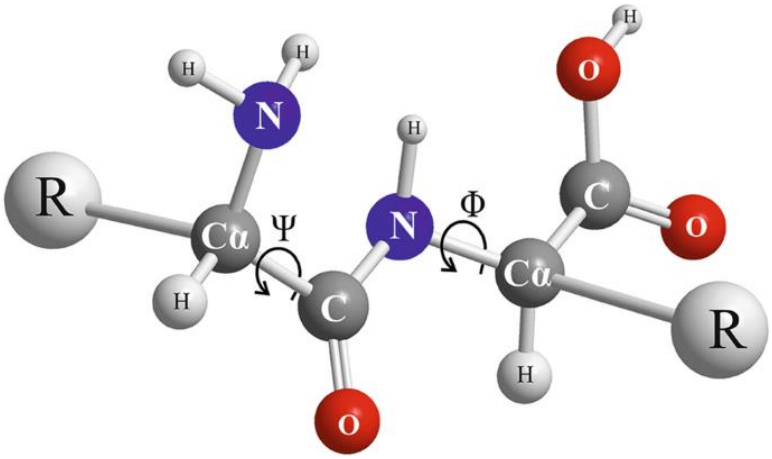


Figure 4. φ and ψ dihedral angles (reproduced without permission from Esfandi B., Atabati M. https://doi.org/10.1007/s10930-020-09961-6)

1.4 β-turns

β–turns can be described as a secondary structure motif which involves four consecutive residues, R_i, R_{i+1}, R_{i+2}, R_{i+3}, where the distance between the atoms Ca_i and Ca_{i+3} is less than 7 Å and the two central residues are not in a helical conformation[4]. They constitute the most frequent non-repetitive structural element recognized in proteins, including an average of 25% of the residues[5]. Furthermore, β-turns have a significant biological role, since they assist in the polypeptide chain folding and can be involved in molecular recognition[6].

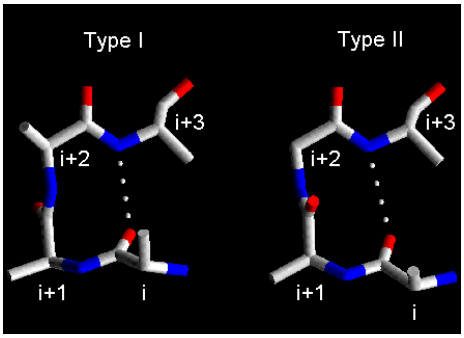


Figure 5. Type I and II β-turns (reproduced without permission from http://www.cryst.bbk.ac.uk/PPS95/course/3_geometry/index.html)

Venkatachalam originally identified six distinct β-turn conformations, I, II, III and their mirror images I', II', III', stabilized by a hydrogen bond between the backbone CO_i and NH_{i+3}[7]. A few years later, Lewis et al. observed that 25% of β-turns do not form an intraturn bond as stated by Venkatachalam and proposed 10 possible β-turn conformations (I, I', II, II', III, III', IV, V, VI and VII) by

incorporating φ and ψ angles and less stringent criteria[8]. However, currently, the classification introduced by Richardson and Thornton et al. is broadly accepted. Richardson and Thornton, while studying the values of φ and ψ angles of residues R_i and R_{i+2}, defined nine different turn types, Type I, Type I', Type II, Type II', Type IV, Type VIa1, Type VIa2, Type VIb and Type VIII[9]. The representative φ and ψ values for each of the nine turn types are shown in Table 1.

Turn Type	Dihedral Angles (°)			
	φ(i+1)	**ψ(i+1)**	**φ(i+2)**	**ψ(i+2)**
I	-60	-30	-90	0
I'	60	30	90	0
II	-60	120	80	0
II'	60	-120	-80	0
IV	-61	10	-53	17
VIa1	-60	120	-90	0
VIa2	-120	120	-60	0
VIb	-135	135	-75	160
VIII	-60	-30	-120	120

Table 1. The nine β-turn types classified by Hutchinson and Thornton[9]

Turn types I, I', II, II' and VIII perfectly meet the criteria stated by the definition of β-turns. For the turn types VIa1, VIa2 and VIb, the presence of a proline as the third residue (R_{i+2}) and Ω_{i+1} equal to 0° is necessary, as in other cases, the distances between Ca_i and Ca_{i+3} are greater than 7.0 Å[4]. Moreover, type IV turns, the miscellaneous category, are the second most common β-turn type, after type I β-turns, and represent 1/3 of β-turn residues[10].

1.5 γ-turns

γ-turns are the second most commonly found turns after β-turns. γ-turns involve only three consecutive residues, R_i, R_{i+1}, R_{i+2}, with a hydrogen bond forming between the backbone CO_i and NH_{i+2}[11]. It is widely believed that the hydrogen bonds forming in γ-turns are weak, due to the difficulty of bridging closely spaced residues. γ-turns are classified into two different types, classic

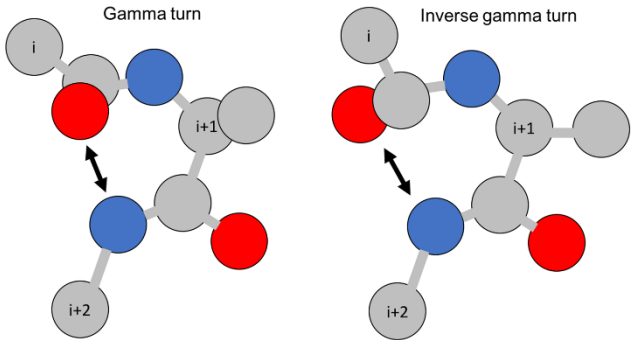


Figure 6. γ-turns (reproduced without permission from Fang et al., https://doi.org/10.1038/s41598-018-34114-2)

and inverse. The main chain atoms of the two forms can be found in two different enantiomers, just like β-turn types I and I' or II and II'[12]. Of the two forms, inverse γ-turns are much more common.

1.6 Solvent-accessible surface area

Hydrophobic interactions take a crucial role in the stability and folding of protein structures. Hydrophobicity is naturally proportional to the solvent-accessible surface area[13]. Solvent-accessible surface area (SASA), or accessible surface area (ASA) is defined as the surface area of a molecule that is accessible to a solvent. Protein amino acid residues can be classified as exposed or buried based on their SASA. There are various types of SASAs ranging from relative solvent accessibility to absolute surface areas[14]. SASA is predominantly estimated by in silico methods involving the rotation of a spherical probe, resembling a water molecule, around a full-atom protein model[15]. The algorithm typically used for the calculation of the SASA was developed by Shrake and Rupley[16] and contains the testing of the overlapping of points on an atom's van der Waals surface with points on the van der Waals surface of neighboring atoms.

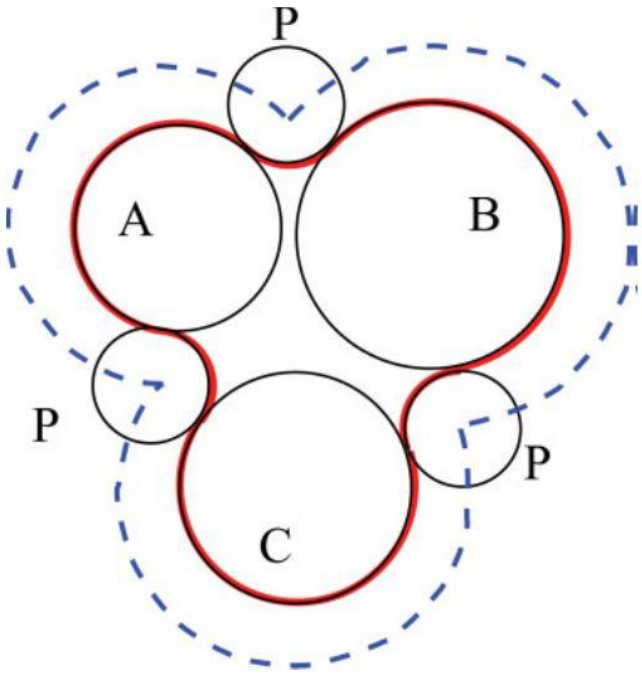


Figure 7. Calculation of SASA. The bold continuous line represents the molecular surface. The bold dashed line shows the SASA. P is the probe sphere and represents the water molecule (reproduced without permission from https://slideplayer.com/slide/4701892/)

1.7 Purpose of the present thesis

Turns are more frequently found on the water-accessible surface of the proteins. Specifically, an isolated β-turn is highly polar, because of the unpaired NH and CO groups. However, there have been reports where turns have been located buried within the hydrophobic core of the protein. At 1983, Rose et al.[17], screened the PDB[18],[19],[20] and found six proteins including buried turns, following Venkatachalam's[7] β-turn conformations. At that time, PDB contained a total of 77 atomic coordinate records for 47 macromolecules[21]. All the turns studied were β-turns type I or II. Rose et al., noticed the presence of one or more firmly complexed water molecules in the local area of each turn. These water molecules were buried within the protein and were bound to the buried turn with at least three hydrogen bonds.

Since then, computational biology has evolved, plenty more protein atomic structures have been studied and uploaded to the PDB and Venkatachalam's β-turn type classifications have been revised. To this current day, the precise structural and functional significance of such occurrences is still unknown. We decided to further study these motifs, using contemporary methods, less stringent criteria and non-redundant protein structures. Therefore, we algorithmically processed a large dataset of existing protein structures for buried β- and, also, γ-turns in the hydrophobic core

and identified these motifs, while observing common structural and functional features shared between them.

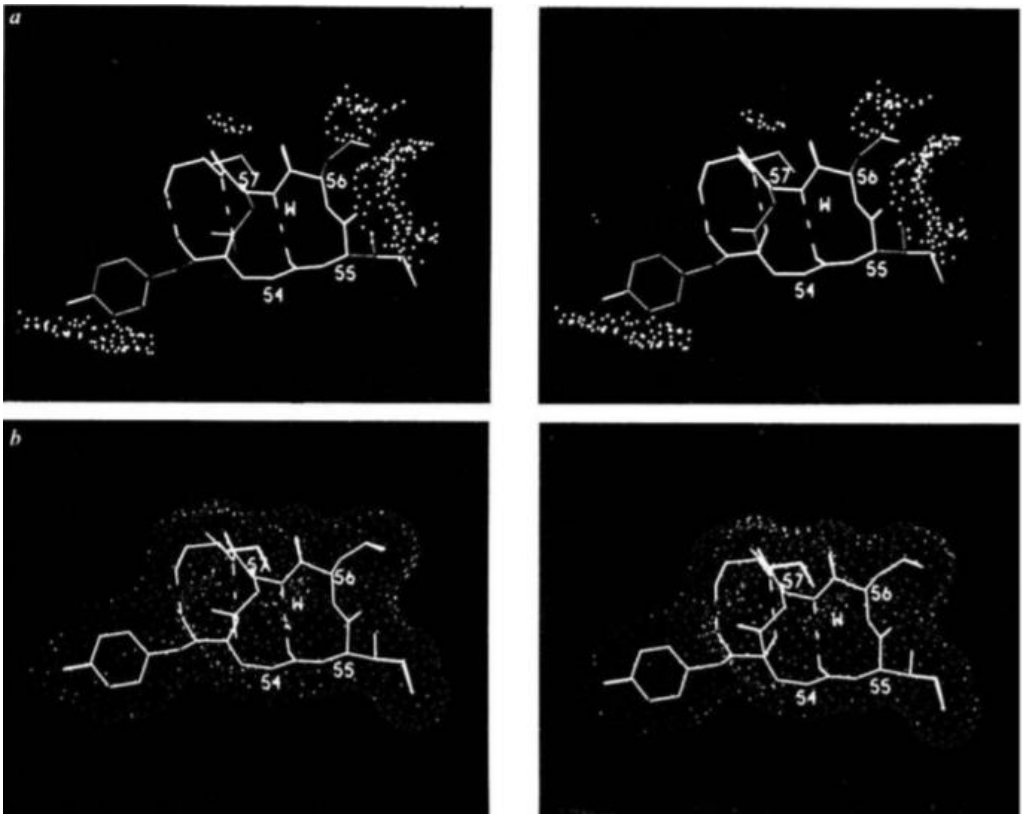


Figure 8. The lysozyme β-turn, showing the SASA. The position of the firmly complexed water molecule is show in W (reproduced without permission from Rose et al., https://doi.org/10.1038/304654a0)

2. Computational Methods

2.1 The Perl programming language

Figure 9. Perl language logo (reproduced without permission from https://www.pngitem.com/)

Perl is a high-level, general-purpose, dynamic, interpreted programming language, introduced by Larry Wall in 1987. It was originally developed for text processing; however, it is now used for a variety of tasks such as system administration, web development, network programming and more. Perl is an open-source software, licensed under the GNU General Public License (GPL), therefore, there is an extensive variety of Perl modules accessible. Although Perl, being an interpreted programming language, does not require the use of a compiler, it is slower compared to compiled languages. Even so, Perl's incorporation of regular expressions has been of great use in bioinformatics and computational biology. The various programs developed for this project are written in Perl and take advantage of Perl's variety of modules and regular expressions.

