NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

BSc THESIS

Clustering approaches for extracting structural determinants of enzyme active sites that dictate ligand binding

Konstantina I. Roka

Supervisors: **Ioannis Emiris,** Professor
Evangelia Chrysina, Senior Researcher ICB-NHRF

ATHENS

SEPTEMBER 2021

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τεχνικές ομαδοποίησης για την εξαγωγή δομικών στοιχείων που καθορίζουν την πρόσδεση στο ενεργό κέντρο του ενζύμου

Κωνσταντίνα Ι. Ρόκα

Επιβλέποντες: **Ιωάννης Εμίρης,** Καθηγητής
Ευαγγελία Χρυσίνα, Κύρια Ερευνήτρια, ΙΧΒ-ΕΙΕ

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2021

BSc THESIS

Clustering approaches for extracting structural determinants of enzyme active sites that dictate ligand binding

Konstantina I. Roka

S.N.: 1115201500139

SUPERVISORS: **Ioannis Emiris,** Professor
Evangelia Chrysina, Senior Researcher ICB-NHRF

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τεχνικές ομαδοποίησης για την εξαγωγή δομικών στοιχείων που καθορίζουν την πρόσδεση στο ενεργό κέντρο του ενζύμου

Κωνσταντίνα Ι. Ρόκα

Α.Μ.: 1115201500139

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ιωάννης Εμίρης,** Καθηγητής
Ευαγγελία Χρυσίνα, Κύρια Ερευνήτρια, ΙΧΒ-ΕΙΕ

ABSTRACT

The study of enzyme binding sites is a very important process which can be very arduous since it requires a lot of experimental work.This is why we need to use informatics and computational tools which could speed up the process. The focus of this thesis is to propose workflows by which we can depict the structure of an enzyme's active site by using computational methods and exploiting geometrical features of the enzyme. The enzyme of glycogen prosphorylase (GP), a validated target for the design of therapeutic agents for the treatment of type 2 diabetes, was selected as use case. With the aim to create a glove representation of the active site of GP, all possible conformations of the active site were artificially generated and clustered, taking into account the rotamers of the individual residues that form this site. The results obtained give new structural insights on a subsite of the catalytic site of GP that may be further exploited for the structure-based design of improved therapeutic agents accelerating the drug discovery process.

SUBJECT AREA: Bioinformatics

KEYWORDS: Active site, pocket, clustering, alpha shape, rotamers, enzyme, glycogen phosphorylase, ligand, superposition, protein-ligand interaction

ΠΕΡΙΛΗΨΗ

Η μελέτη των κέντρων πρόσδεσης των ενζύμων είναι μια πολύ σημαντική διαδικασία που μπορεί να αποδειχτεί επίπονη γιατί απαιτείται πλήθος εργαστηριακών μελετών. Για αυτό το λόγο και με σκοπό να επιταχύνουμε την έρευνα στο συγκεκριμένο τομέα, χρειάζεται να καταφύγουμε στην πληροφορική και υπολογιστικές μεθόδους. Στην εργασία αυτή επικεντρωνόμαστε στο να αποτυπώσουμε μια διαδικασία μέσω της οποίας θα μπορούμε να ορίσουμε την δομή του ενεργού κέντρου ενός ενζύμου, χρησιμοποιώντας υπολογιστικές μεθόδους και εκμεταλλευόμενοι τα γεωμετρικά χαρακτηριστικά του ενζύμου. Το ένζυμο της φωσφορυλάσης του γλυκογόνου (GP) είναι ένας αναγνωρισμένος μοριακός στόχος για τον σχεδιασμό θεραπευτικών μέσων για την αντιμετώπιση του σακχαρώδους διαβήτη τύπου 2 και για αυτόν τον λόγο επιλέχθηκε ως παράδειγμα στην παρούσα εργασία. Με στόχο τη δημιουργία μιας σχηματικής αναπαράστασης του ενεργού κέντρου-"γαντιού" της GP, δημιουργήθηκαν με υπολογιστικές μεθόδους και στη συνέχεια ομαδοποιήθηκαν, όλες οι πιθανές διαμορφώσεις του ενεργού κέντρου, λαμβάνοντας υπ' όψιν τα διαμορφομερή των επιμέρους αμινοξέων που συνθέτουν το κέντρο. Τα αποτελέσματα δίνουν νέες χρήσιμες δομικές πληροφορίες για μια υποπεριοχή του καταλυτικού κέντρου της GP, η οποία μπορεί να αξιοποιηθεί περαιτέρω για τον κατευθυνόμενο-από τη δομή-σχεδιασμό βελτιωμένων θεραπευτικών μέσων, επιταχύνοντας την διαδικασία εύρεσης φαρμάκων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ενεργό κέντρο, Κοιλότητα, Ομαδοποίηση, Διαμορφομερή, Ένζυμο, Φωσφορυλάση του γλυκογόνου, Προσδέτης, Υπέρθεση, Αλληλεπίδραση πρωτεΐνης-προσδέτη

ACKNOWLEDGEMENTS

I would like to thank my supervisors Prof. Ioannis Emiris and Dr Evangelia Chrysina for their guidance, as well as the opportunity they provided to me to work in this very exciting project. I would also like to thank Ms Ismini Stamatelou for her help, guidance and productive cooperation throughout this project.

CONTENTS

1 **INTRODUCTION** . . . 13

2 **BACKGROUND AND RELATED WORK** . . . 14

2.1 **Background** . . . 14

2.1.1 Amino acids, proteins and enzymes . . . 14

2.1.2 Ligands and Ligand-Binding sites . . . 14

2.1.3 Glycogen Phosphorylase . . . 14

2.1.4 Steric clashes . . . 15

2.1.5 Dihedral angles . . . 15

2.1.6 Superposition . . . 16

3 **DATASET CREATION** . . . 17

3.1 **Protein structure selection** . . . 17

3.1.1 PDB file selection . . . 17

3.1.2 Superposition . . . 18

3.1.3 Average protein . . . 18

3.1.4 Pdb structure with minimum and maximum active site volume . . . 18

3.2 **Amino acids and generation of rotamers** . . . 18

3.2.1 Amino acid selection . . . 18

3.2.2 Preparation of Richardson rotamers . . . 19

3.2.3 Rotamers selection . . . 19

3.3 **Conformations** . . . 20

3.3.1 Detection of steric clashes . . . 20

3.3.2 Properties and representation of final conformations . . . 21

4 **CLUSTERING** . . . 22

4.1 **Clustering Algorithm** . . . 22

4.1.1 Algorithm . . . 22

4.1.1.1 k-medoids (PAM) . . . 22

4.1.1.2 CLARA (Clustering Large Application) . . . 22

4.1.2 Metrics . . . 22

4.2 **Dataset with coordinates** . . . 23

4.2.1 Dataset formation . . . 23

4.2.2 Optimal k . . . 23

4.3 **Dataset with chi angles** . . . 27
4.3.1 Dataset formation . . . 27
4.3.2 Optimal k . . . 27
4.4 **Dataset with phi, psi, chi1 angles and minimum distances** . . . 30
4.4.1 Dataset formation . . . 30
4.4.2 Optimal k and representative pdbs . . . 30
4.4.3 Further examination . . . 31
4.5 **Intersection of representatives** . . . 31

5 **OUTPUT** . . . 33
5.1 **Output** . . . 33
5.2 **Visualization** . . . 33

6 **TOOLS** . . . 35

7 **LIMITATIONS AND ALTERNATIVE APPROACHES** . . . 36
7.1 **Limitations** . . . 36
7.2 **Alternative approaches** . . . 36
7.3 **Assumptions** . . . 36

8 **CONCLUSION AND FUTURE WORK** . . . 37
8.1 **Conclusion** . . . 37
8.2 **Future work** . . . 38

ABBREVIATIONS - ACRONYMS . . . 39

APPENDICES . . . 39

A **ROTAMERS' APPEARANCE FREQUENCY** . . . 40

B **STERIC CLASHES BETWEEN ROTAMERS (SUPERPOSITION ON AVERAGE PROTEIN)** . . . 42

REFERENCES . . . 48

LIST OF FIGURES

2.1 Dihedral angles 16

4.1 Clustering on coord dataset superposed on average structure 1. 24
4.2 Clustering on coord dataset superposed on average structure 2. 24
4.3 Clustering on coord dataset superposed on maximum volume 1. 25
4.4 Clustering on coord dataset superposed on maximum volume 2. 25
4.5 Clustering on coord dataset superposed on minimum volume 1. 26
4.6 Clustering on coord dataset superposed on minimum volume 2. 26
4.7 Clustering on angles dataset superposed on average 1. 27
4.8 Clustering on angles dataset superposed on average structure 2. 28
4.9 Clustering on angles dataset superposed on maximum volume 1. 28
4.10 Clustering on angles dataset superposed on maximum volume 2. 29
4.11 Clustering on angles dataset superposed on minimum volume 1. 29
4.12 Clustering on angles dataset superposed on minimum volume 2. 30
4.13 Clustering on ensemble 31
4.14 Intersection of clustering representatives 32

5.1 Alpha shape of active site of 2G9R.pdb 34
5.2 Alpha shape of active site of 4gpb.pdb 34

8.1 Workflow 1 37
8.2 Workflow 2 38

A.1 Appearance frequency of rotamers in the ensemble 41

B.1 Steric Clashes between asn282 and others 43
B.2 Steric Clashes between his341 and others 44
B.3 Steric Clashes between arg292 and others 44
B.4 Steric Clashes between asp339 and others 45
B.5 Steric Clashes between asn282 and others 45
B.6 Steric Clashes between thr378 and others 46
B.7 Steric Clashes between glu 88 and others 46
B.8 Steric Clashes between asp283 and others 46
B.9 Steric Clashes between phe285 and others 47
B.10 Steric Clashes between leu136 and others 47

LIST OF TABLES

3.1 Removed rotamers causing steric clashes (dataset superposed on average structure) 20
3.2 Removed rotamers causing steric clashes (dataset superposed on maximum structure) 21
3.3 Removed rotamers causing steric clashes (dataset superposed on minimum structure) 21

PREFACE

This thesis was carried out as a part of the requirements for the acquisition of a Bachelor's degree in the department of Informatics and Telecommunications of the National and Kapodistrian University of Athens. The duration of the project was two academic semesters, under the supervision of Professor Ioannis Emiris and Senior Researcher at the Institute of Chemical Biology, National Hellenic Research Foundation (ICB-NHRF) Dr Evangelia Chrysina.