2.2 Protein Data Bank

The Protein Data Bank (PDB)[18],[19],[20] is an online database that stores 3D biomolecular structures of proteins, nucleic acids and complex assemblies, studied by researchers around the globe. In our research, we used the PDB, in order to obtain the protein structures needed. At the time this thesis is written, PDB contains 182,624 structures. An example of a PDB formatted file is shown in **Figure 10** and a screenshot of the PDB user interface is shown in **Figure 11**.

```
HEADER    TRANSFERASE/TRANSFERASE INHIBITOR       01-MAR-12   4E00
TITLE     CRYSTAL STRUCTURE OF BRANCHED-CHAIN ALPHA-KETOACID DEHYDROGENASE
TITLE    2 KINASE/3,6-DICHLOROBENZO[B]THIOPHENE-2-CARBOXYLIC ACID COMPLEX WITH
TITLE    3 ADP
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: [3-METHYL-2-OXOBUTANOATE DEHYDROGENASE [LIPOAMIDE]] KINASE,
COMPND   3 MITOCHONDRIAL;
COMPND   4 CHAIN: A;
COMPND   5 SYNONYM: BRANCHED-CHAIN ALPHA-KETOACID DEHYDROGENASE KINASE, BCKD-
COMPND   6 KINASE, BCKDHKIN;
COMPND   7 EC: 2.7.11.4;
COMPND   8 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: RATTUS NORVEGICUS;
SOURCE   3 ORGANISM_COMMON: RAT;
SOURCE   4 ORGANISM_TAXID: 10116;
SOURCE   5 GENE: BCKDK;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 511693;
SOURCE   8 EXPRESSION_SYSTEM_STRAIN: BL21GROESL;
SOURCE   9 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE  10 EXPRESSION_SYSTEM_PLASMID: PTRCKHISB
KEYWDS    GHKL PROTEIN KINASE, ALLOSTERIC KINASE INHIBITOR,BRANCHED-CHAIN
KEYWDS   2 ALPHA-KETOACID, BRANCHED-CHAIN AMINO ACIDS, MAPLE SYRUP URINE
KEYWDS   3 DISEASE,DIABETES AND OBESITY, BERGERAT NUCLEOTIDE-BINDING FOLD,
KEYWDS   4 PROTEIN KINASE, TRANSFERASE-TRANSFERASE INHIBITOR COMPLEX
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.C.TSO,J.L.CHUANG,W.J.GUI,R.M.WYNN,J.LI,D.T.CHUANG
REVDAT   1   27-MAR-13 4E00    0
```

Figure 10. Format of a PDB file

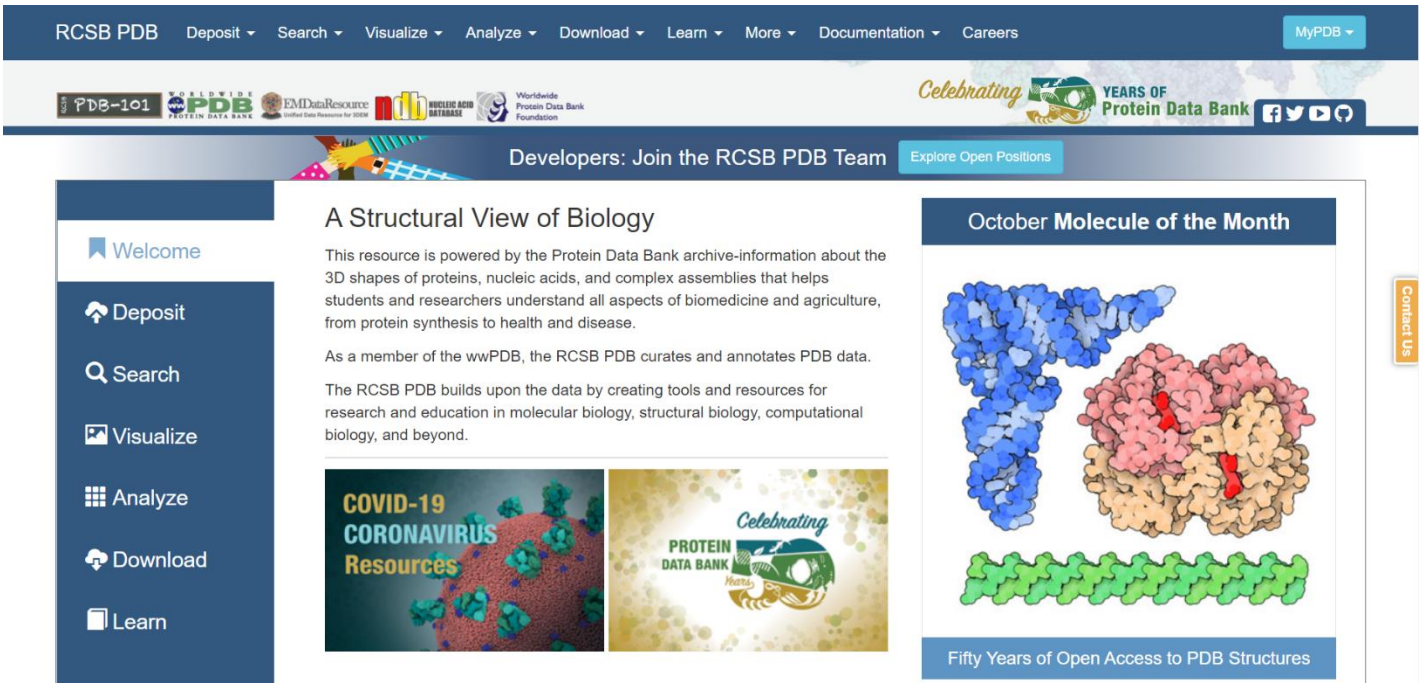


Figure 11. PDB user interface

2.3 PISCES culling server

PISCES[22] is a public protein sequence culling server from PDB by structural quality criteria and sequence identity. PISCES can provide better lists than servers that use BLAST, since BLAST has been found to align only well-conserved fragments, overestimating sequence identity, and is incapable of identifying many relationships below 40% sequence identity. PISCES is a tool that can be used to obtain lists of non-redundant (not repeating) entries from the entire PDB, using user-provided input criteria.

A large, non-redundant sample of proteins needed to be obtained. In order to avoid false results or artefacts, by identifying a possible conformational pattern, due to an iteration of a sequence in many identical structures, the exclusion of PDB's identical entries was vital. A list containing 26388 entries was created using PISCES, using the following criteria: maximum 2.2 Å resolution, maximum R-factor 1.0 and 70% identity cut–off.

2.4 FTP: File Transfer Protocol

FTP is a network protocol for transferring computer files from a server to a client. The files obtained from the PDB were downloaded via FTP. The list created by PISCES contains PDB IDs in a five letter, uppercase format. **Script 1** (lc.pl)(source code in **Appendix**) modifies the list from PISCES and creates a new list containing names in a four letter, lowercase format. This new list was used as input for **Script 2** (ftp.pl), which downloads all the entries in the list from the PDB server (http://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/). 24951 compressed PDB files were

downloaded. The deviation in the number of files downloaded from PDB from the number of ID's given by PISCES, is due to multiple IDs for each polypeptide chain in the list created by PISCES and, moreover, some files were only found in the mmCIF format.

2.5 Identification of β- and γ- turns: PROMOTIF

PROMOTIF[23] is a program that provides information about secondary structural motifs in the protein. Specifically, PROMOTIF can currently identify disulphide bridges, β- and γ- turns, β-strands, β-bulges, β-hairpins, β-α-β units, ψ loops, β-sheet topology, helical geometry, helical interactions and main chain hydrogen bonding patterns. PROMOTIF uses a slightly modified version of the DSSP[5] algorithm, in which, if possible, one extra amino acid residue is counted at the ends of each helix or strand and designated as lowercase "h" and "e". For this project, only the identification of β- and γ-turns was necessary, so **Script 3** (promotif_results.pl) was created, which incorporates PROMOTIF's p_sstruc and p_turn, runs these programs using as input all the PDB files saved in the working directory and saves the results in a personalized format. In this format the chain ID, first and last residue and turn type are printed and, only for β-turns, whether there is a hydrogen bond involved in the main chain.

2.6 Calculation of solvent accessibility: STRIDE

STRIDE[24] is an open-source, algorithm for secondary structure identification from atomic resolution protein structures. Compared to the DSSP[5] algorithm, STRIDE makes combined use of hydrogen bonding patterns and backbone geometry. The STRIDE server can be accessed from http://webclu.bio.wzw.tum.de/stride/. **Script 4** (stride_results.pl) uses the PROMOTIF[23] results file as input, implements the STRIDE algorithm in order to calculate the solvent accessible area of the residues involved in the β- and γ-turns found and saves these results in a personalized format. Subsequently, **Script 5** (chain.pl) was then used, which uses a modification of the list made by PISCES[22] and only saves the chain IDs dictated. **Script 6** (format.pl) was used for the modification of the list, in order to acquire the desirable format.

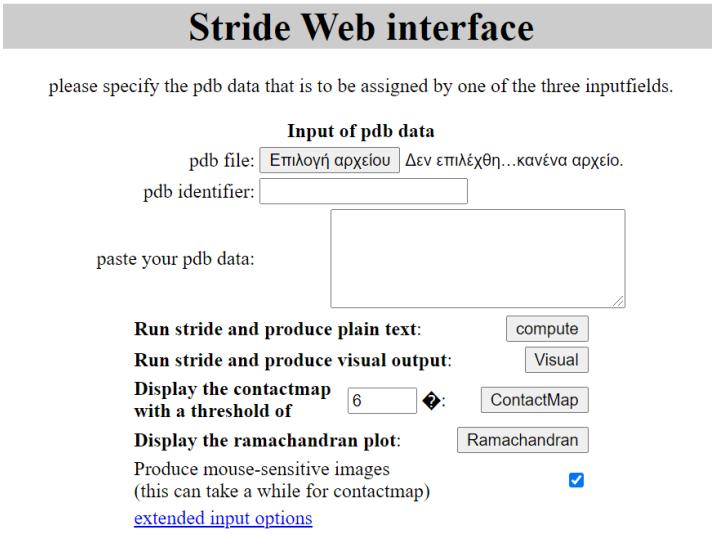


Figure 12. STRIDE web interface

2.7 Construction of histograms: Gnuplot

Gnuplot[25],[26] is a free, command-driven, multi-platform data plotting program. Gnuplot was used for plotting the density distributions. **Script 7** (max.pl) calculates the maximum solvent accessibility number of the residues involved in a β- or γ-turn and prints these numbers at a separate

file. **Script 8** (histogram_data.pl) organizes these maximum numbers in the optimal format needed by Gnuplot for the construction of a histogram. Specifically, **Script 8** first calculates the full range of the values given and, then, this range is divided into different sub-ranges using a user-selected bin number and calculates the sum of the values that fall into each sub-range. The output results are printed in a "sub-range sum" format.

2.8 Hydrophobic turns: PyMOL

Script 9 (hydrophobic_turns.pl) was used in order to only locate the β- or γ-turns with a maximum solvent accessibility value of 0.0. **Script 9** is a variant of **Script 7**. At first, **Script 9** calculates the maximum solvent accessibility of the residues involved in the turn, and afterwards checks if the value is in the range 0.0-0.2 and if this condition is true, it prints the file name and the hydrophobic turn. PyMOL[27], an open-source, user-sponsored, molecular visualization system, was then used to study and showcase some of the found hydrophobic turns.

3. Results

3.1 Histograms

Four histograms with bin numbers 0.2, 0.4, 0.6 and 0.8 showing the distribution of maximum solvent accessibility in the found turns were plotted using Gnuplot[25],[26]. All of the histograms plotted revealed a peak at 0.0, a drop at around 1.0, another peak at around 2.0, a plateau from 2.0 till 60.0 and then a greater peak at around 100.0. The histograms are presented below.

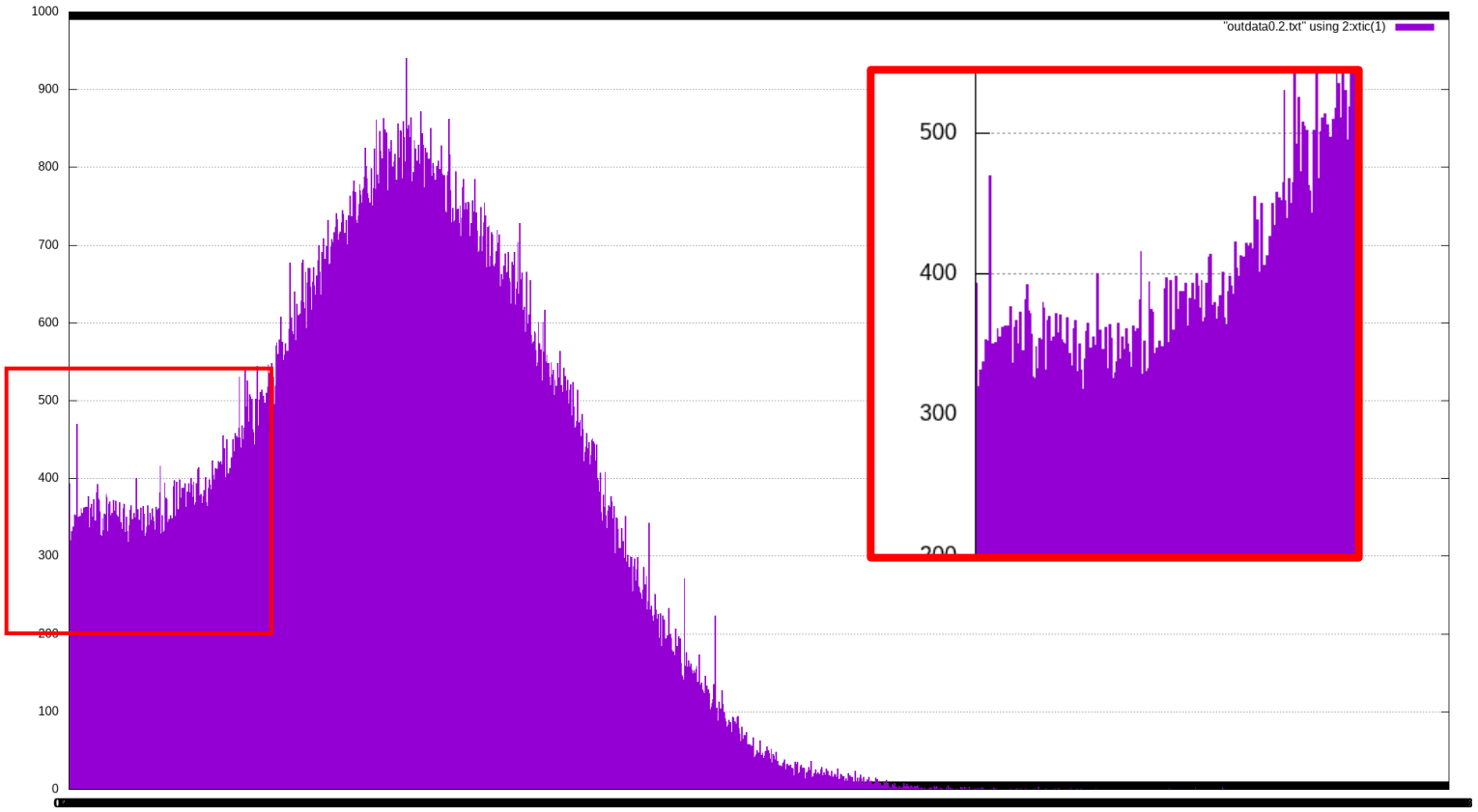


Figure 13. Histogram with bin 0.2, zooming at area 0.0~70.0

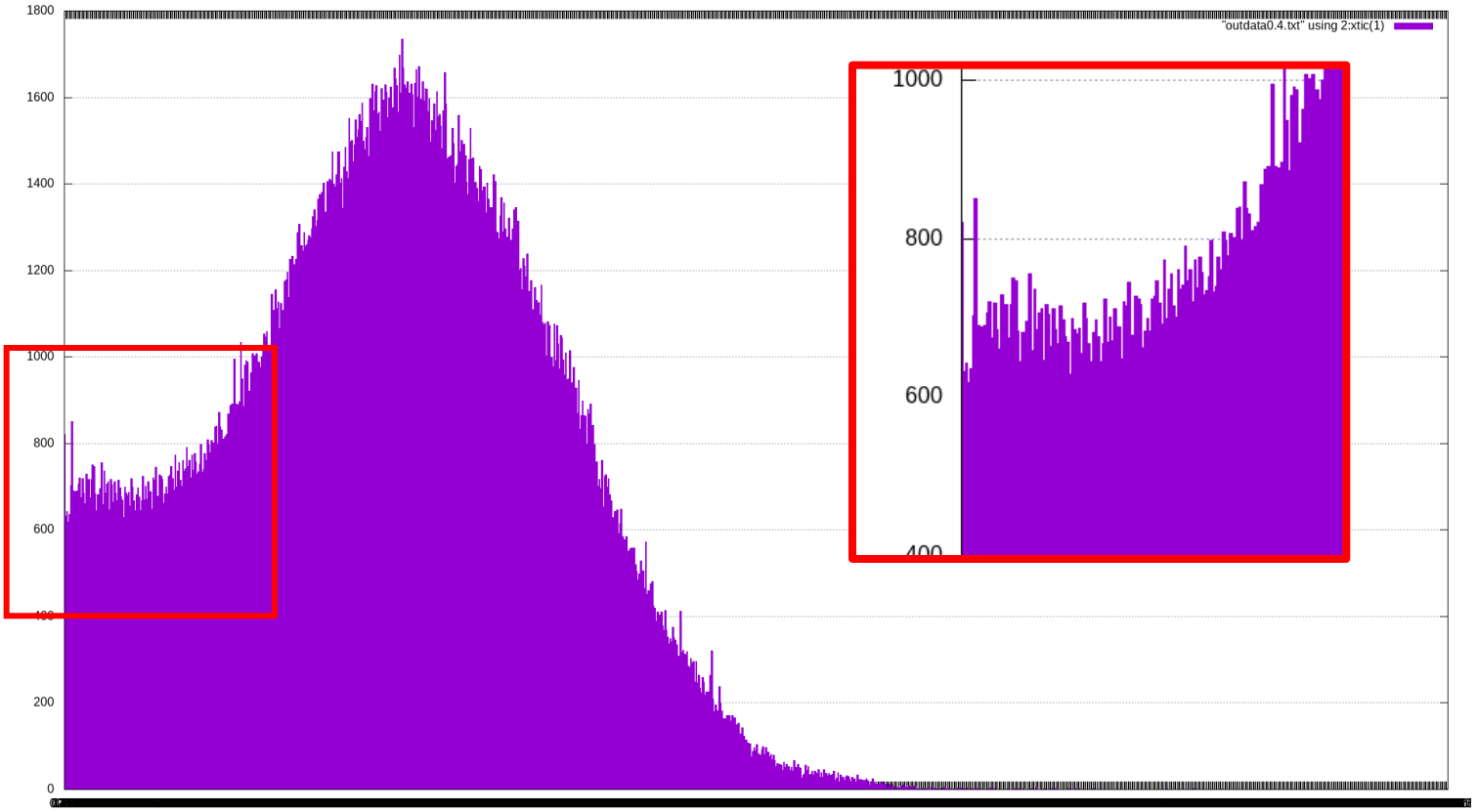


Figure 15. Histogram with bin 0.4, zooming at area 0.0~70.0

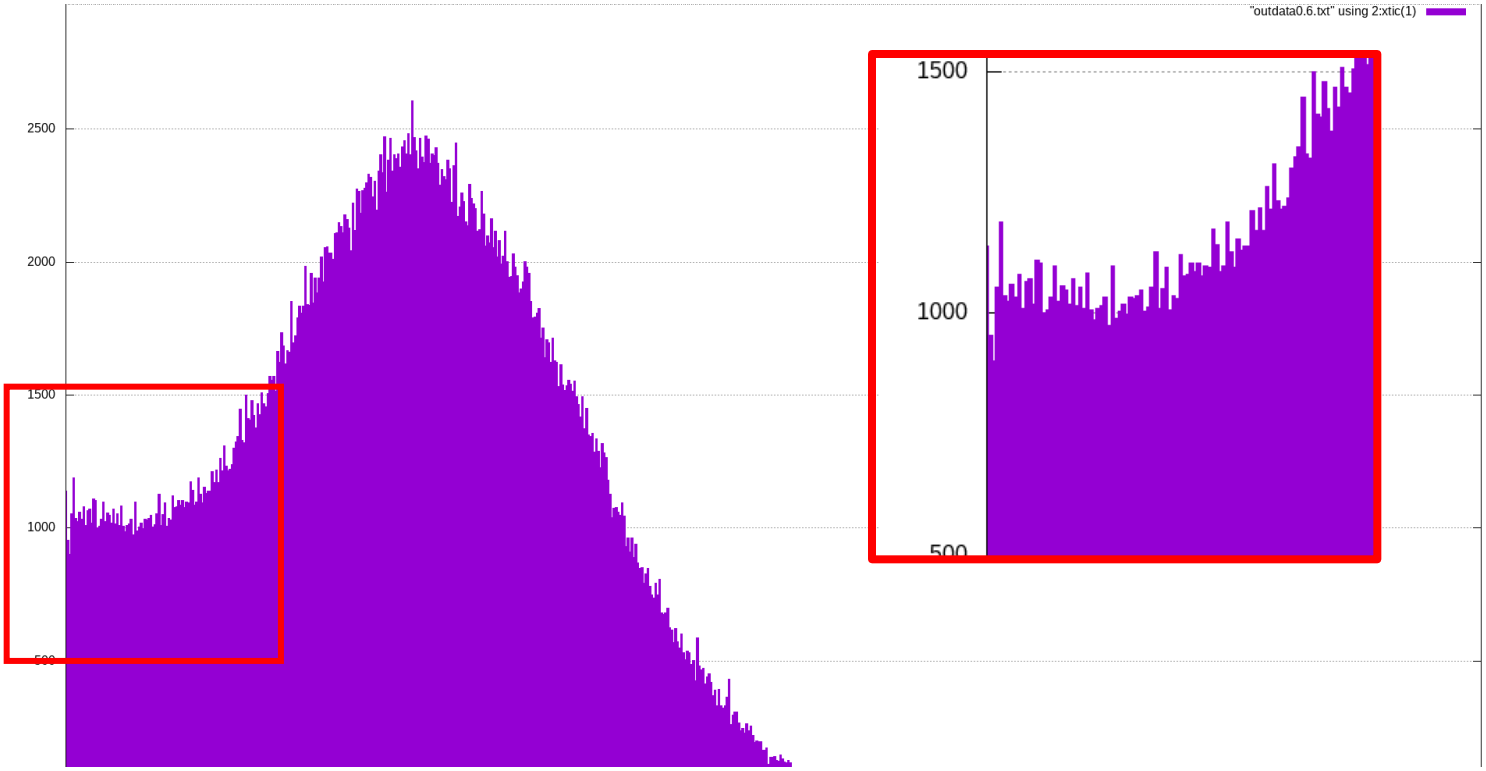


Figure 14. Histogram with bin 0.6, zooming at area 0.0~70.0

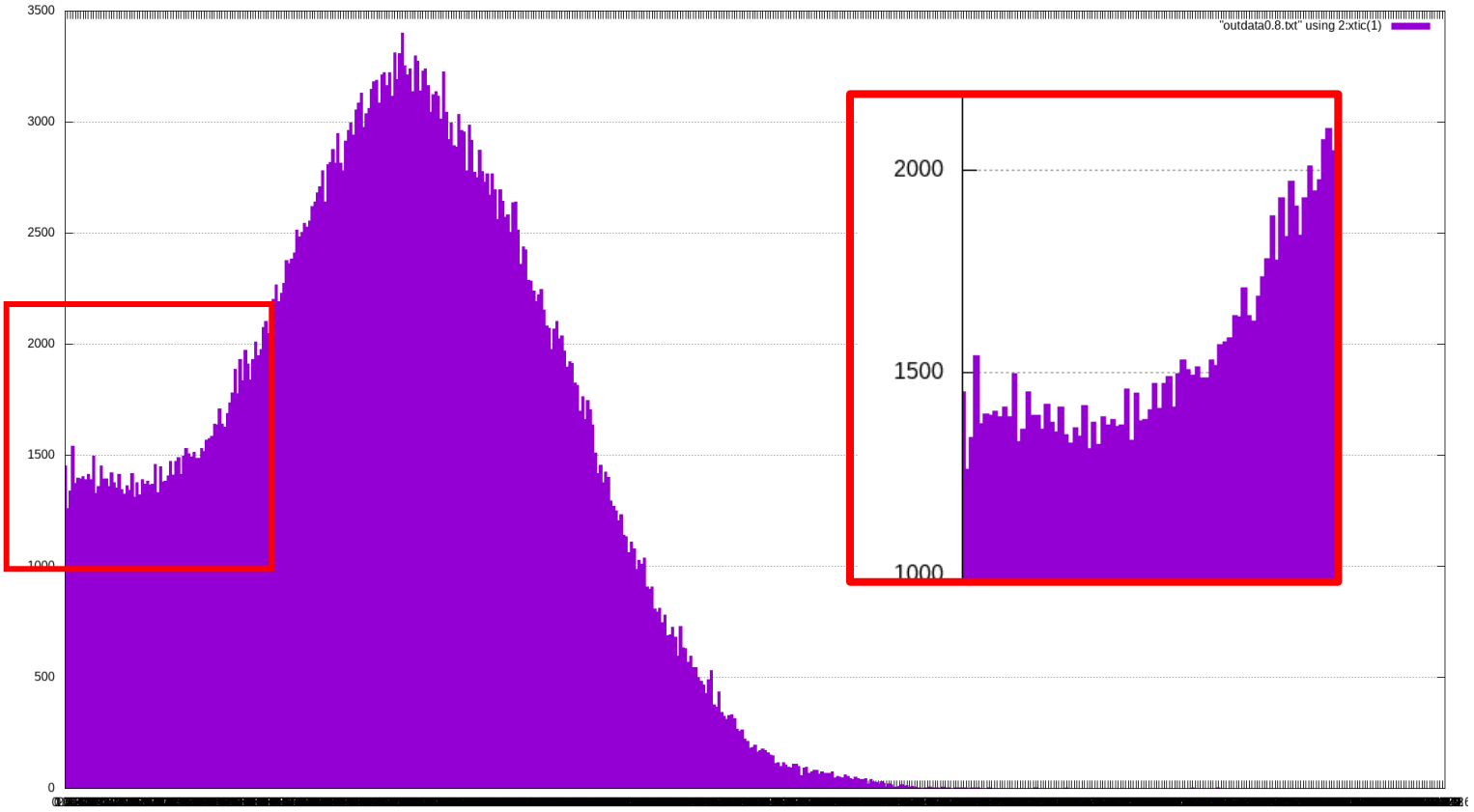


Figure 16. Histogram with bin 0.8, zooming at area 0.0~70.0

3.2 Hydrophobic turns

A total of 427 β– and γ–turns with a maximum solvent accessibility number of the residues involved equal to 0.0 were found. These 427 turns involve a total of 397 proteins out of the 24921 proteins in the initial sample, i.e., an average of 1.6% of protein structures included one or more hydrophobic turns. 26 protein structures (6.5%) were found to have multiple hydrophobic turns (23 structures had two hydrophobic turns, 2 structures had three and 1 had four). These turns were often found to be overlapping (e.g., residues 42-45 β-turn and residues 43-45 γ-turn). A large sample of the hydrophobic turns found were studied using PyMOL[27] and 18 of them are showcased below.