1. INTRODUCTION

The study of enzyme binding sites is a very important process which can be very arduous since it requires a lot of experimental work. Knowing the structure of an active site is very important, in order to properly introduce the molecules that can bind to this protein and alter its function in the desired way. This knowledge can be used for accelerating the research in the development of new specific drugs for the molecular target.

The process of determining a protein structure, from an experimental point of view, is very time consuming not to mention costly and difficult. As time passes, more new proteins are identified and sequenced and the need to find a way to determine their structure in a more efficient way increases.

This is the reason why new computational methods are introduced and computer science plays a huge part in the research in this area.

The focus of this thesis is to propose workflows by which we can map the structure of an enzyme's active site by using computational methods and exploiting geometrical features of the enzyme.

Using the enzyme of glycogen phopshorylase (GP), a validated target for the design of therapeutic agents for the treatment of type 2 diabetes, all possible conformations of the active site are going to be artificially created and clustered with the aim to create a glove representation of its active site.

The process of creating artificial conformations and acquiring the glove of the representatives is going to be explained thoroughly in the rest of the thesis.

The results aim to offer new structural insights of the catalytic site, that may be further exploited for the structure-based design of improved therapeutic agents, accelerating the drug discovery process.

2. BACKGROUND AND RELATED WORK

2.1 Background

2.1.1 Amino acids, proteins and enzymes

Amino acids are simple molecules consisting of an amine group and an acidic group. Each amino acid consists of two parts, the backbone and the side chain. The backbone contains two Carbons (C and C_a), Nitrogen (N) and Oxygen (O). The side chain differs in each amino acid and is the most agile because it can change place in space with less energy than the backbone. When the side chain of an amino acid moves in space, we have a different rotamer for the aminoacid. Both backbone and side chain contain hydrogens which, in this work, are not taken into consideration.

When we combine 150 or more amino acids, we have new structures which we call proteins and are very important for all physiological functions. Proteins play a key role as the functional and structural basis for cells. The proteins the function of which is to regulate biochemical reactions are called enzymes. [3]

In nature, we can come across more than 100 amino acids, but only 20 of them appear in protein structures. Individual amino acids and proteins have many features, like energy, bonds and polarity [7]. In the context of the current work we shall focus only on some geometric characteristics such as dihedral angles, coordinates and distances of the atoms.

2.1.2 Ligands and Ligand-Binding sites

Once the amino acids combine and fold giving proteins a shape, they acquire some properties and can execute specific procssec. The alteration of this shape can modify its function drastically.

Enzymes have some regions called ligand-binding sites where ligands bind. The ligand binding site aiding in the catalysis of a biochemical reaction is called the active site of the enzyme.

In biochemistry and pharmacology, a ligand is a substance that forms a complex with a biomolecule to serve a biological purpose. In protein-ligand binding, the ligand is usually a molecule which produces a signal by binding to a site on a target protein. The binding typically results in a change of conformational isomerism (conformation) of the target protein. [2]

The focus of our work lies in the mapping of the space that may be exploited for drug design, taking into consideration all possible conformations of the binding site.

2.1.3 Glycogen Phosphorylase

Glycogen Phosphorylase is a very important enzyme which is responsible for the catalysis of the phosphorolytic breakdown of glycogen, which is eventually converted to glucose and is releases in the blood stream tissues. Isoforms appear in many places in our bodies, like brain, liver and muscles and their active sites show homology in the sequence at a rate of 80-100%. The enzyme is allosteric and exists in at least two states, an active and an

inactive one. The entrance to the catalytic site of the enzyme is blocked by a loop region, residues 282-287 (280's loop) which is rather flexible and acts as a toll gate when the enzyme is in the active (open) or inactive (closed) state. The conformation of the 280's loop will be used as selection criterion for the dataset creation.[5]

2.1.4 Steric clashes

Proteins consist of a combination of many aminoacids. These amino acids are connected and are very close in 3D space. When two non bonding atoms overlap in a protein structure a steric clash is formed.

2.1.5 Dihedral angles

One way to describe the position of the amino acids in 3D-space is to use the dihedral angles between amino acids in a conformations. Dihedral angles can be used to describe both backbone and side chain atoms. The angles describing the backbone are phi and psi and refer to C_{i-1},N_i,CA_i,C_i and N_i,CA_i,C_i,N_{i+1} atoms respectively, when the amino acid we currently examine is i, i-1 is the one in the previous position and i+1 is the one in the next position. The angles describing the side chain are called chi angles and depend on the atoms comprising the chain. Below we can see the atoms comprising the phi, psi, chi angles of selected amino acids as well as a figure describing the backbone angles:[1]

- **Arg:**
  - chi1: N-CA-CB-CG,
  - chi2: CA-CB-CG-CD,
  - chi3: CB-CG-CD-NE,
  - chi4: CG-CD-NE-CZ
  - chi5: CD-NE-CZ-NH1
- **Asn, Asp:**
  - ch1: N-CA-CB-CG,
  - chi2: CA-CB-CG-OD1
- **Glu:**
  - ch1: N-CA-CB-CG,
  - chi2: CA-CB-CG-CD,
  - chi3: CB-CG-CD-OE1
- **His, Phe, Leu:**
  - ch1: N-CA-CB-CG,
  - chi2: CA-CB-CG-ND1
- **Thr:**
  - chi1: N-CA-CB-OG1

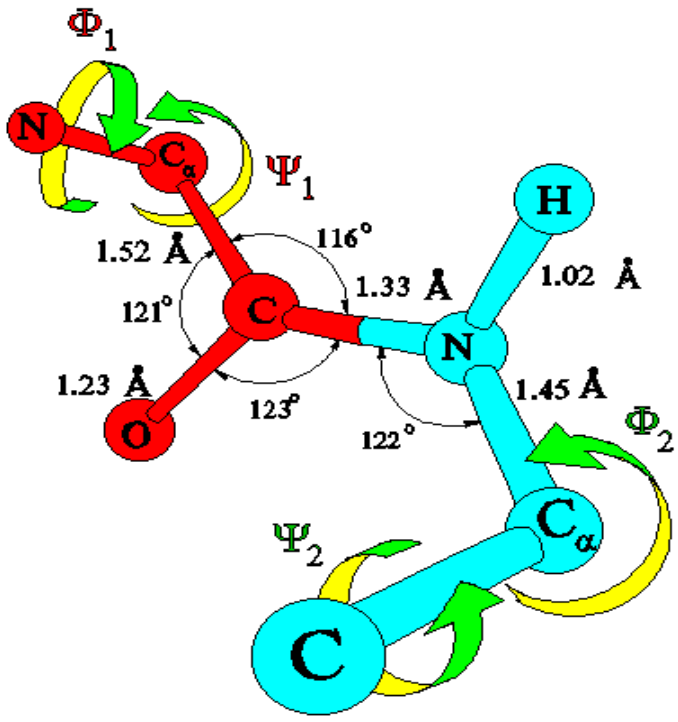


Figure 2.1: Phi(Φ) and psi(Ψ) angles in an amino acid[1]

2.1.6 Superposition

The exact same amino acid can appear in the 3D space in different positions. If its co-ordinates were to be compared, it would not be obvious that we are looking at the same amino acids. For that reason, before any checks for similarity or any other actions, a superposition must occur. Superposition is the process by which the coordinated of a protein are trasformed so that its backbone lies on top of the backbone of a reference protein.

3. DATASET CREATION

3.1 Protein structure selection

3.1.1 PDB file selection

Glycogen phosphorylase exists in three isoforms and is locate in differenct tissues such as liver, muscle and brain. We focused on GP found in rabbits (scientific name: Oryctolagus cuniculus), the structure of which has been acquired by X-ray diffraction. The files resulting from this query, contain pdbs which present a dissimilar conformation in the 280's loop compared to the rest of the files. These pdbs refer to instances where the enzyme was crystallized as tetramer and are removed from the ensemble. So, the query is refined to produce only the T-state glycogen phosphorylase structures in complex with small molecules (ligands) This crystallizes in the tetragonal lattice as a monomer and in the biological assembly (in solution) it is a dimer. The list generated by the search also includes some additional pdb files that appear because glycogen phosphorylase is their counterpart. By removing these files and some other files with missing residues (2G9Q), we end up with 90 different pdbs, as follows:

- 1E1Y
- 6S4R
- 1FU8
- 1FS4
- 3L7C
- 1K08
- 6S52
- 1Z6Q
- 1P29
- 1WW3
- 1BX3
- 1H5U
- 2GPN
- 5OX3
- 1P4G
- 2G9R
- 1WUY
- 1AXR
- 4YUA
- 5OX4
- 3L7A
- 1FTY
- 1WW2
- 5MEM
- 1FTW
- 2FET
- 4CTO
- 1C50
- 1XKX
- 5GPB
- 6S51
- 2IEG
- 1Z62
- 2FF5
- 5MCB
- 2PRI
- 3L7D
- 1A8I
- 1FU7
- 1P2D
- 1Z6P
- 5OX1
- 6S4K
- 1NOJ
- 4GPB
- 1K06
- 5OWY
- 2IEI
- 5LRC
- 1KTI
- 6GPB
- 1XL0
- 1GGN
- 3L79
- 3GPB
- 1WV0
- 1HLF
- 1P2G
- 2FFR
- 8GPB
- 1XL1
- 1XC7
- 3L7B
- 1UZU
- 6S4P
- 1FU4
- 1LWN
- 1GPY
- 1B4D
- 1WV1
- 5LRF
- 1C8K

- 2GPB
- 4Z5X
- 3EBP
- 4CTM
- 5LRE
- 1GG8
- 5OWZ
- 1P4J
- 5OX0
- 3ZCR
- 5O50
- 2PRJ
- 1FTQ
- 1WUT
- 1NOK
- 1LWO
- 1GPB
- 1P2B

3.1.2 Superposition

Having created our ensemble we have to superpose the pdbs. The pdb used as the base of the superposition is 5MEM and the zones selected were:

- 24-249
- 261-281
- 289-313
- 326-549
- 558-830

The zone selection was based on the need to exclude all flexible and missing regions. Superposition was performed over CA atoms of chain A and using CCP4 [9].

3.1.3 Average protein

In preparation for the rest of our work we had to reduce our ensemble to one protein structure. To achieve that, out of the main chain atoms, the average coordinates were calculated using the N, C, CA atoms of all the pdbs, of specific amino acids.

3.1.4 Pdb structure with minimum and maximum active site volume

Mapping the full space in which the active site of our enzyme could reside, demanded the identification of the pdbs of our ensemble which correspond the maximum and minimum active site volume. CASTp 3.0[13] was chosen for this process and we found out that 2G9R has the minimum active site volume (287.524Å), while 4gpb has the maximum active site volume (3934.462Å).

It is important to mention that the calculated volume is not the volume of the molecule, but the available space inside the glove, which is what CASTp calculates.

The data given to CASTp for calculation were pdbs containing only the amino acids of the active site which are specified subsequently.

3.2 Amino acids and generation of rotamers

3.2.1 Amino acid selection

For the construction of the pocket, only those residues known to form the active site were chosen. The selection criterion was the direct or water-mediated interactions with the

ligand atoms presented in Mamais et al. [8] other than glucose. The selected amino acids were Glu88, Leu136, Asn282, Asp283, Asn284, Phe285, Arg292, Asp339, His341, His377, and Thr378. This group of aminoacids will be referred to as glove.

3.2.2 Preparation of Richardson rotamers

The source of our initial rotamers is Richardson library. [11]. After we acquired the required protein and the rotamers for each amino acid, we had to bring the rotamers as close as possible to the protein backbone. For each selected amino acid of the binding site, we superposed the rotamers provided by Richardson, using the Ca, C and N atoms of the selected protein structures (average, minimum and maximum volume) in the respective position, creating three sets of rotamers. It is important to note that while some amino acids are repeated in the sequence (ex. Asn282, Asn284), their final coordinates differ because of the difference in configuration of the proteins. We also removed the hydrogens from our files, because they average resolution that the structures were determined was around 2Å.

3.2.3 Rotamers selection

Richardson library[11] contains the following number of rotamers for each amino acid:

- Arg: 28
- Asn: 7
- Asp: 5
- Glu: 7
- His: 8
- Leu: 4
- Phe: 4
- Thr: 3

If we wanted to produce all the possible conformations, we would have in hand 737587200 conformations. With the aim to reduce the number of rotamers, the frequency that they appear in the dataset was examined, as follows.

Algorithm 1 Evaluation of the frequency that the rotamers appear in the dataset with GP structures

```
for each of the 11 amino acids do
    for each of the 90 pdb files do
        Find which rotamer is closest to the respective amino acid of the pdb,
        after temporary superposition
        For the closest rotamer, rotamer_i_appearance += 1
    end for
    Divide with 90, to find the appearance_frequency
    Remove all rotamers with appearance_frequency == 0
end for
```

After this process the number of remaining rotamers for each amino acid is as follows (same for all 3 superposition structures):

- Arg292: 4 (24 removed)
- Asn282: 5 (2 removed)
- Asn284: 5 (2 removed)
- Asp283: 4 (1 removed)
- Asp339: 2 (3 removed)
- Glu88: 2 (5 removed)
- His341: 1 (7 removed)
- His377: 2 (6 removed)
- Leu136: 3 (1 removed)
- Phe285: 4 (0 removed)
- Thr378: 3 (0 removed)

As we can see in the corresponding appendix, the same amino acid in different position, is represented by different rotamers.

As expected, the most frequent rotamers remain the same regardless the choice of the base structure.

After the reduction the number of possible conformations is 115200. As we can see, the total number is significantly reduced. Depending on our computational power we can choose to reduce this number further by removing rotamers with appearance frequency smaller than a threshold (ex. 2%) or increase it by calculating the n closest rotamers (ex. the 2 closest each time).

One possible deduction is that His in position 341 does not play a huge part in the variation of the glove but this can be shot down if we choose to maintain more rotamers.

3.3 Conformations

3.3.1 Detection of steric clashes

Amino acids occupy space. Alternating their position in space by choosing different rotamers can result in steric clashes. That is why we carried out a throughout check to detect and eliminate these clashes. We started by checking pairs of rotamers and then we created the conformations containing all 11 amino acids. In the corresponding appendix we include matrices showing which rotamers are safe to use together as pairs. The safe pairs are shown with ”True” value and the unsafe with ”False”.
(Example of steric clash check result for superposition on average protein, in appendix ii)

From these calculations we deducted some rotamers that are invalid, no matter the conformations, observations concuring with the notes in ”The penultimate rotamer library” [11].

Table 3.1: Removed rotamers causing steric clashes (dataset superposed on average structure)

Amino acid combination which clashes	Amino acid with the clashing Rotamer	Discarded rotamer library id
Asn282-Phe285	Phe285	4
Asn284-Asp283	Asn282	6
Asn284-Phe285	Phe285	1

Table 3.2: Removed rotamers causing steric clashes (dataset superposed on maximum structure)

Amino acid combination which clashes	Amino acid with the clashing Rotamer	Discarded rotamer library id
Asn282-Phe285	Phe285	4

Table 3.3: Removed rotamers causing steric clashes (dataset superposed on minimum structure)

Amino acid combination which clashes	Amino acid with the clashing Rotamer	Discarded rotamer library id
Asn282-Phe285	Phe285	4
His341-Asn284	Asn284	4

The total number of possible gloves after our reductions equals **46080** for superposition on the average protein, **72000** for the maximum volume structure, and **63360** for the minimum volume structure.

Regarding the results for the superposition on average protein, we can observe that, besides the rotamers which clash with everything, there are no other conformations that we can deem as invalid.

3.3.2 Properties and representation of final conformations

Following the described workflow we end up with all valid combinations of the most common rotamers, without steric clashes. Each conformation has a key value describing it which comprises the rotamer number of each amino acid in the order they appear in the glove sequence. The rotamer number maintains the numeration of ”The penultimate rotamer library” for easier identification.

4. CLUSTERING

4.1 Clustering Algorithm

4.1.1 Algorithm

4.1.1.1 k-medoids (PAM)

The algorithm we wanted to use for clustering was k-medoids and is given below. We did not prefer k-means because we wanted our centers to be part of the dataset.

Algorithm 2 k-medoids

```
1: Select k random points from the dataset as medoids
2: Assign each point of the dataset to the closest medoid
3: do
4:     for Each cluster do
5:         if A data point of the cluster, different from the medoid, decreases
6:         the cost function then
7:             Assign the data point as the new medoid
8:         end if
9:     end for
10: while The medoids change
```

The algorithm calculates the cost by finding the euclidean distances between the points and the medoid of each cluster.

4.1.1.2 CLARA (Clustering Large Application)

CLARA is a variation of PAM algorithm. As previously shown, our dataset can be very large, depending on our filters in rotamer selection. For this reason we need to reduce the computer time and RAM storage needs and that is what CLARA does.

Algorithm 3 CLARA

```
1: Select multiple fixed-size subsets of the original datase
2: for Each subset do
3:     Apply PAM and find the optimal medoids
4:     Assign each point of the original dataset to the closest medoid
5:     Calculate the cost function
6: end for
7: Choose the medoids of the partition with the best score
```

4.1.2 Metrics

We chose to use CLARA for the clustering process, but we still don't know what is the suitable number of clusters (k) for our datasets. Aiming to find the best k, we used silhouette.

Silhouette is a metric which takes values in range [-1,1] and calculates how close are the points of a cluster to the points of the neighboring clusters. Higher silhouette score implies greater distance from the other clusters and consequently a good clustering.

Algorithm 4 Silhouette

1: **for** Each point i of the dataset **do**
2: We calculate the mean distance $a(i)$ from the other points in the same cluster as i
3: **for** Each cluster of which i is not a member **do**
4: We calculate the mean distance between point i and the points of the cluster
5: Keep the minimum mean distance $b(i)$
6: **end for**
7: Calculate $s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$
8: **end for**
9: Calculate mean $s(i)$

s(i) represents the silhouette score of one specific point and mean s(i) for all data points represents the silhouette score of the dataset.

While searching for the optimal k in the big datasets, we tested a wide range of numbers (0-900) with step 100. Then we chose the optimal by testing with step 10 in a range around the best value of each dataset given by the previous test.

When the silhouette score is similar for two or more k values, the smallest value is chosen as optimal, in an attempt to minimize the number of representatives.

4.2 Dataset with coordinates

4.2.1 Dataset formation

Our first choice was to represent the glove using the coordinates of the amino acids. It is important to note that with this approach, the coordinates of atoms of the backbone do not have to be considered because they are the same across the rotamers.

Each entry of the dataset consists of the conformation id, as described before, and the coordinates for all the atoms of the side chain for each amino acid in the glove.

4.2.2 Optimal k

The process of clustering was repeated for many different k in each of the three datasets and the optimal k value was chosen. As we can see in the plot below, the optimal k was different in each dataset.