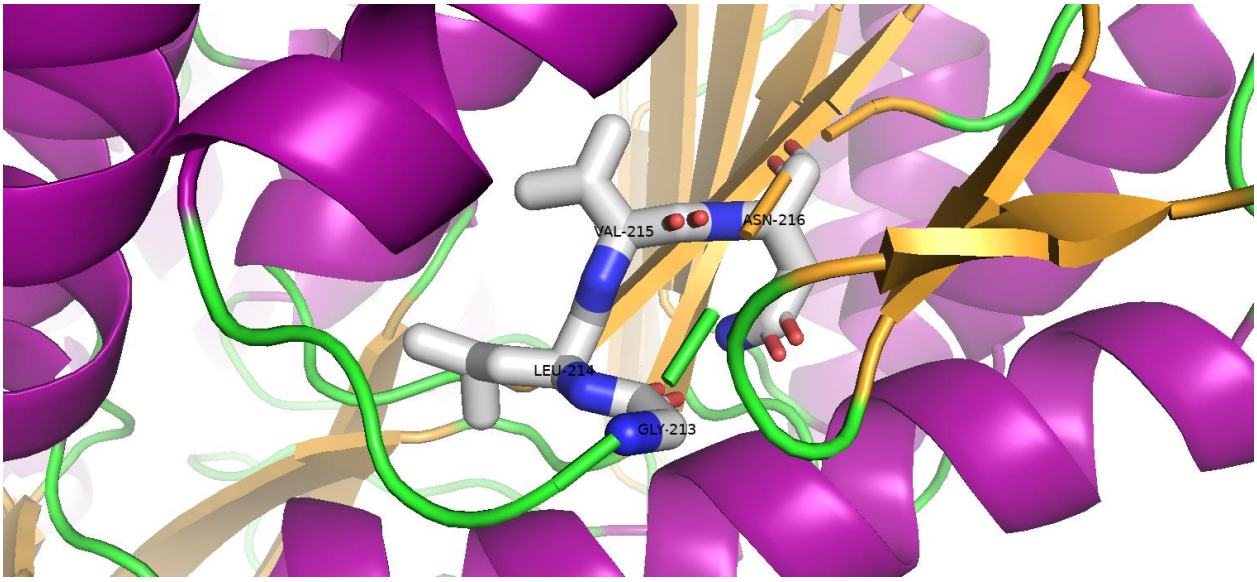


Figure 17. 1a4s, β-turn type VIII, involving residues 213-216 in chain A

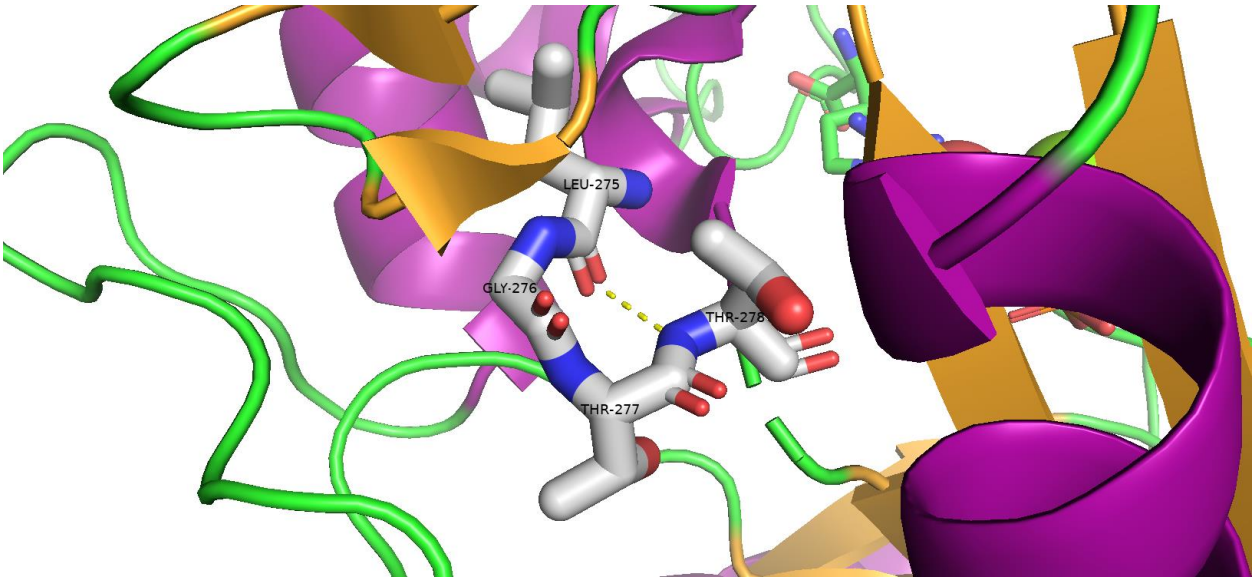


Figure 18. 1m15, β-turn type II' involving residues 275-278 in chain A

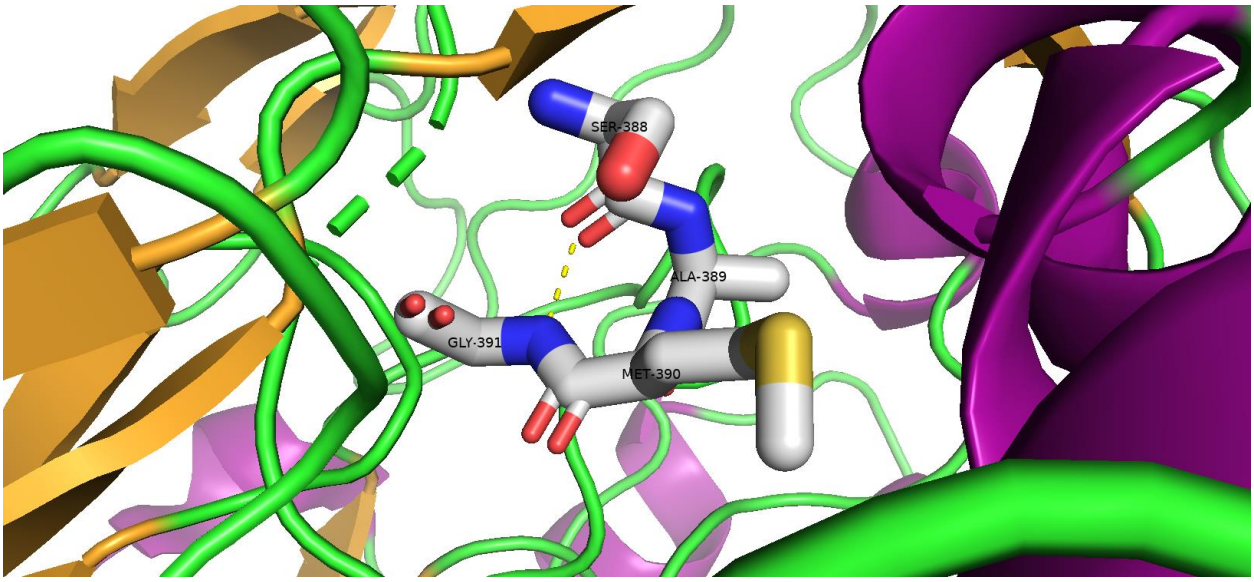


Figure 19. 1w6s, β-turn type I, involving residues 388-391 in chain A

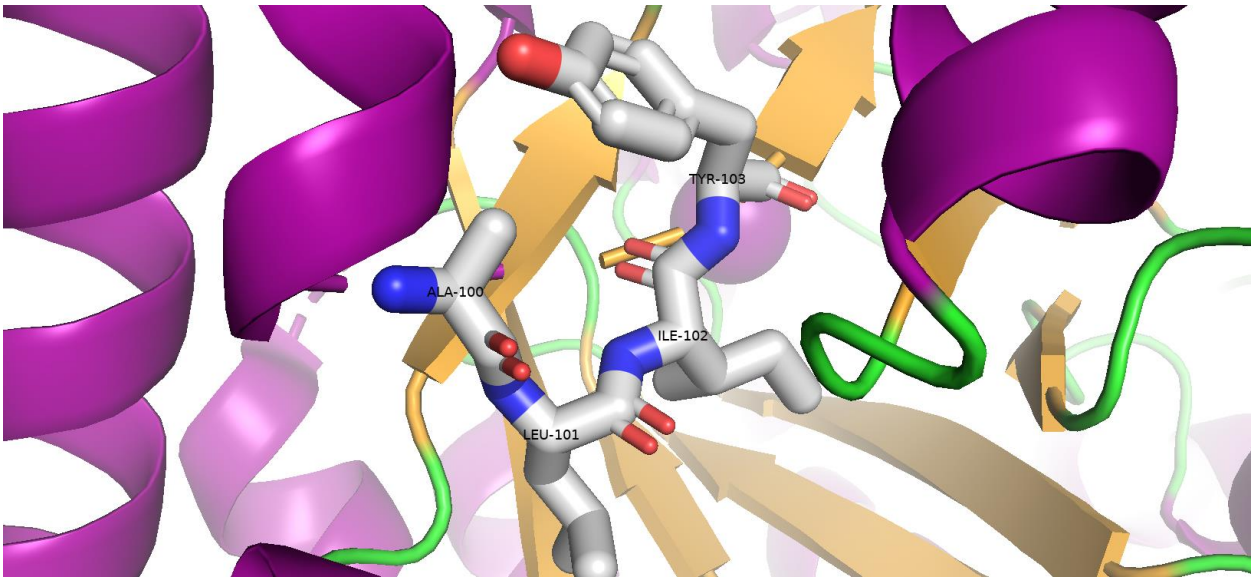


Figure 20. 2fds, β-turn type IV, involving residues 100-103 in chain B

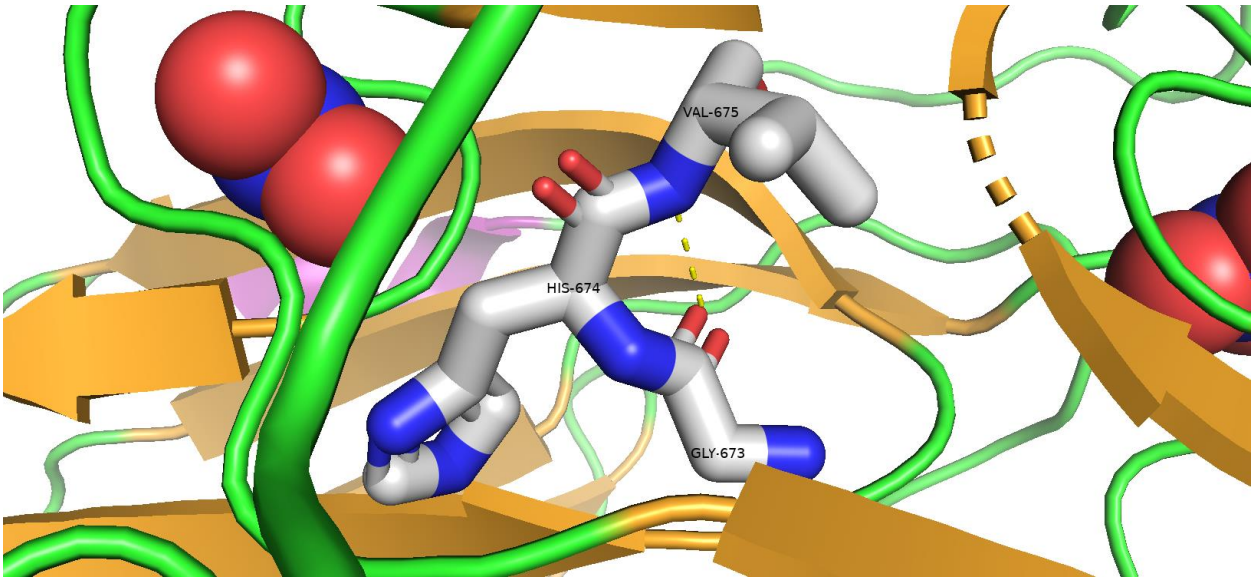


Figure 21. 2wt1, inverse γ-turn, involving residues 673-675 in chain A

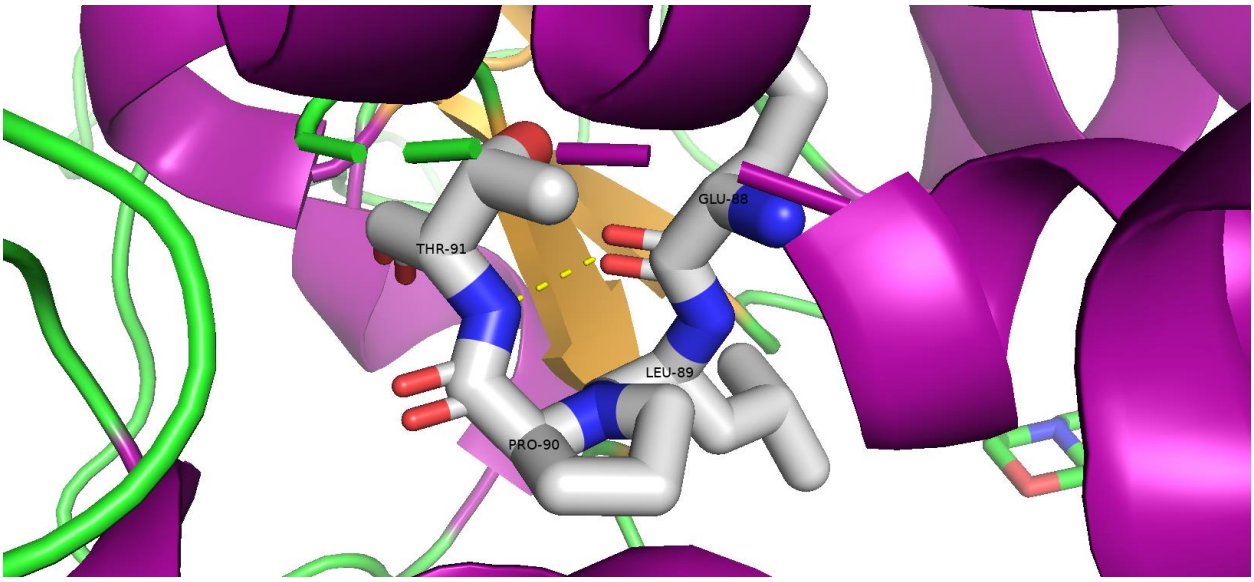


Figure 22. 3ckc, β-turn type I, involving residues 88-91 in chain A

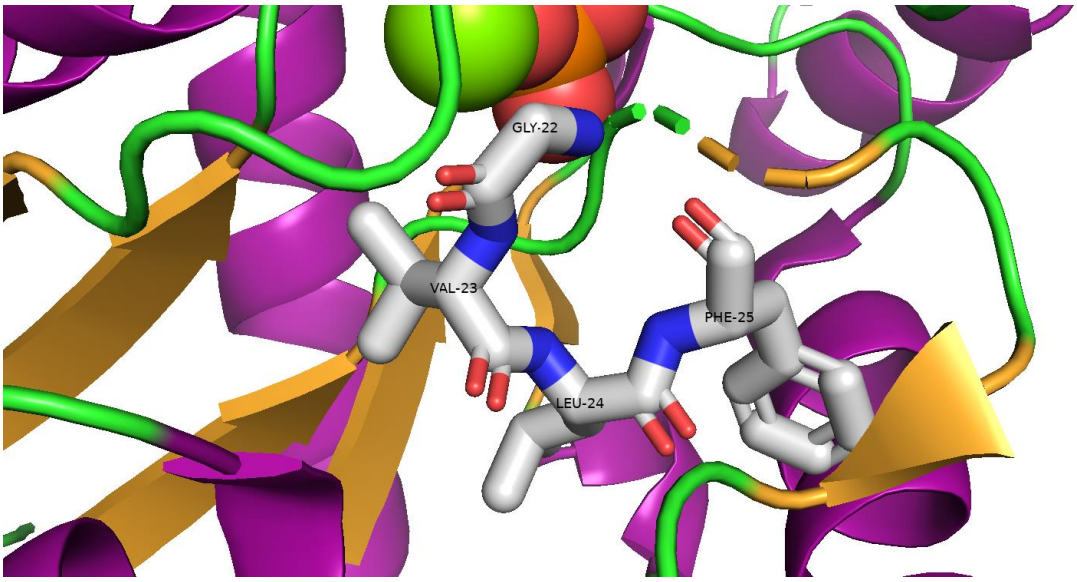


Figure 23. 3kc2, β-turn type IV, involving residues 22-25 in chain B

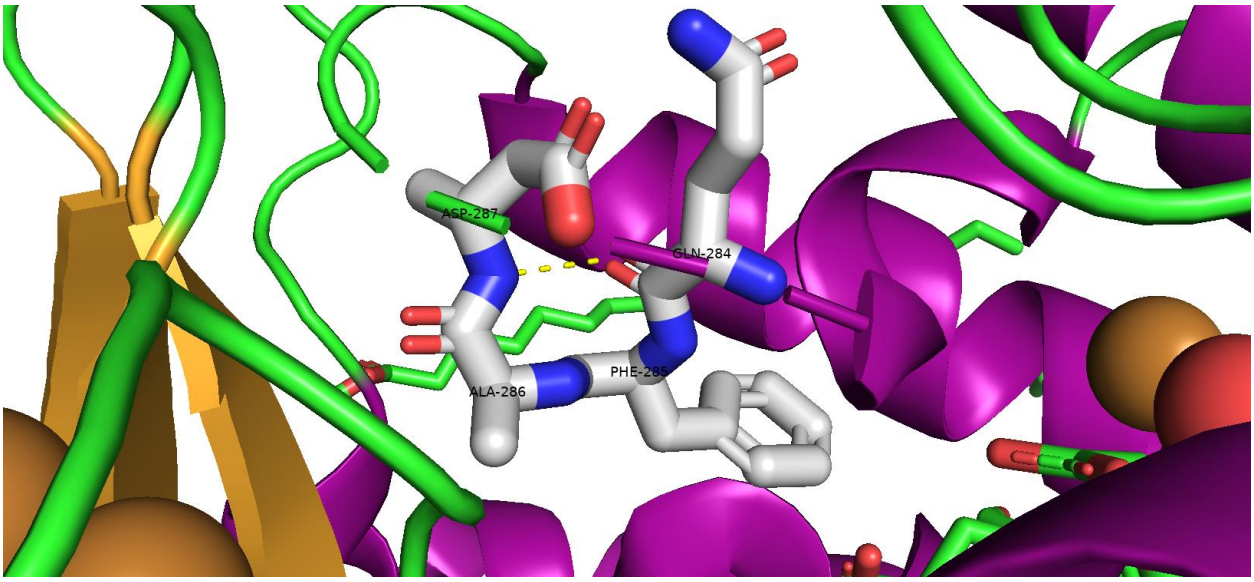


Figure 24. 3s8g, β-turn type I, involving residues 264-287 in chain A

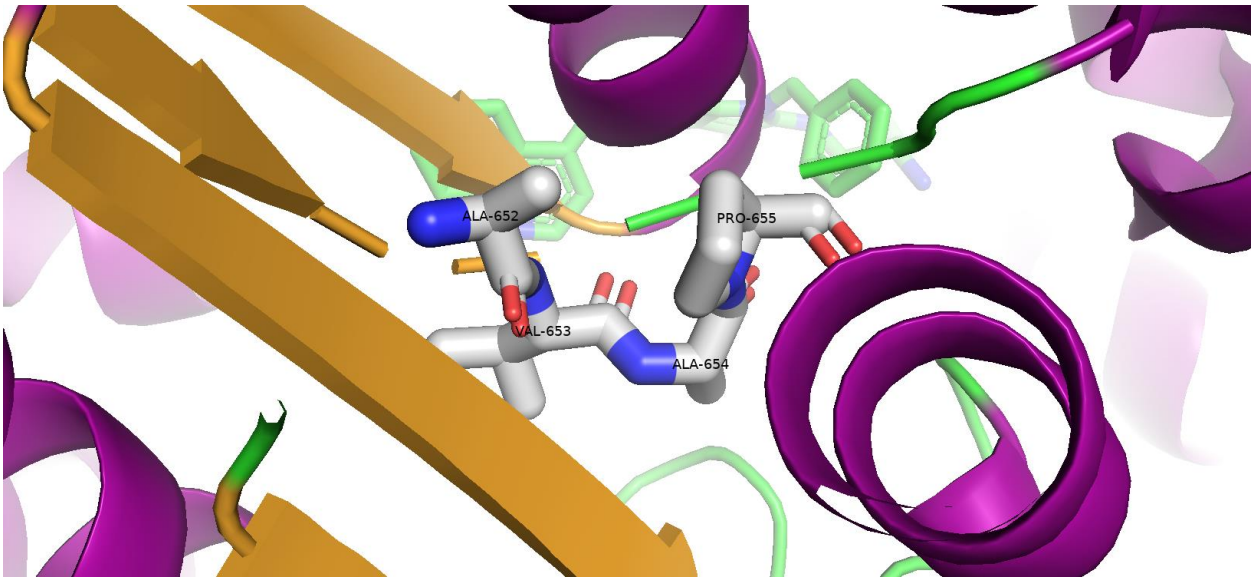


Figure 25. 4a5s, β-turn type IV, involving residues 652-655 in chain B

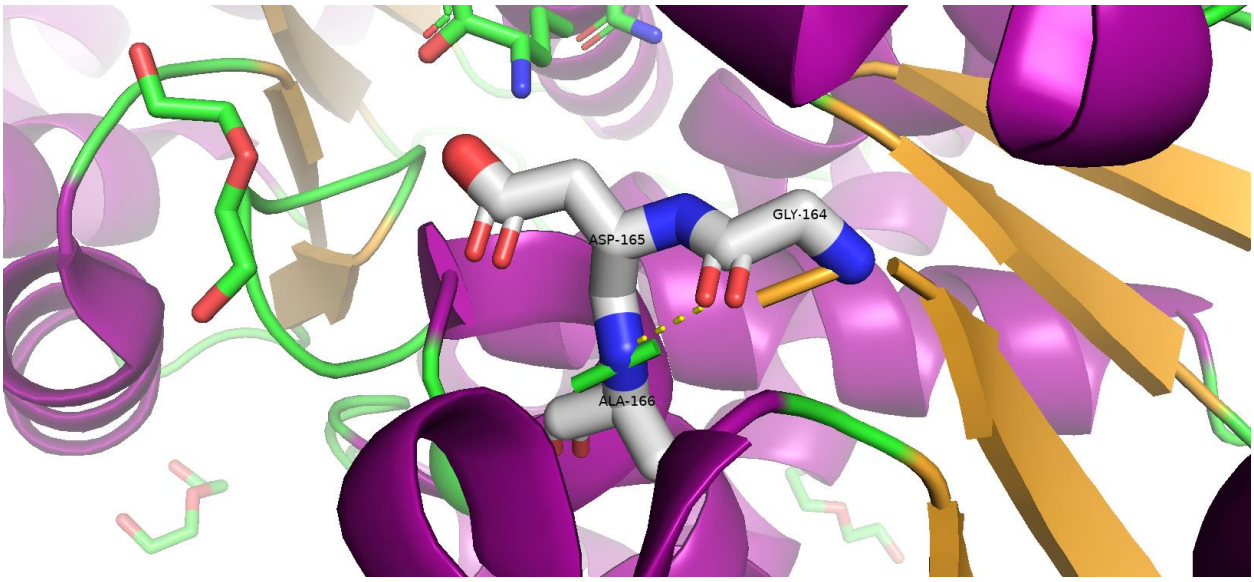


Figure 26. 4h31, inverse γ-turn, involving residues 164-166 in chain A

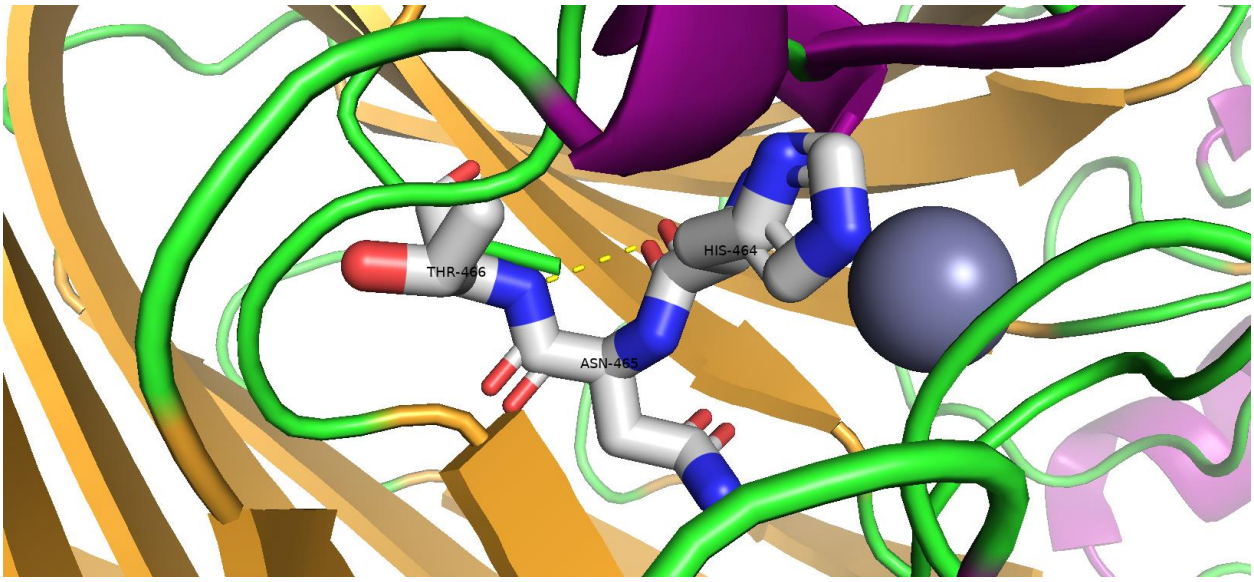


Figure 27. 4ok4, inverse γ-turn, involving residues 464-466 in chain A

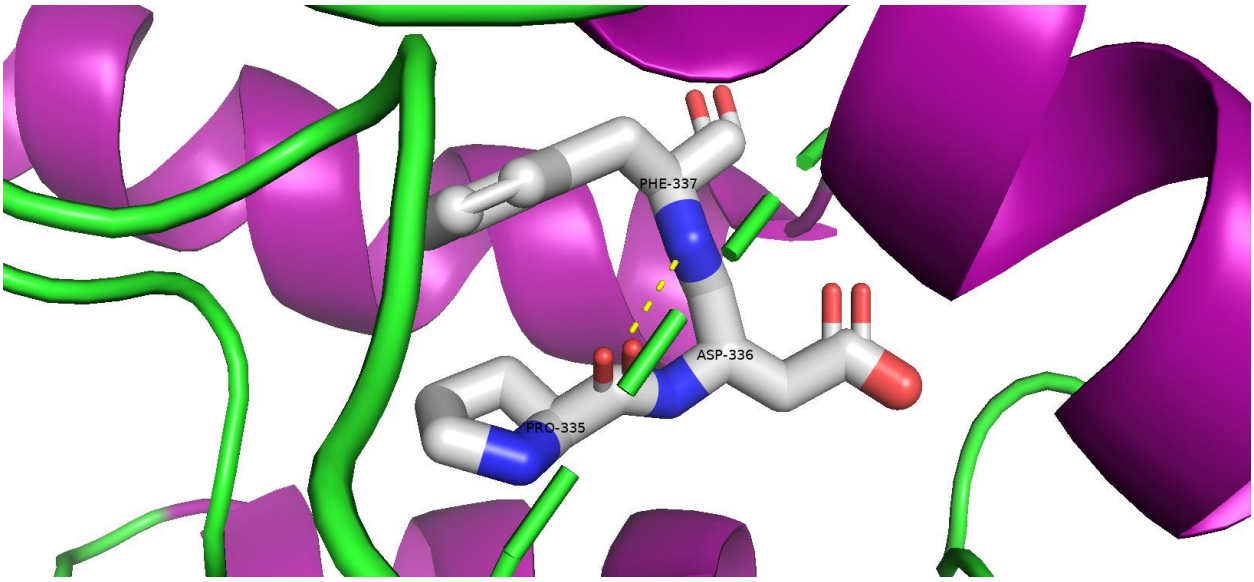


Figure 28. 4xpq, inverse γ-turn, involving residues 335-337 in chain A

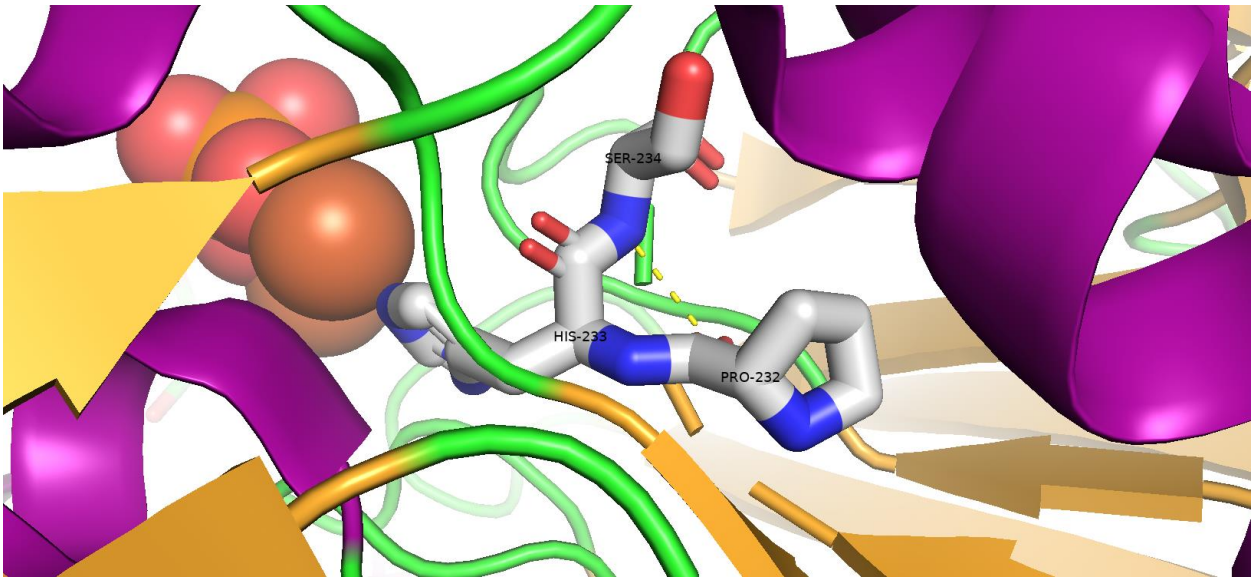


Figure 29. 5eqv, inverse γ-turn, involving residues 232-234 in chain A

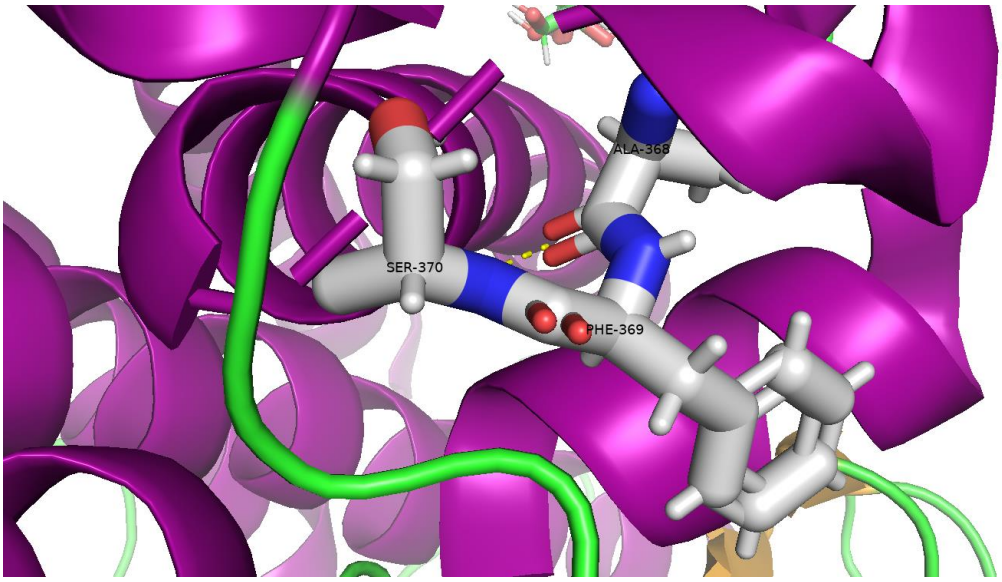


Figure 30. 5jsk, inverse γ-turn, involving residues 368-370 in chain B

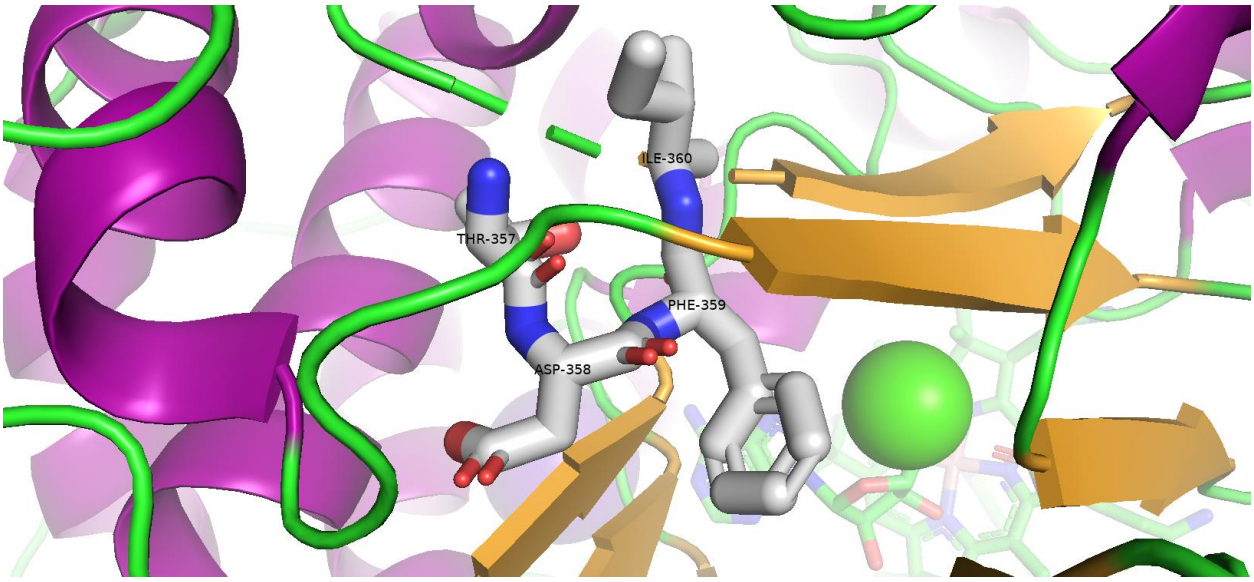


Figure 31. 5yrv, β-turn type VIII, involving residues 357-360 in chain J

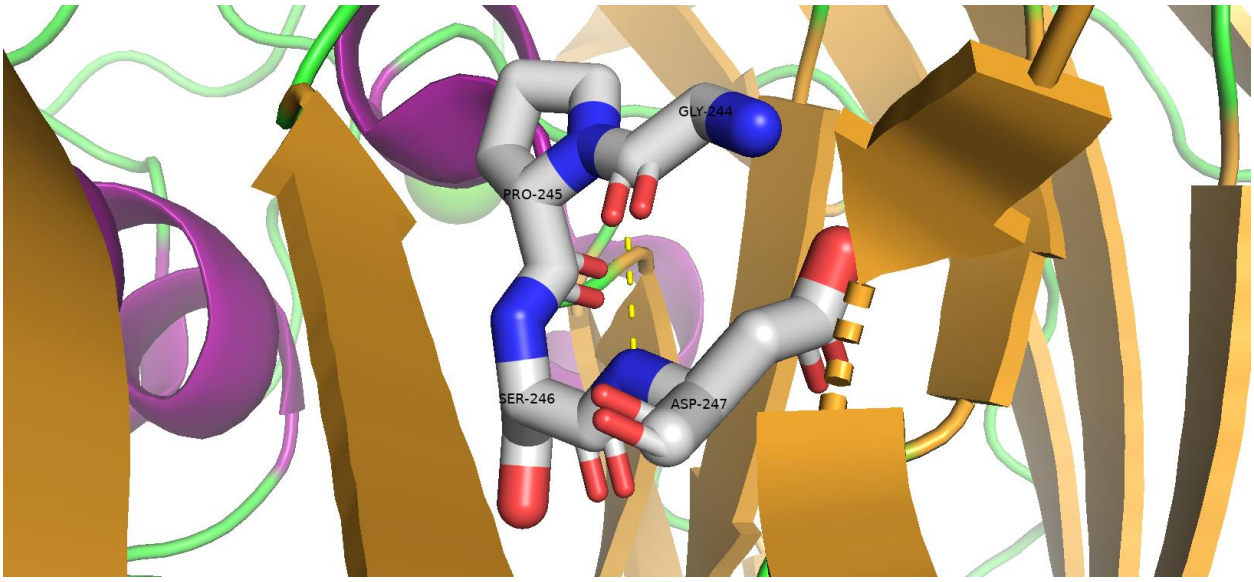


Figure 32. 6hbe, β-turn type II, involving residues 244-247 in chain A

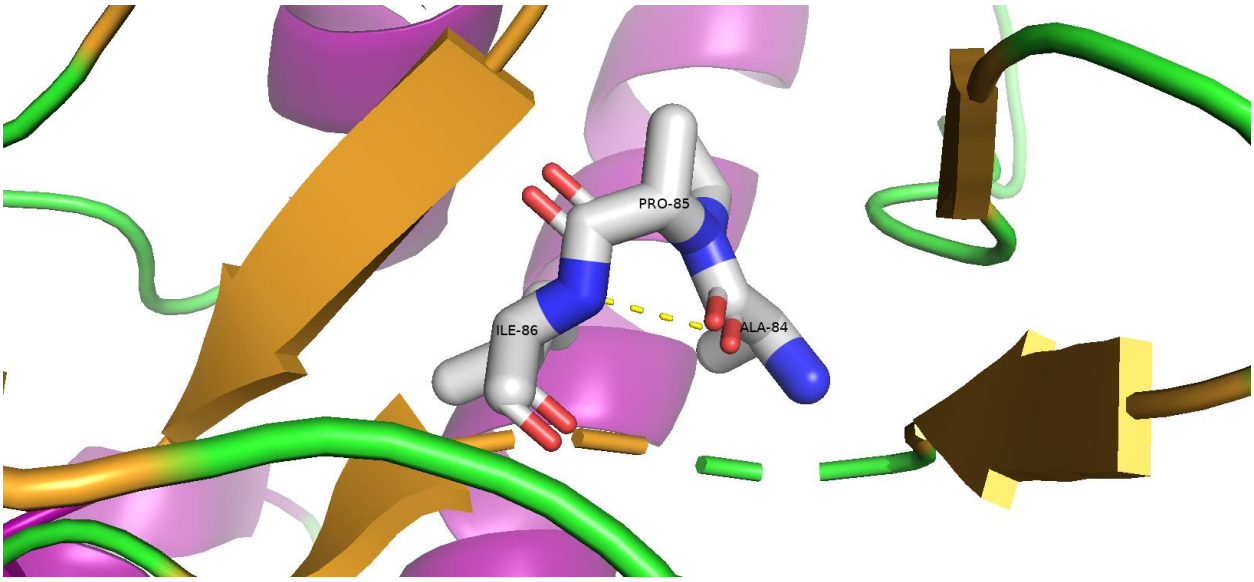


Figure 33. inverse γ-turn, involving residues 84-86 in chain A

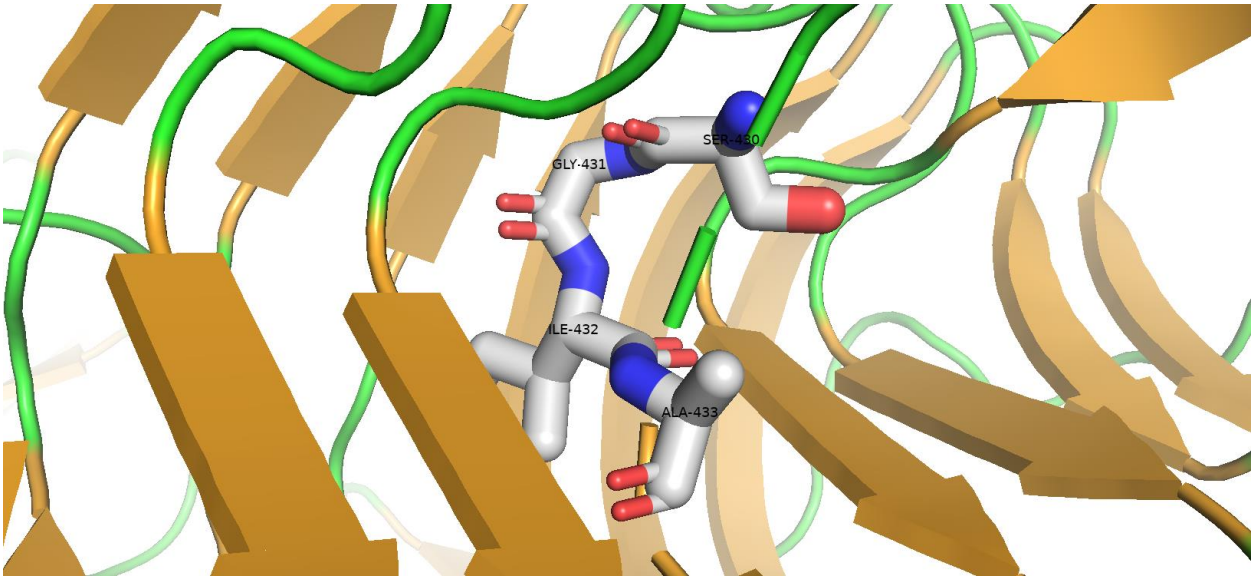


Figure 34. 7c7d, β-turn type VIII, involving residues 430-433 in chain B

3.3 Protein families

Table 2 shows the PDB IDs of the proteins found with hydrophobic turns and their classification in PDB[18],[19],[20]. Some families seem to be overrepresented. In detail, 45% of the proteins fall into the hydrolase family, 15% in the oxidoreductase, 13% in the transferase and 7% in the lyase. At the time this thesis is written, a 4% of the proteins have unknown functions. Moreover, 88% of the proteins have catalytic properties.

PDB IDs	Classification
1a4s	OXIDOREDUCTASE
1apy	HYDROLASE
1bu8	HYDROLASE
1ccw	ISOMERASE
1dl2	HYDROLASE
1duv	TRANSFERASE
1ffy	LIGASE/RNA
1fzy	LIGASE