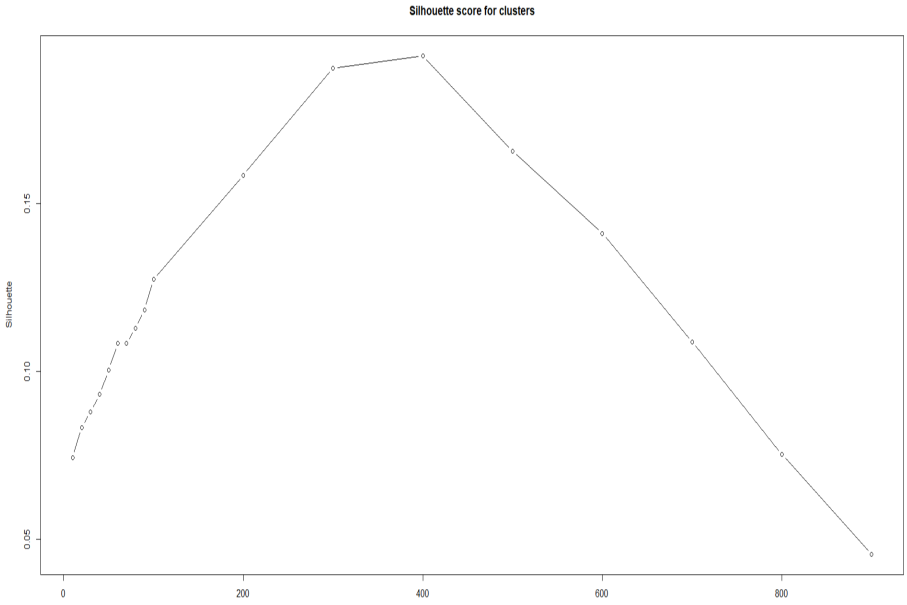


Figure 4.1: Silhouette score for different k in clustering of coordinates dataset for amino acids superposed on the average structure. k tested with step 100.[1]

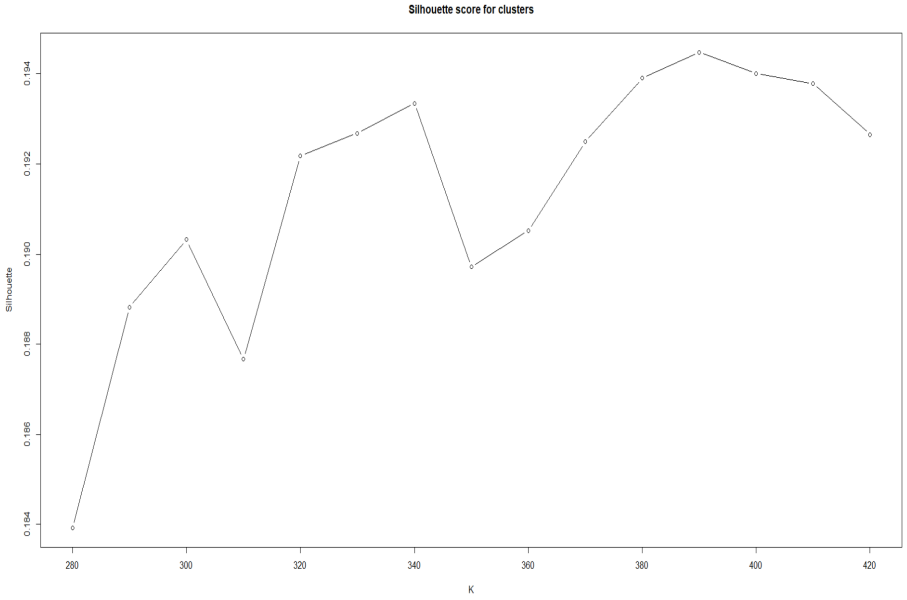


Figure 4.2: Silhouette score for different k in clustering of coordinates dataset for amino acids superposed on the average structure. k tested with step 10. [2]

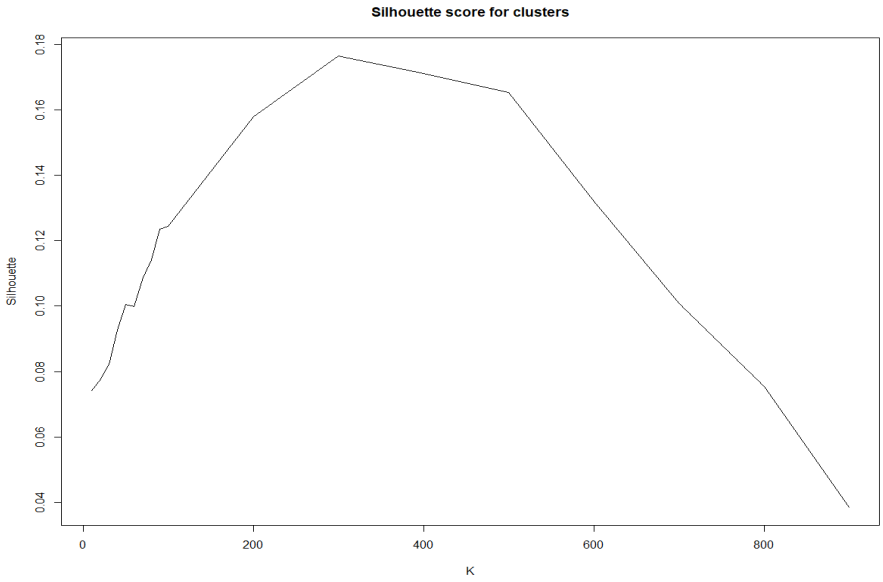


Figure 4.3: Silhouette score for different k in clustering of coordinates dataset for amino acids superposed on the pdb structure of maximum volume. k tested with step 100.[3]

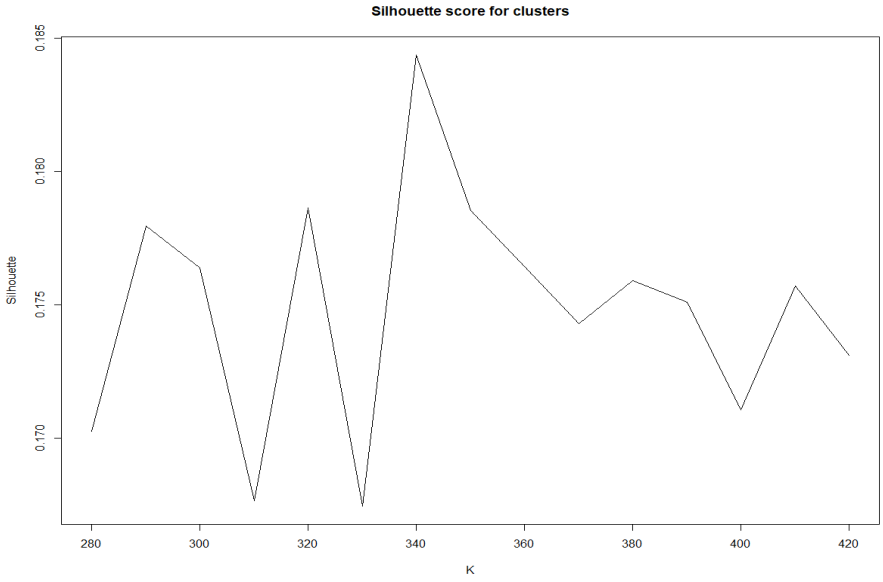


Figure 4.4: Silhouette score for different k in clustering of coordinates dataset for amino acids superposed on the pdb structure of maximum volume. k tested with step 10. [4]

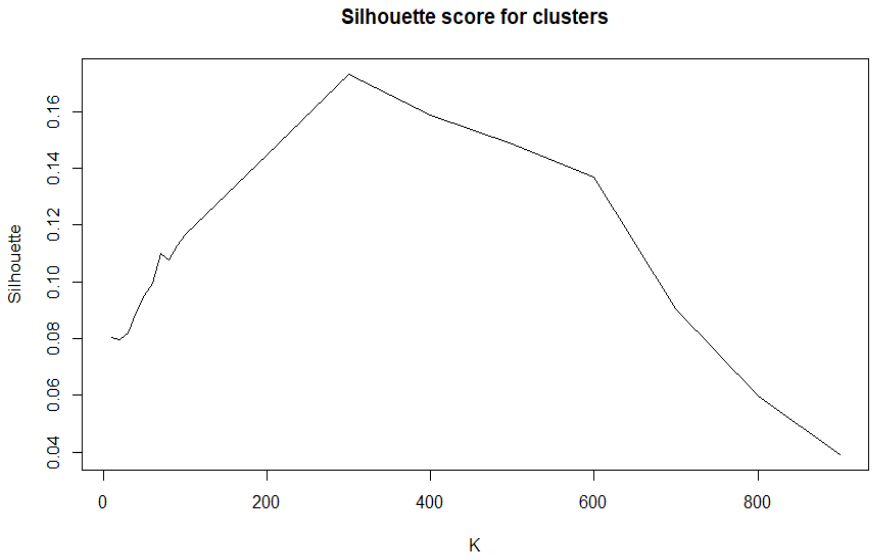


Figure 4.5: Silhouette score for different k in clustering of coordinates dataset for amino acids superposed on the pdb structure of minimum volume. k tested with step 100. [5]

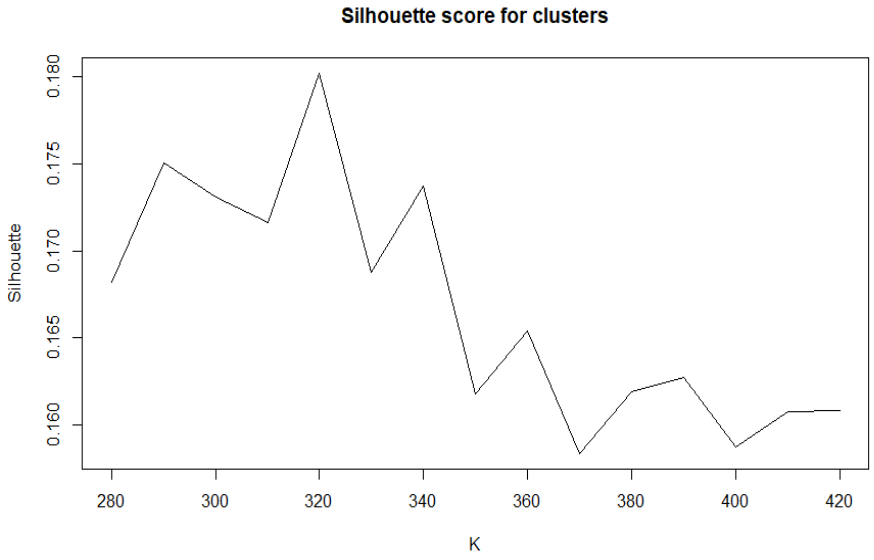


Figure 4.6: Silhouette score for different k in clustering of coordinates dataset for amino acids superposed on the pdb structure of minimum volume. k tested with step 10.[6]

4.3 Dataset with chi angles

4.3.1 Dataset formation

The second representation we chose was dihedral angles of the aminoacids. As before, we do not need the backbone and that is why we use only chi angles.

Each entry of the dataset consists of the conformation id, as described before, and the chi angles for all the amino acids in the glove.

4.3.2 Optimal k

As before, the process of clustering was repeated for many different k in each of the three datasets and the result were compared in order to find the optimal k value. As we can see in the plot below, the optimal k was different in each dataset.