1g9g	HYDROLASE
1gde	TRANSFERASE
1gpl	SERINE ESTERASE
1h0h	ELECTRON TRANSPORT
1hkh	HYDROLASE
1iyn	OXIDOREDUCTASE
1jc9	SUGAR BINDING PROTEIN
1jji	HYDROLASE
1jnr	OXIDOREDUCTASE
1kap	ZINC METALLOPROTEASE
1kbl	TRANSFERASE
1kjq	TRANSFERASE
1kol	OXIDOREDUCTASE
1kqf	OXIDOREDUCTASE
1llf	HYDROLASE
1m15	TRANSFERASE
1ml4	TRANSFERASE
1mzh	STRUCTURAL GENOMICS, ALDOLASE
1n62	OXIDOREDUCTASE
1nnw	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
1o20	OXIDOREDUCTASE
1odz	HYDROLASE
1oi6	ISOMERASE
1p1m	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
1px5	TRANSFERASE
1qh4	TRANSFERASE

1r88	IMMUNE SYSTEM
1ra0	HYDROLASE
1rp1	HYDROLASE
1ru4	LYASE
1t61	STRUCTURAL PROTEIN
1u7g	TRANSPORT PROTEIN
1uas	HYDROLASE
1uxo	HYDROLASE
1vr6	TRANSFERASE
1w2t	HYDROLASE
1w6s	OXIDOREDUCTASE
1wdd	LYASE
1weh	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
1xl7	TRANSFERASE
1y4j	SUGAR BINDING PROTEIN
1yiq	OXIDOREDUCTASE
1yq2	HYDROLASE
1zxx	TRANSFERASE
2a9s	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
2ad6	OXIDOREDUCTASE
2b2h	TRANSPORT PROTEIN
2bhu	HYDROLASE
2bii	OXIDOREDUCTASE
2bog	HYDROLASE
2c61	HYDROLASE
2c6q	OXIDOREDUCTASE
2czq	HYDROLASE
2d81	HYDROLASE
2dgk	LYASE
2dyu	HYDROLASE
2esb	HYDROLASE
2etj	HYDROLASE
2f48	TRANSFERASE
2fds	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
2ff4	TRANSCRIPTION
2fzw	OXIDOREDUCTASE

2gmw	HYDROLASE
2gwg	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
2h1v	LYASE
2i33	HYDROLASE
2ivf	OXIDOREDUCTASE
2j5g	HYDROLASE
2jbv	OXIDOREDUCTASE
2je6	HYDROLASE
2jhf	OXIDOREDUCTASE
2nli	OXIDOREDUCTASE
2nuw	LYASE
2p5x	CELL CYCLE
2p6w	TRANSFERASE
2pl5	TRANSFERASE
2po1	HYDROLASE/RNA
2q9o	OXIDOREDUCTASE
2q9u	OXIDOREDUCTASE
2qgy	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
2uz0	HYDROLASE
2vat	TRANSFERASE
2vch	TRANSFERASE
2wi8	TRANSPORT PROTEIN
2wt1	VIRAL PROTEIN
2x3l	LYASE
2x98	HYDROLASE
2x9g	OXIDOREDUCTASE
2xn2	HYDROLASE
2y3c	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
2y4r	LYASE
2y8n	LYASE
2yeq	HYDROLASE
2yfo	HYDROLASE
2yim	ISOMERASE
2yn0	TRANSCRIPTION
2yn2	HYDROLASE
2yyk	OXIDOREDUCTASE

2z3z	HYDROLASE
2zo6	LUMINESCENT PROTEIN
2zuy	LYASE
2zws	HYDROLASE
3a5v	HYDROLASE
3aam	HYDROLASE
3ahy	HYDROLASE
3azo	HYDROLASE
3bmx	HYDROLASE
3c9f	HYDROLASE
3cij	TRANSPORT PROTEIN
3ckc	SUGAR BINDING PROTEIN
3e0x	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
3e2d	HYDROLASE
3edy	HYDROLASE
3eqn	HYDROLASE
3erp	OXIDOREDUCTASE
3fak	HYDROLASE
3fcx	HYDROLASE
3fj1	ISOMERASE
3g5w	METAL BINDING PROTEIN
3gg7	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
3gmi	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
3gnp	HYDROLASE
3gve	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
3gyl	HYDROLASE
3hg3	HYDROLASE
3i6y	HYDROLASE
3ib7	HYDROLASE
3ik4	ISOMERASE
3ilv	LIGASE
3ima	HYDROLASE/HYDROLASE INHIBITOR
3iog	HYDROLASE
3k28	ISOMERASE/TRANSFERASE
3k2w	OXIDOREDUCTASE

3kc2	HYDROLASE
3kgb	TRANSFERASE
3l4y	HYDROLASE
3l8a	LYASE
3liu	CELL ADHESION
3lrk	HYDROLASE
3ls2	HYDROLASE
3mc1	HYDROLASE
3mk1	HYDROLASE
3n17	HYDROLASE
3n3m	LYASE
3n6z	HYDROLASE
3nqp	SUGAR BINDING PROTEIN
3o4h	HYDROLASE
3okp	TRANSFERASE
3ozy	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
3pgu	LIPID TRANSPORT
3qfm	HYDROLASE
3qpa	HYDROLASE
3qx3	ISOMERASE/DNA/ISOMERASE INHIBITOR
3rqz	HYDROLASE
3s8g	OXIDOREDUCTASE
3s9c	HYDROLASE
3sd7	HYDROLASE
3sqg	TRANSFERASE
3sqr	OXIDOREDUCTASE
3sza	OXIDOREDUCTASE
3szy	HYDROLASE
3trd	HYDROLASE
3two	OXIDOREDUCTASE
3u1j	HYDROLASE/HYDROLASE INHIBITOR
3uk0	TRANSPORT PROTEIN
3uko	OXIDOREDUCTASE
3uue	HYDROLASE
3vca	OXIDOREDUCTASE
3vsv	HYDROLASE
3w5f	HYDROLASE

3wiw	HYDROLASE
3x0d	TRANSFERASE
3x2m	HYDROLASE
4a0s	OXIDOREDUCTASE
4a35	ISOMERASE
4a4a	HYDROLASE
4a5s	HYDROLASE
4arc	LIGASE/RNA
4b6g	HYDROLASE
4bf5	ISOMERASE
4bjh	HYDROLASE/TRANSFERASE
4ccd	HYDROLASE
4ccw	HYDROLASE
4ccy	HYDROLASE
4ckk	ISOMERASE
4da2	DNA BINDING PROTEIN
4do4	HYDROLASE/HYDROLASE INHIBITOR
4ex6	HYDROLASE
4exk	TRANSPORT PROTEIN
4f32	TRANSFERASE/ANTIBIOTIC
4f4h	LIGASE
4fbc	HYDROLASE
4fkb	HYDROLASE
4fol	HYDROLASE
4g4g	HYDROLASE
4gfi	ISOMERASE
4gkv	OXIDOREDUCTASE
4gy7	HYDROLASE
4h31	TRANSFERASE
4hwv	LYASE
4i3f	HYDROLASE
4i3v	OXIDOREDUCTASE
4i7e	TRANSFERASE
4iib	HYDROLASE
4ino	TRANSPORT PROTEIN
4iqb	TRANSFERASE
4jcm	TRANSFERASE
4jjj	HYDROLASE

4k8y	HYDROLASE/HYDROLASE INHIBITOR
4ktp	TRANSFERASE
4lus	ISOMERASE
4m7h	SUGAR BINDING PROTEIN
4m7t	METAL BINDING PROTEIN
4m85	TRANSFERASE
4mr0	PLASMIN AND FIBRONECTIN BINDING PROTEIN
4my4	ISOMERASE
4nnb	LYASE
4nno	TRANSPORT PROTEIN
4nur	HYDROLASE
4nzf	HYDROLASE
4nzj	HYDROLASE
4o0c	HYDROLASE
4oju	STRUCTURAL GENOMICS, UNKNOWN FUNCTION
4ok4	LYASE
4opm	HYDROLASE
4p8v	SUGAR BINDING PROTEIN
4pzv	TRANSFERASE
4q05	HYDROLASE
4q3o	HYDROLASE
4q9w	FLUORESCENT PROTEIN
4qgk	OXIDOREDUCTASE
4qnw	OXIDOREDUCTASE
4r3n	OXIDOREDUCTASE
4rgy	HYDROLASE
4rot	HYDROLASE
4ru4	STRUCTURAL PROTEIN
4s28	LYASE
4tz1	HYDROLASE
4usa	OXIDOREDUCTASE
4uzs	HYDROLASE
4wk0	CELL ADHESION/IMMUNE SYSTEM
4wkz	STRUCTURAL PROTEIN
4wm8	VIRUS
4wpt	LYASE/TRANSFERASE

4x00	HYDROLASE
4x4a	HYDROLASE
4x8e	OXIDOREDUCTASE
4xpq	HYDROLASE
4xzw	HYDROLASE
4y2e	HYDROLASE
4yhe	HYDROLASE
4z4d	GENE REGULATION/RNA
4z9m	TRANSFERASE
4zac	LYASE
4zxo	HYDROLASE
5a71	HYDROLASE
5ah1	HYDROLASE
5aoo	VIRUS
5bza	HYDROLASE
5c1e	HYDROLASE
5cg0	HYDROLASE
5cgq	LYASE/LYASE INHIBITOR
5cm0	TRANSFERASE
5cx6	TRANSFERASE
5d7w	HYDROLASE
5dl5	MEMBRANE PROTEIN
5ehf	OXIDOREDUCTASE
5eqv	HYDROLASE
5ew0	HYDROLASE
5ey5	LYASE
5fya	HYDROLASE
5g0x	HYDROLASE
5g1a	HYDROLASE
5g2u	HYDROLASE
5g2v	HYDROLASE
5ghf	TRANSFERASE
5gkv	HYDROLASE
5gp4	LYASE
5gsm	HYDROLASE
5gt5	LYASE
5gw8	HYDROLASE
5gwn	PROTEIN BINDING

5gzk	HYDROLASE
5h3h	HYDROLASE
5h82	OXIDOREDUCTASE
5hea	TRANSFERASE
5hha	HYDROLASE
5izd	OXIDOREDUCTASE
5jpd	TRANSPORT PROTEIN
5jrh	LIGASE
5jsk	OXIDOREDUCTASE
5jxm	TRANSFERASE
5k7a	OXIDOREDUCTASE
5kxj	OXIDOREDUCTASE
5l8x	TRANSFERASE
5m9b	MEMBRANE PROTEIN
5msx	HYDROLASE
5nlm	TRANSFERASE
5o3u	HYDROLASE
5ohy	HYDROLASE
5olu	HYDROLASE
5opq	HYDROLASE
5oq3	HYDROLASE
5r61	IMMUNE SYSTEM
5uao	OXIDOREDUCTASE
5ui3	LYASE
5v2j	TRANSFERASE
5vc1	CELL ADHESION
5vn5	HYDROLASE
5wan	OXIDOREDUCTASE
5wzq	HYDROLASE
5x88	HYDROLASE
5xd7	ISOMERASE
5ypv	TRANSFERASE
5yrv	LYASE
5z0c	OXIDOREDUCTASE
5zru	HYDROLASE
6aib	STRUCTURAL PROTEIN
6ask	FLAVOPROTEIN/OXIDOREDUCTASE
6bsu	TRANSFERASE

6c02	HYDROLASE
6c9u	TRANSFERASE/IMMUNE SYSTEM
6coj	HYDROLASE
6d4b	OXIDOREDUCTASE
6d6w	HYDROLASE
6da7	LYASE
6dqh	HYDROLASE
6dxs	LYASE/LYASE INHIBITOR
6dxu	HYDROLASE
6f5z	TRANSFERASE
6f6d	OXIDOREDUCTASE
6fkx	HYDROLASE
6fmx	HYDROLASE
6g1u	HYDROLASE
6g4b	TRANSFERASE
6git	HYDROLASE
6gnf	TRANSFERASE
6hbe	OXIDOREDUCTASE
6hpf	HYDROLASE
6hwr	METAL BINDING PROTEIN
6idy	LIPID BINDING PROTEIN
6ix2	HYDROLASE
6j66	HYDROLASE
6jca	TRANSFERASE
6jd9	HYDROLASE
6jeb	HYDROLASE
6jow	HYDROLASE
6k0p	HYDROLASE
6k38	OXIDOREDUCTASE
6k5j	HYDROLASE
6ksi	HYDROLASE
6ktk	OXIDOREDUCTASE
6ljh	OXIDOREDUCTASE
6m5a	HYDROLASE
6m6l	HYDROLASE
6m74	OXIDOREDUCTASE
6mzo	HYDROLASE
6n9j	HYDROLASE/HYDROLASE INHIBITOR

6owe	OXIDOREDUCTASE
6pnr	HYDROLASE/HYDROLASE INHIBITOR
6pzl	TOXIN
6q75	HYDROLASE
6r33	LYASE
6r5t	HYDROLASE
6t40	VIRUS
6ub6	HYDROLASE
6v1r	NEUROTRANSMITTER BINDING PROTEIN
6v3w	TRANSFERASE
6vow	OXIDOREDUCTASE
6w04	HYDROLASE
6w4q	HYDROLASE
6wyn	HYDROLASE
6xic	HYDROLASE/INHIBITOR
6xu3	TRANSFERASE
6zeg	HYDROLASE
6zjs	HYDROLASE
6zti	HYDROLASE
7abn	LIGASE
7bv3	TRANSFERASE
7c7d	HYDROLASE
7cd1	SIGNALING PROTEIN
7cy3	HYDROLASE
7d5c	LIGASE/INHIBITOR
7d9e	TRANSFERASE
7jwf	HYDROLASE
7l01	OXIDOREDUCTASE/OXIDOREDUCTASE INHIBITOR
7l5i	OXIDOREDUCTASE
7lqx	HYDROLASE
7lr8	HYDROLASE

Table 2. PDB IDs of the proteins found with hydrophobic turns and their classification

3.4 Turn types

Table 3 shows the number of turn types found in the hydrophobic turns. 60% of the turns found are β-turns and 40% γ-turns. IV, the miscellaneous type, is the most common turn type found in our data (56%), followed by type I (15%), type VIII (14%) and type II (8%). Comparing these frequencies with the frequencies of β-turn types found regardless of hydrophobicity (38.2% for type I, 31.7% for type IV, 11,8% for type II, 9.8% for type VIII[10]), types IV and VIII seem to be more common in hydrophobic turns. For γ-turns, as expected, inverse type is the most common type (96%).

Turn Type per Turn			
β-turns		*γ-turns*	
I	39	CLASSIC	7
I'	4	INVERSE	163
II	21		
II'	8		
IV	144		
VIa1	1		
VIa2	0		
VIb	3		
VIII	37		

Table 3. Frequency of turn types

4. Conclusions and Discussion

To conclude, turns are secondary structural motifs where the direction of the polypeptide chain is reversed. β-turns are more commonly found than γ-turns, and involve four residues, while γ-turns involve three. As a globular protein folds, a hydrophobic area gets buried in the core, while a more hydrophilic area remains accessible by the solvent. Turns are more frequently located at the solvent-accessible area, although there have been instances of turns found inside the hydrophobic core. The primary aim for this project was to uncover the functional and structural implications of the buried hydrophobic turns. For this cause, we obtained a sample of non-redundant protein structures, searched for the motifs of interest and performed a statistical analysis on those results.

A small percentage of proteins from our sample of protein structures obtained from the PDB[18],[19],[20] contained buried β- or γ-turns. Hydrophobic β-turns were found slightly more frequently that hydrophobic γ-turns, with the predominant type for β-turns being the miscellaneous IV type and the inverse type for γ-turns, as expected with inverse being the most commonly found γ-turn type. While Rose et al.[17], only located type I or II buried β-turns, our results clearly show that type IV is the most common type for these motifs. Moreover, types IV and VIII are more prevalent in hydrophobic β-turns comparing to the general β-type frequency.

Hydrophobic buried turns were most commonly found in proteins with enzymatic properties and, in particular, in proteins belonging in the hydrolase protein family. The proteins including buried turns found by Rose et al., with three of them were homologous serine proteases, also functioned as enzymes. These results may be indicative of a preserved function associated with buried turns; however, further evaluation is necessary.

Our future intentions involve the construction of an RMSD- or TM-score-based[28] distance matrix with the aim of further reducing the structural redundancy that may be present in our data set. Furthermore, we will analyze our data in terms of the function of the proteins that we have identified with the aim of characterizing any putative functional significance of the buried turns. Additionally, we will analyze the structural context within which these turns are located with the aim of identifying any systematic trend in terms of their structural environment.