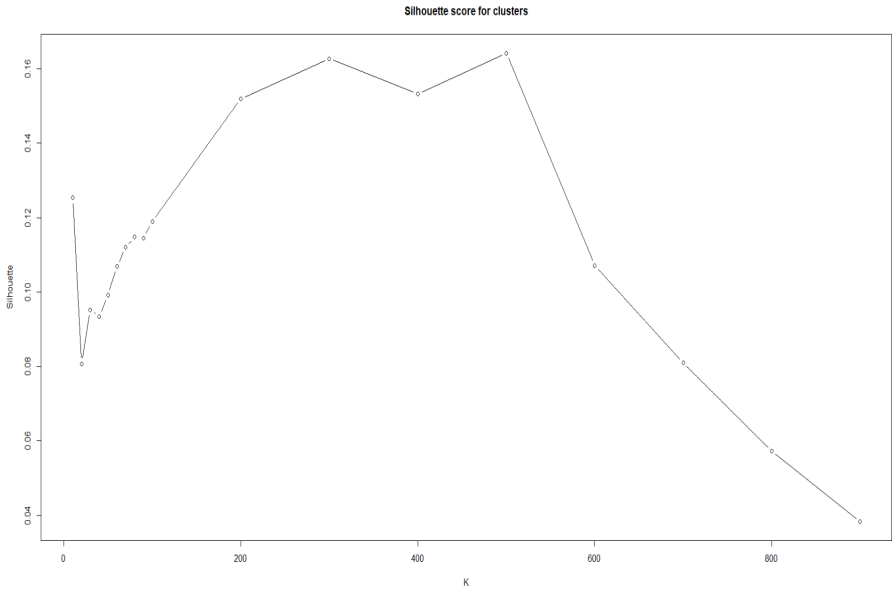


Figure 4.7: Silhouette score for different k in clustering of angles dataset for amino acids superposed on the average structure. k tested with step 100.[7]

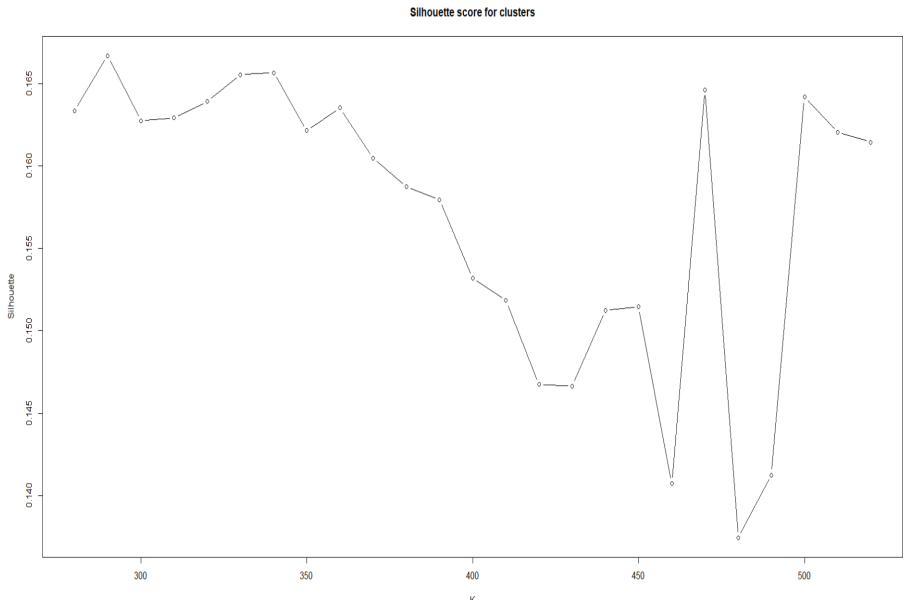


Figure 4.8: Silhouette score for different k in clustering of angles dataset for amino acids superposed on the pdb structure of minimum volume. k tested with step 10. [8]

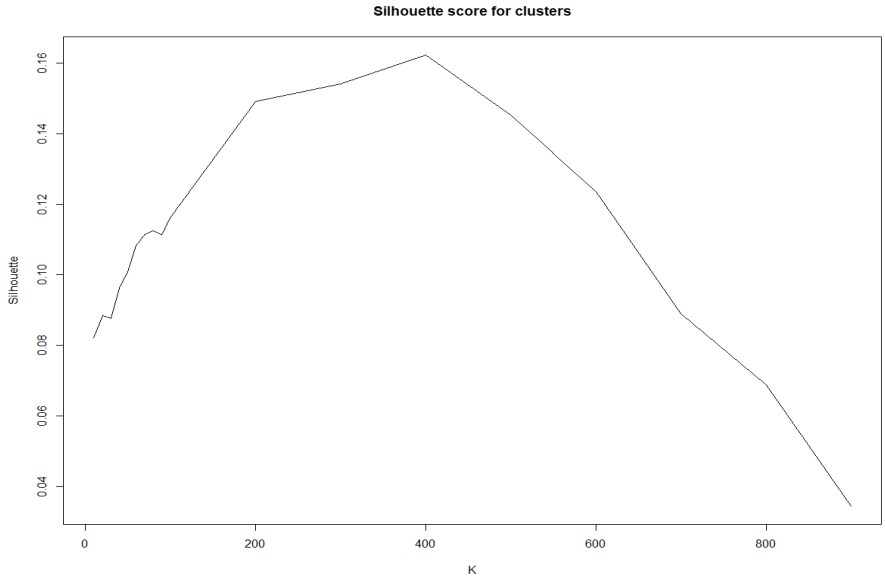


Figure 4.9: Silhouette score for different k in clustering of angles dataset for amino acids superposed on the pdb structure of maximum volume. k tested with step 100.[9]

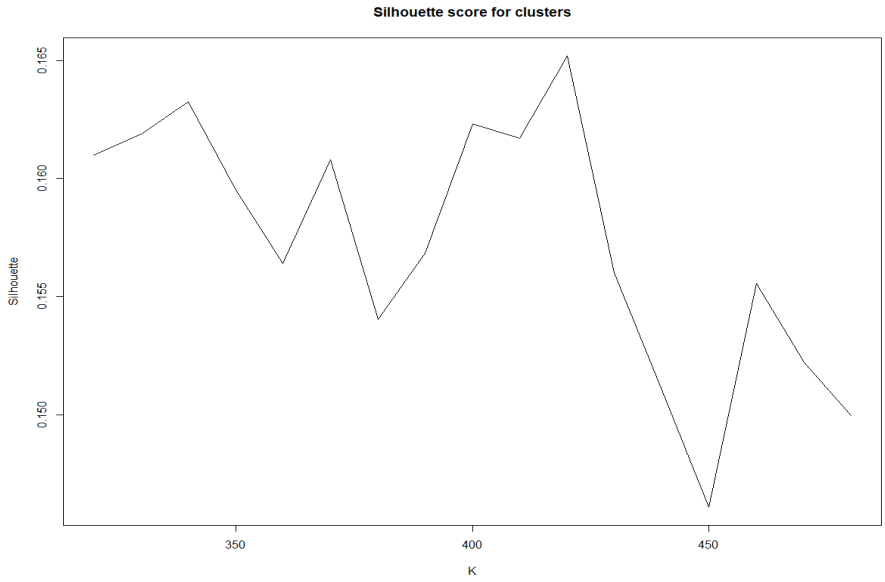


Figure 4.10: Silhouette score for different k in clustering of angles dataset for amino acids superposed on the pdb structure of maximum volume. k tested with step 10. [10]

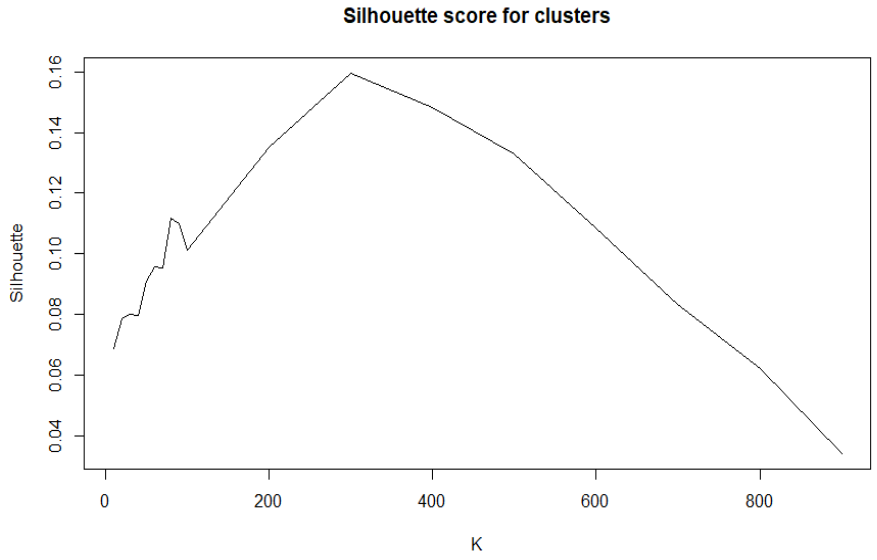


Figure 4.11: Silhouette score for different k in clustering of angles dataset for amino acids superposed on the pdb structure of minimum volume. k tested with step 100. [11]

Figure 4.12: Silhouette score for different k in clustering of angles dataset for amino acids superposed on the pdb structure of minimum volume. k tested with step 10.[12]

4.4 Dataset with phi, psi, chi1 angles and minimum distances

4.4.1 Dataset formation

For the third dataset, we used a very different approach. We wanted to cluster the initial ensemble and that is why we created a dataset where each entry consists of the pdb id and the phi, psi and chi1 angles of the glove's amino acids, as well as the minimum distances between each pair of amino acids in the glove.

This technique requires information about the previous and the following amino acid in the sequence for the calculation of phi and psi angles. There is also a need for the backbone to be dissimilar in each entry of the dataset. For this reason, it is not applicable in the glove created by alternating rotamers.

4.4.2 Optimal k and representative pdbs

The representatives of this clustering were the following 14:

- 1XC7
- 1P4J
- 1WW2
- 2GPB
- 4YUA
- 1WUY
- 1K06
- 4GPB
- 6GPB
- 3EBP
- 8GPB
- 1P2B
- 2PRJ
- 1LWN

The optimal k was found by one test this time because of the small size of the dataset.