References

1. Dietzen, D. J. Amino Acids, Peptides, and Proteins. *Princ. Appl. Mol. Diagnostics* 345–380 (2018) doi:10.1016/B978-0-12-816061-9.00013-8.
2. Luo, Y. *et al.* Characterization and Analysis of Biopharmaceutical Proteins. *Sep. Sci. Technol.* **10**, 283–359 (2011).
3. Rehman, I., Farooq, M. & Botelho, S. Biochemistry, Secondary Protein Structure. *StatPearls* (2020).
4. Chou, K. C. & Blinn, J. R. Classification and prediction of β-turn types. *J. Protein Chem.* **16**, 575–595 (1997).
5. W, K. & C, S. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
6. GD, R., AR, G., GJ, L., RH, L. & MH, Z. Hydrophobicity of amino acid residues in globular proteins. *Science* **229**, 834–838 (1985).
7. CM, V. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**, 1425–1436 (1968).
8. PN, L., FA, M. & HA, S. Chain reversals in proteins. *Biochim. Biophys. Acta* **303**, 211–229 (1973).
9. Hutchinson, E. G. & Thornton, J. M. A revised set of potentials for β-turn formation in proteins. *Protein Sci.* **3**, 2207–2216 (1994).
10. AG, de B. Extension of the classical classification of β-turns. *Sci. Rep.* **6**, (2016).
11. Fang, C., Shang, Y. & Xu, D. Improving Protein Gamma-Turn Prediction Using Inception Capsule Networks. *Sci. Rep.* **8**, (2018).
12. Milner-White, E. J. Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites. *J. Mol. Biol.* **216**, 385–397 (1990).
13. Gromiha, M. M., Nagarajan, R. & Selvaraj, S. Protein Structural Bioinformatics: An Overview. *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.* **1**–**3**, 445–459 (2019).
14. SA, A., MI, H., A, I. & F, A. A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and unfolded states. *Curr. Protein Pept. Sci.* **15**, 456–476 (2014).
15. Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R. & Meiler, J. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J. Mol. Model.* **15**, 1093 (2009).
16. A, S. & JA, R. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, (1973).
17. Rose, G. D., Young, W. B. & Gierasch, L. M. Interior turns in globular proteins. *Nature* **304**, 654–657 (1983).
18. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
19. H, B., K, H., H, N. & JL, M. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**, (2007).
20. H, B., K, H. & H, N. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).
21. Bernstein, F. C. *et al.* The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542 (1977).

22. G, W. & RL, D. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).
23. Hutchinson, E. G. & Thornton, J. M. PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci.* **5**, 212 (1996).
24. Heinig, M. & Frishman, D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32**, 500–502 (2004).
25. JANERT, P. K. Gnuplot in Action SECOND EDITION. *Manning Shelter Isl.* 511 (2016).
26. gnuplot homepage. http://www.gnuplot.info/.
27. PyMOL | pymol.org. https://pymol.org/2/.
28. Y, Z. & J, S. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).

Appendix

Script 1: lc.pl

```
@ARGV == 1 || die "Usage: lc.pl filename\n";
open(INFILE, "<$ARGV[0]") || die "Cannot open input file\n";
open(OUTFILE, ">>lc.txt") || die "Cannot open output file\n";
$i = 0;
while($line = <INFILE>)
{
	if($i > 0)
	{
		$pdb_code = lc substr("$line", 0, 4);
		print OUTFILE "$pdb_code\n";
	}
	$i++;
}
close(INFILE);
close(OUTFILE);
exit(0);
```

Script 2: ftp.pl

```
use Net::FTP;

@ARGV == 1 || die "Usage: ftp.pl filename\n";
$ftp = Net::FTP->new("ftp.wwpdb.org", Debug => 0) || die "Cannot connect to ftp.wwpdb.org\n";
$ftp->login("anonymous", '-anonymous@') || die "Cannot login\n";
$ftp->passive(1);
my $fetching_directory = "/pub/pdb/data/structures/all/pdb/";
$ftp->cwd($fetching_directory);
open(INFILE, "$ARGV[0]") || die "Cannot open input file\n";
my $line;
my $file_to_transfer;
while($line = <INFILE>)
{
	chop $line;
	$file_to_transfer = "pdb$line.ent.gz";
	print "Getting $file_to_transfer\n";
	$ftp->get($file_to_transfer) || warn "Couldn't get $file_to_transfer\n", $ftp->message;
}
$ftp->quit;
close(INFILE);
exit(0);
```

Script 3: promotif_results.pl

```
use File::Find;

open(PRM, ">>promotif_results.txt");
my $motifdir = "/usr/local/promotif";
system("cp $motifdir/phipsi.mat ./phipsi.mat");
system("cp $motifdir/promotif2.prm ./promotif2.prm");
find(\&bturns, ".");
close(PRM);
exit(0);

sub bturns
{
	my $file = $_;
	if($file =~ /(\w*)(\.pdb)$/ || $file =~ /(\w*)(\.ent)$/)
	{
		$pdb = $file;
		$struc = `echo $pdb | $motifdir/p_sstruc3`;
		$turn = `echo $pdb | $motifdir/p_turn3`;
		print "$1 $struc\n";
		print "$1 $turn\n";
#If the output of p_turn3 is not empty then there is an error and there's no point continuing
		if($turn =~ /\S/)
		{
		}
		else
		{
			print PRM "$1 ";
			open(BT, "<$1.bturns");
			$i = 0;
			while(my $line = <BT>)
			{
				if($i > 1)
				{
#If the line matches the following regex, then $1 is chain ID, $2 is FirstResidue, $3 is LastResidue and $4 is bturn type
					if($line =~ /(\w?)\s*(\-?\d+\w?)\s*\D\s*\-?\d+\w?\s*\D\s*\-?\d+\w?\s*\D\s*(\-?\d+\w?)\s*\D\s(\w+[\']*)\s+/)
					{
						$cid = $1;
						$res1 = $2;
						$res4 = $3;
						$tt = $4;
						if($' =~ /\s+(Y|N)/)
						{
```

```
                                        $yn = $1;
                                }
#If there's no chain ID then assign "z"
                                if($cid eq "")
                                {
                                        $cid = "z";
                                }
                        }
                        print PRM "$cid:$res1:$res4:$yn:$tt ";
                }
                $i++;
        }
        close(BT);
        open(GT, "<$1.gturns");
        $i = 0;
        while(my $line = <GT>)
        {
                if($i > 1)
                {
#If the line matches the following regex, then $1 is chain ID, $2 is FirstResidue, $3 is LastResidue, $4 is gturn type
                        if($line =~ /(\w?)\s*(\-?\d+\w?)\s*\D\s*\-?\d+\w?\s*\D\s*(\-?\d+\w?)\s*\D\s(\w+)\s+/)
                        {
                                $cid = $1;
                                $res1 = $2;
                                $res3 = $3;
                                $tt = $4;
                                if($cid eq "")
                                {
                                        $cid = "z";
                                }
                        }
                        print PRM "$cid:$res1:$res3:",":$tt ";
                }
                $i++;
        }
        close(GT);
        print PRM "\n";
        system("rm $1.sst $1.bturns $1.gturns");
    }
  }
}
```

Script 4: stride_results.pl

```
@ARGV == 1 || die "Need input file\n";
open(PRM, "<$ARGV[0]") || die "Cannot open input\n";
open(OUT, ">>results.txt") || die "Cannot open output file\n";
while($line = <PRM>)
{
	@prm = split(' ', $line);
	$pdb = "$prm[0].ent";
	$stride = `stride $pdb`;
	print OUT "$prm[0]";
	@stride = split('\n', $stride);
	$i = 1;
	while($prm[$i] =~ /(\w+)\:(\-?\w+)\:\-?\w+\:([Y|N])*\:\w+/)
	{
		$cid = $1;
		$res1 = $2;
#This value is used to differentiate between beta and gamma turns
		$b_g = $3;
		foreach(@stride)
		{
#Information about solvent accessibility start with ASG
			if($_ =~ /^ASG/)
			{
				@stline = split(' ', $_);
				if($stline[3] eq $res1 && $stline[2] eq $cid)
				{
					$a = $stline[4];
					$b = $a + 1;
					$c = $b + 1;
					$d = $c + 1;
					$saa = $stline[9];
				}
				elsif($stline[4] == $b && $stline[2] eq $cid)
				{
					$sab = $stline[9];
				}
				elsif($stline[4] == $c && $stline[2] eq $cid)
				{
					$sac = $stline[9];
				}
				elsif($stline[4] == $d && $stline[2] eq $cid)
				{
					$sad = $stline[9];
				}
			}
```

```
                }
                if($b_g =~ /\S/)
                {
                        print OUT " $prm[$i]:$saa:$sab:$sac:$sad";
                }
                else
                {
                        print OUT " $prm[$i]:$saa:$sab:$sac";
                }
                $i++;
        }
        print OUT "\n";
}
close(PRM);
close(OUT);
exit(0);
```

Script 5: chain.pl

```
@ARGV == 2 || die "Usage: chain.pl input results.txt\n";
open(RES, "<$ARGV[1]") || die "Cannot open results.txt\n";
open(OUT, ">>results_final.txt") || die "Cannot open output file\n";
while($res = <RES>)
{
        @res = split(' ', $res);
        print OUT "$res[0]";
        open(IN, "<$ARGV[0]") || die "Cannot open input file\n";
        while($line = <IN>)
        {
                if($line =~ /(\w+\w+\w+\w+)(\w+)/)
                {
                        $name = $1;
                        $chain = $2;
                }
                if($res[0] =~ /^pdb$name/)
                {
                        foreach(@res)
                        {
                                if($_ =~ /^$chain/)
                                {
                                        print OUT " $_";
                                }
                        }
                }
        }
        close(IN);
```

```
        print OUT "\n";
}
close(RES);
close(OUT);
exit(0);
```

Script 6: format.pl

```
@ARGV == 1 || die "Usage: format.pl filename\n";
open(IN, "<$ARGV[0]") || die "Cannot open input file\n";
open(OUT, ">>format.txt") || die "Cannot open output file\n";
while($line = <IN>)
{
        if($line =~ /^(\w\w\w\w)(\w)/)
        {
                $lc = lc($1);
                print OUT "$lc"."$2\n";
        }
}
close(IN);
close(OUT);
exit(0);
```

Script 7: max.pl

```
@ARGV == 1 || die "Need input file\n";
open(IN, "<$ARGV[0]") || die "Cannot open input file\n";
open(OUT, ">>data.txt") || die "Cannot open output file\n";
while($line = <IN>)
{
        @prm = split(' ', $line);
        foreach(@prm)
        {
                if($_ =~ /\w+\:\-?\w+\:\-?\w+\:[Y|N]+\:\w+\'*\:(\d+\.\d*)\:(\d+\.\d*)\:(\d+\.\d*)\:(\d
+\.\d*)/)
                {
                        $a = $1;
                        $b = $2;
                        $c = $3;
                        $d = $4;
                        if($a >= $b && $a >= $c && $a >= $d)
                        {
                                $max = $a;
                        }
                        elsif($b >= $a && $b >= $c && $b >= $d)
                        {
```

```
                        $max = $b;
                }
                elsif($c >= $a && $c >= $b && $c >= $d)
                {
                        $max = $c;
                }
                else
                {
                        $max = $d;
                }
        }
        elsif($_ =~ /\w+\:\-?\w+\:\-?\w+\:\:\w+\:(\d+\.\d*)\:(\d+\.\d*)\:(\d+\.\d*)/)
        {
                $a = $1;
                $b = $2;
                $c = $3;
                if($a >= $b && $a >= $c)
                {
                        $max = $a;
                }
                elsif($b >= $a && $b >= $c)
                {
                        $max = $b;
                }
                else
                {
                        $max = $c;
                }
        }
        else
        {
                next;
        }
        print OUT "$max\n";
    }
}
close(IN);
close(OUT);
exit(0);
```

Script 8: histogram_data.pl

```
@ARGV == 2 || die "Need input file and bin number\n";
open(IN, "<$ARGV[0]") || die "Cannot open input file\n";
open(OUT, ">>outdata.txt") || die "Cannot open output file\n";
$max = 0.0;
```

```
$min = 100.0;
$bin = $ARGV[1];
while($line = <IN>)
{
	chomp($line);
	if($line > $max)
	{
		$max = $line;
	}
	if($line < $min)
	{
		$min = $line;
	}
}
#Changing the values into decimal
$max = $max*10;
$min = $min*10;
$bin = $bin*10;
$i = $min;
close(IN);
while($i <= $max)
{
	$k = $i + $bin;
	$a = 0;
i	open(IN, "<$ARGV[0]") || die "$!\n";
	while($line = <IN>)
	{
		chomp($line);
		$line = $line*10;
		if($line >= $i && $line < $k)
		{
			$a++;
		}
	}
	print OUT $i/10,"~",$k/10," ",$a,"\n";
	$i = $k;
	close(IN);
}
close(OUT);
exit(0);
```

Script 9: hydrophobic_turns.pl

```
@ARGV == 1 || die "Need input file\n";
open(IN, "<$ARGV[0]") || die "Cannot open input file\n";
open(OUT, ">>data.txt") || die "Cannot open output file\n";
```

```
while($line = <IN>)
{
	@prm = split(' ', $line);
	foreach(@prm)
	{
		if($_ =~ /\w+\:\-?\w+\:\-?\w+\:[Y|N]+\:\w+\'*\:(\d+\.\d*)\:(\d+\.\d*)\:(\d+\.\d*)\:(\d
+\.\d*)/)
		{
			$a = $1;
			$b = $2;
			$c = $3;
			$d = $4;
			if($a >= $b && $a >= $c && $a >= $d)
			{
				$max = $a;
			}
			elsif($b >= $a && $b >= $c && $b >= $d)
			{
				$max = $b;
			}
			elsif($c >= $a && $c >= $b && $c >= $d)
			{
				$max = $c;
			}
			else
			{
				$max = $d;
			}
		}
		elsif($_ =~ /\w+\:\-?\w+\:\-?\w+\:\:\w+\:(\d+\.\d*)\:(\d+\.\d*)\:(\d+\.\d*)/)
		{
			$a = $1;
			$b = $2;
			$c = $3;
			if($a >= $b && $a >= $c)
			{
				$max = $a;
			}
			elsif($b >= $a && $b >= $c)
			{
				$max = $b;
			}
			else
			{
				$max = $c;
			}
```

```
                }
                else
                {
                        next;
                }
                if($max < 0.2)
                {
                        print OUT "$prm[0] $_\n";
                }
        }
}
close(IN);
close(OUT);
exit(0);
```