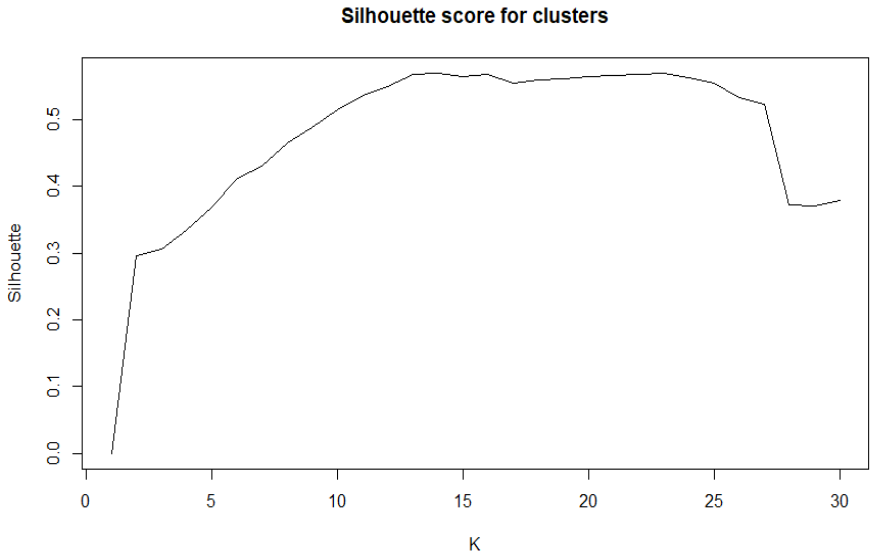


Figure 4.13: Silhouette score for different k in clustering of pdb ensemble. k tested with step 10.[13]

4.4.3 Further examination

After finding the representantive pdbs regarding the initial ensemble, we repeated some parts of the glove creation, introducing gloves, with each of the representative pdbs used as a base protein for superposition. The clustering step was not included.

This is a different point of view, with the purpose of focusing on the most prevailing pdbs, as resulted from clustering, and not the ones chosen for their volume.

4.5 Intersection of representatives

After clustering our data we wanted to examine the intersection of the representatives of the first two datasets (coordinates and angles). The third dataset is not included in the comparison because it contains conformations acquired directly from the ensemble pdbs and not artificially, while it also refers to different dihedral angles.

The intersection of the first two was found and visualized as follows:

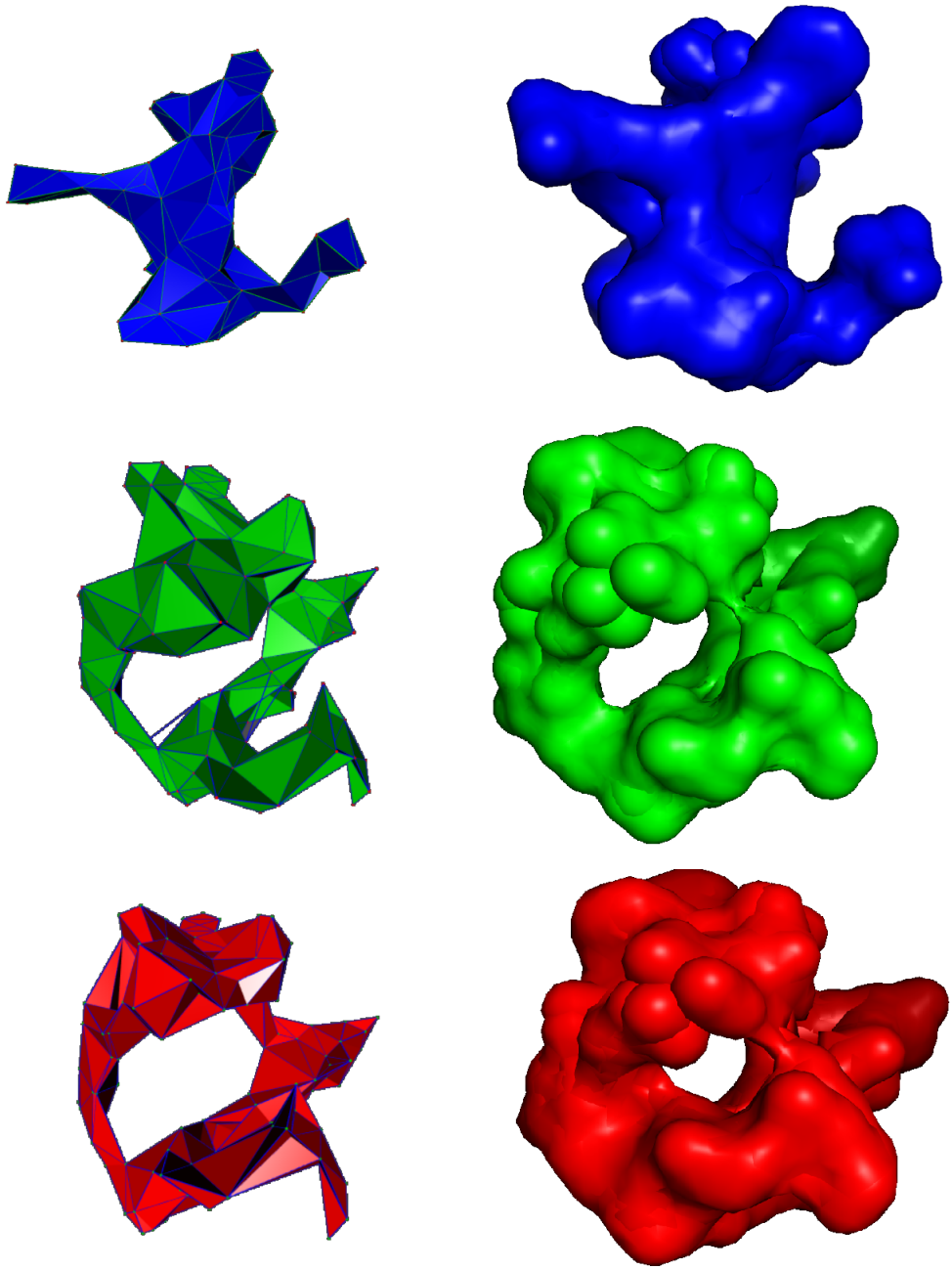


Figure 4.14: Visualization of the intersection of the representatives of the two clusterings based on the average, the maximum volume and the minimum volume structure respectively. On the left we have the alpha shape [6] with α = 2.5 and on the right we have the classic molecular representation with default solvent radius (1.4Å) [12]. [14]

5. OUTPUT

5.1 Output

During the process of creating our glove we obtain important information such as:

- pdb files describing the different conformations of amino acids that could alter the available space of the binding site.
- Statistics describing the probability of appearance for the different rotamers, based on the ensemble of files.
- All the possible conformations of the binding site categorised as viable or non viable (steric clashes).
- Prevalent set of conformations of binding site depending on specific properties (ex. coordinates, angles) as provided by clustering.
- Prevalent set of conformations of binding site as provided by clustering, non dependent on properties (same for all clusterings).
- Gloves describing the binding site in its maximum and minimum volume it may adopt.

5.2 Visualization

While pdb files give us a lot of information, we wanted to be able to visualize our results. For this purposed, we decided to use alpha shapes [6], a visualization method based on Delaunay triangulation of a point set, which uses a parameter α to control the detail of the produced shape.

Alpha shapes are wildly used in the field of bioinformatics, with the purpose of solving the difficult problem of the description of the molecular shape.

The figures included here and in previous chapters describe the active sites of minimum and maximum volume as well as the gloves produced by these structures, but do not need to be limited to these. All the outputs of our workflow can be visualized by this tool as a means to better understand the structure of the selected enzyme.

Figure 5.1: Alpha shape of active site of 2G9R.pdb, which is the file with the minimum active site volume in our ensemble. [1]

Figure 5.2: Alpha shape of active site of 4gpb.pdb, which is the file with the maximum active site volume in our ensemble. [2]

6. TOOLS

- CCP4 suit for superposition of pdb files [9]
- Biopython
- Castp for volume calculation [13]
- Chimera for min/max protein visualization [10]
- Alpha Shapes for visualization [6]
- PyMol for representantives' intersection visualization [12]

7. LIMITATIONS AND ALTERNATIVE APPROACHES

7.1 Limitations

Our project was subjected to some limitations. The major one was related to computational resources. That is the reason why we chose to reduce the number of rotamers to such extend.

We also had to reduce the number of resulting representatives. We chose to acquire the intersection instead of the whole set, with the aim to provide a small number of structures to be used in future research from biochemists.

Additionally, because of the negligible differences in some biochemical properties of our conformations, due to their very similar structure, some algorithms widely used in protein comparison could not be used in our project.

7.2 Alternative approaches

- Ensemble superposition with alternative software (eg.Matlab). We chose CCP4 because it was easier to specify the zones.
- Different tools for volume calculation, such as ProteinVolume [4] and others.
- Disregard the frequency of appearance of the rotamers or use the n closest rotamers for the frequency calculation.
- Use different base pdbs for superposition (currently depending on volume).
- Different tools for visualization of results.

7.3 Assumptions

- Minimum safe distance before a steric clash occurs: 2Å.
- Chosen protein files.
- Chosen amino acids.
- Deduction of rotamers based on appearance frequency.
- Atoms chosen for superposition.

8. CONCLUSION AND FUTURE WORK

8.1 Conclusion

The focus of this work is to use a well-studied enzyme, such as glycogen phosphorylase, with the aim to introduce a new process for calculating the available space within the active center of a protein, in order to pave the way for further studies in drug design and docking.

Both of methods result in a glove describing the space occupied by our enzyme, by producing artificial set of conformations superposed on a selected pdb file. Their basic difference lies on the levels of dependence on the beginning ensemble. While in the first approach the basic superposition structure (average) is produced by the summation of the files, in the second approach the base pdbs do not include information for the ensemble, besides their rating as a clustering representative.

The first proposed workflow can be summarized as follows:

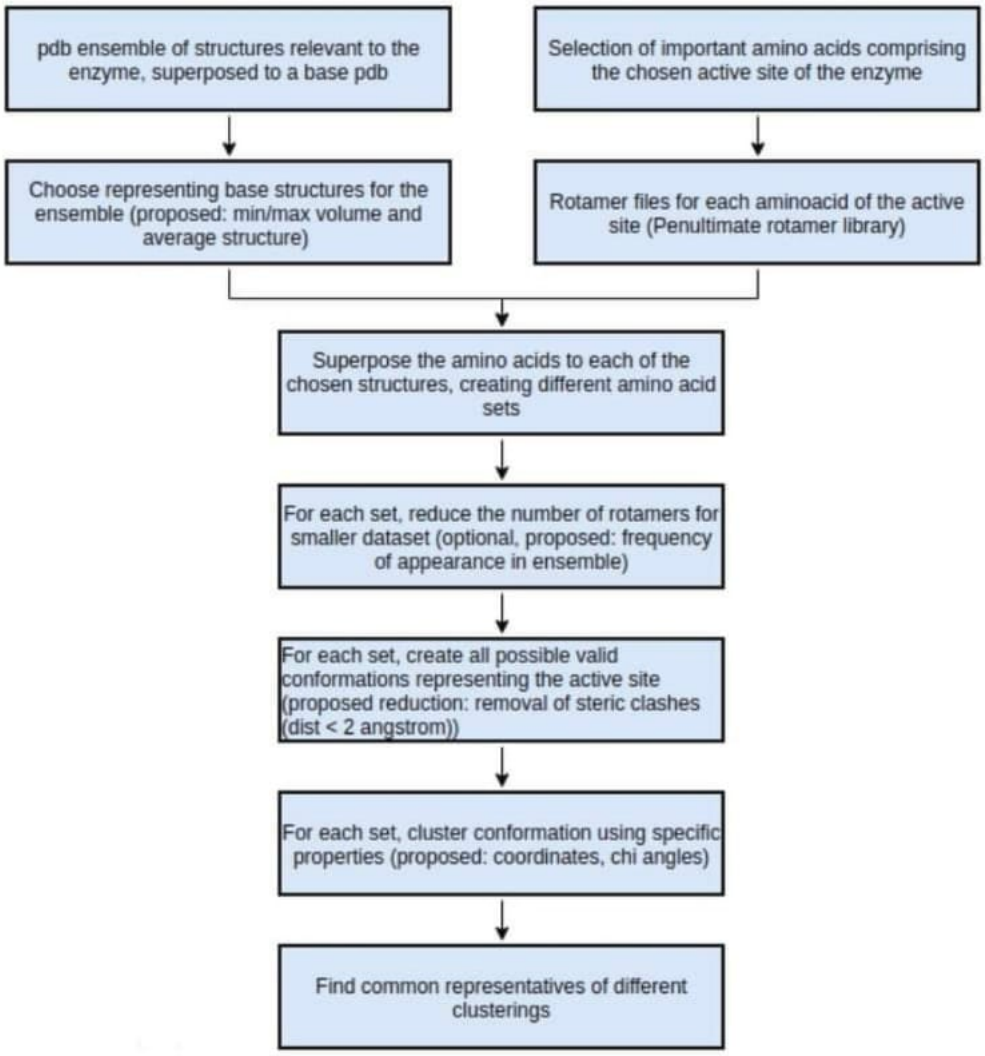


Figure 8.1: First proposed workflow. Clustering the artificial pockets.[1]

The second proposed workflow can be summarized as follows:

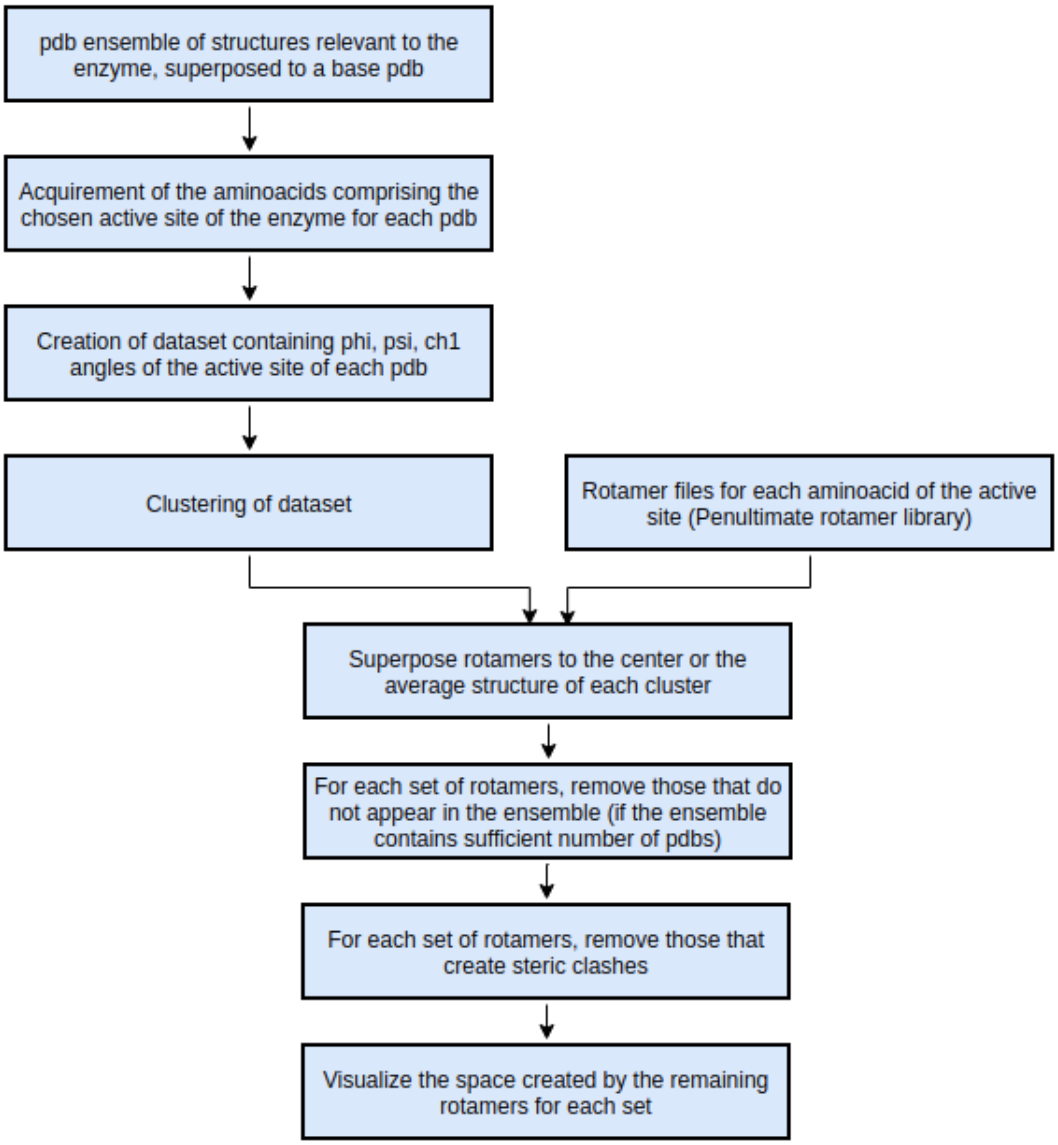


Figure 8.2: Second proposed workflow. Artificial pockets created based on ensemble clustering.[2]

8.2 Future work

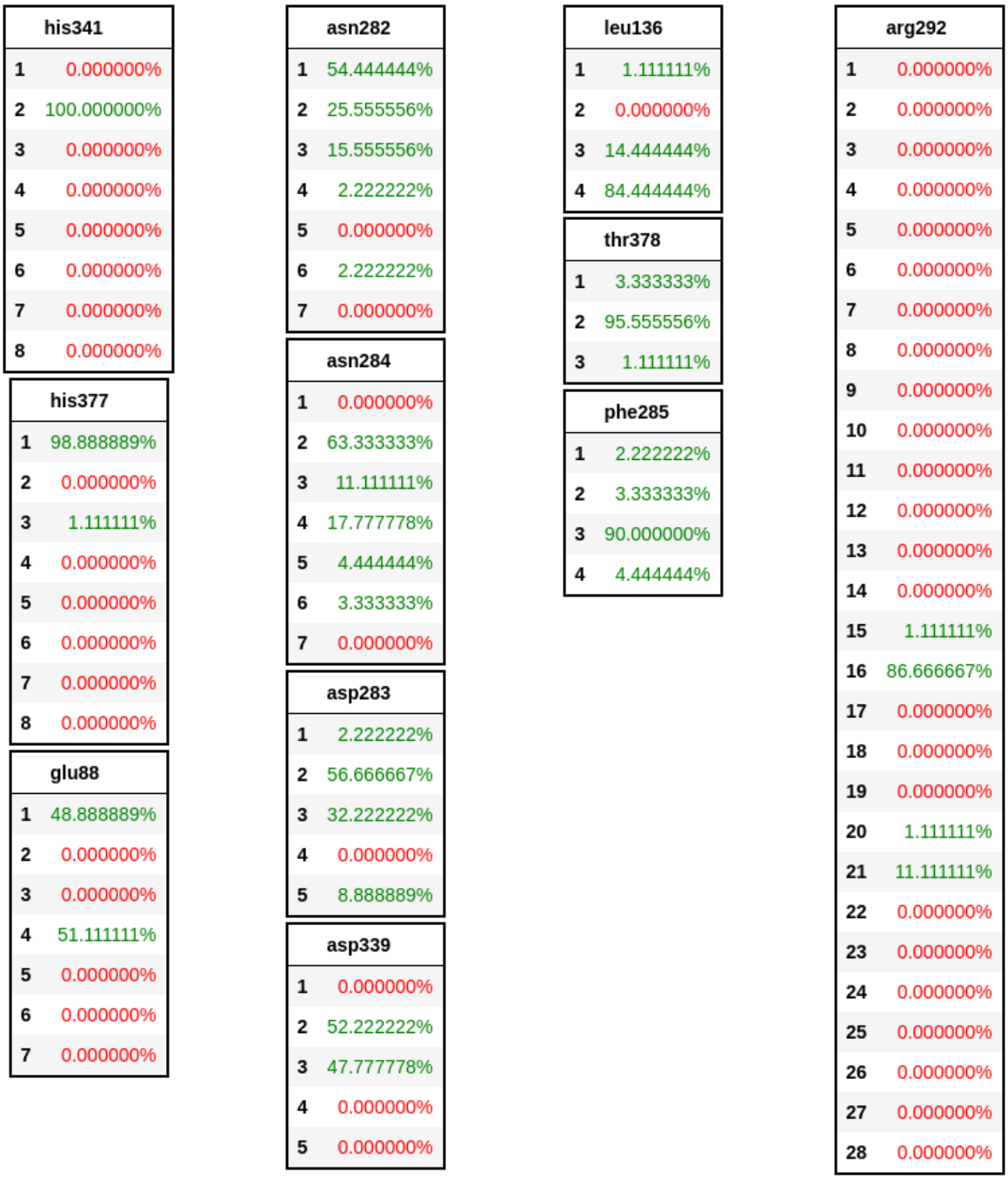
Some ideas for future work include:

- Exploring the biochemical properties which could be included in our clustering dataset.
- Experimenting with different clustering algorithms and evaluation metrics.
- Introduce a ranking system for the conformations depending on the probability to appear in nature.
- Different visualization of the space occupied by our enzyme.
- Chosen base file for protein superposition (5MEM).

ABBREVIATIONS - ACRONYMS

GP	Glycogen Phosphorylase
Arg	Arginine
Asn	Asparagine
Asp	Aspartic Acid
Glu	Glutamic Acid
His	Histidine
Leu	Leucine
Phe	Phenylalanine
Thr	Threonine
RMSD	Root Mean Square Deviation
PDB	Protein DataBank

APPENDIX A. ROTAMERS' APPEARANCE FREQUENCY



his341	
1	0.000000%
2	100.000000%
3	0.000000%
4	0.000000%
5	0.000000%
6	0.000000%
7	0.000000%
8	0.000000%

his377	
1	98.888889%
2	0.000000%
3	1.111111%
4	0.000000%
5	0.000000%
6	0.000000%
7	0.000000%
8	0.000000%

glu88	
1	48.888889%
2	0.000000%
3	0.000000%
4	51.111111%
5	0.000000%
6	0.000000%
7	0.000000%

asn282	
1	54.444444%
2	25.555556%
3	15.555556%
4	2.222222%
5	0.000000%
6	2.222222%
7	0.000000%

asn284	
1	0.000000%
2	63.333333%
3	11.111111%
4	17.777778%
5	4.444444%
6	3.333333%
7	0.000000%

asp283	
1	2.222222%
2	56.666667%
3	32.222222%
4	0.000000%
5	8.888889%

asp339	
1	0.000000%
2	52.222222%
3	47.777778%
4	0.000000%
5	0.000000%

leu136	
1	1.111111%
2	0.000000%
3	14.444444%
4	84.444444%

thr378	
1	3.333333%
2	95.555556%
3	1.111111%

phe285	
1	2.222222%
2	3.333333%
3	90.000000%
4	4.444444%

arg292	
1	0.000000%
2	0.000000%
3	0.000000%
4	0.000000%
5	0.000000%
6	0.000000%
7	0.000000%
8	0.000000%
9	0.000000%
10	0.000000%
11	0.000000%
12	0.000000%
13	0.000000%
14	0.000000%
15	1.111111%
16	86.666667%
17	0.000000%
18	0.000000%
19	0.000000%
20	1.111111%
21	11.111111%
22	0.000000%
23	0.000000%
24	0.000000%
25	0.000000%
26	0.000000%
27	0.000000%
28	0.000000%

Figure A.1: Appearance frequency of each amino acid's rotamers in the ensemble.[1]

APPENDIX B. STERIC CLASHES BETWEEN ROTAMERS (SUPERPOSITION ON AVERAGE PROTEIN)

	asn284_1	asn284_2	asn284_3	asn284_4	asn284_5
asn282_1	True	True	True	True	True
asn282_2	True	True	True	True	True
asn282_3	True	True	True	True	True
asn282_4	True	True	True	True	True
asn282_5	True	True	True	True	True

	asp339_1	asp339_2
asn282_1	True	True
asn282_2	True	True
asn282_3	True	True
asn282_4	True	True
asn282_5	True	True

	phe285_1	phe285_2	phe285_3	phe285_4
asn282_1	True	True	True	False
asn282_2	True	True	True	False
asn282_3	True	True	True	False
asn282_4	True	True	True	False
asn282_5	True	True	True	False

	his377_1	his377_2
asn282_1	True	True
asn282_2	True	True
asn282_3	True	True
asn282_4	True	True
asn282_5	True	True

	asp283_1	asp283_2	asp283_3	asp283_4
asn282_1	True	True	True	True
asn282_2	True	True	True	True
asn282_3	True	True	True	True
asn282_4	True	True	True	True
asn282_5	True	True	True	True

	thr378_1	thr378_2	thr378_3
asn282_1	True	True	True
asn282_2	True	True	True
asn282_3	True	True	True
asn282_4	True	True	True
asn282_5	True	True	True

	arg292_1	arg292_2	arg292_3	arg292_4
asn282_1	True	True	True	True
asn282_2	True	True	True	True
asn282_3	True	True	True	True
asn282_4	True	True	True	True
asn282_5	True	True	True	True

	leu136_1	leu136_2	leu136_3
asn282_1	True	True	True
asn282_2	True	True	True
asn282_3	True	True	True
asn282_4	True	True	True
asn282_5	True	True	True

	his341_1
asn282_1	True
asn282_2	True
asn282_3	True
asn282_4	True
asn282_5	True

	glu88_1	glu88_2
asn282_1	True	True
asn282_2	True	True
asn282_3	True	True
asn282_4	True	True
asn282_5	True	True

Figure B.1: Check for steric clashes between asn282 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with ”True” and the ones that clash with ”False”.[1]

	asn284_1	asn284_2	asn284_3	asn284_4	asn284_5
his341_1	True	True	True	True	True

	glu88_1	glu88_2
his341_1	True	True

	phe285_1	phe285_2	phe285_3	phe285_4
his341_1	True	True	True	True

	his377_1	his377_2
his341_1	True	True

	asp283_1	asp283_2	asp283_3	asp283_4
his341_1	True	True	True	True

	asp339_1	asp339_2
his341_1	True	True

	arg292_1	arg292_2	arg292_3	arg292_4
his341_1	True	True	True	True

	thr378_1	thr378_2	thr378_3
his341_1	True	True	True

	leu136_1	leu136_2	leu136_3
his341_1	True	True	True

Figure B.2: Check for steric clashes between his341 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with "True" and the ones that clash with "False". The pairs already shown do not appear again.[2]

	asn284_1	asn284_2	asn284_3	asn284_4	asn284_5
arg292_1	True	True	True	True	True
arg292_2	True	True	True	True	True
arg292_3	True	True	True	True	True
arg292_4	True	True	True	True	True

	glu88_1	glu88_2
arg292_1	True	True
arg292_2	True	True
arg292_3	True	True
arg292_4	True	True

	asp283_1	asp283_2	asp283_3	asp283_4
arg292_1	True	True	True	True
arg292_2	True	True	True	True
arg292_3	True	True	True	True
arg292_4	True	True	True	True

	his377_1	his377_2
arg292_1	True	True
arg292_2	True	True
arg292_3	True	True
arg292_4	True	True

	phe285_1	phe285_2	phe285_3	phe285_4
arg292_1	True	True	True	True
arg292_2	True	True	True	True
arg292_3	True	True	True	True
arg292_4	True	True	True	True

	asp339_1	asp339_2
arg292_1	True	True
arg292_2	True	True
arg292_3	True	True
arg292_4	True	True

	thr378_1	thr378_2	thr378_3
arg292_1	True	True	True
arg292_2	True	True	True
arg292_3	True	True	True
arg292_4	True	True	True

	leu136_1	leu136_2	leu136_3
arg292_1	True	True	True
arg292_2	True	True	True
arg292_3	True	True	True
arg292_4	True	True	True

Figure B.3: Check for steric clashes between arg292 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with "True" and the ones that clash with "False". The pairs already shown do not appear again.[3]

	asn284_1	asn284_2	asn284_3	asn284_4	asn284_5
asp339_1	True	True	True	True	True
asp339_2	True	True	True	True	True

	glu88_1	glu88_2
asp339_1	True	True
asp339_2	True	True

	asp283_1	asp283_2	asp283_3	asp283_4
asp339_1	True	True	True	True
asp339_2	True	True	True	True

	his377_1	his377_2
asp339_1	True	True
asp339_2	True	True

	phe285_1	phe285_2	phe285_3	phe285_4
asp339_1	True	True	True	True
asp339_2	True	True	True	True

	thr378_1	thr378_2	thr378_3
asp339_1	True	True	True
asp339_2	True	True	True

	leu136_1	leu136_2	leu136_3
asp339_1	True	True	True
asp339_2	True	True	True

Figure B.4: Check for steric clashes between asp339 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with "True" and the ones that clash with "False". The pairs already shown do not appear again.[4]

	asp283_1	asp283_2	asp283_3	asp283_4
asn284_1	True	True	True	True
asn284_2	True	True	True	True
asn284_3	True	True	True	True
asn284_4	True	True	True	True
asn284_5	False	False	False	False

	leu136_1	leu136_2	leu136_3
asn284_1	True	True	True
asn284_2	True	True	True
asn284_3	True	True	True
asn284_4	True	True	True
asn284_5	True	True	True

	phe285_1	phe285_2	phe285_3	phe285_4
asn284_1	False	True	True	True
asn284_2	False	True	True	True
asn284_3	False	True	True	True
asn284_4	False	True	True	True
asn284_5	False	True	True	True

	thr378_1	thr378_2	thr378_3
asn284_1	True	True	True
asn284_2	True	True	True
asn284_3	True	True	True
asn284_4	True	True	True
asn284_5	True	True	True

	his377_1	his377_2
asn284_1	True	True
asn284_2	True	True
asn284_3	True	True
asn284_4	True	True
asn284_5	True	True

	glu88_1	glu88_2
asn284_1	True	True
asn284_2	True	True
asn284_3	True	True
asn284_4	True	True
asn284_5	True	True

Figure B.5: Check for steric clashes between asn284 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with "True" and the ones that clash with "False". The pairs already shown do not appear again.[5]

	asp283_1	asp283_2	asp283_3	asp283_4
thr378_1	True	True	True	True
thr378_2	True	True	True	True
thr378_3	True	True	True	True

	glu88_1	glu88_2
thr378_1	True	True
thr378_2	True	True
thr378_3	True	True

	phe285_1	phe285_2	phe285_3	phe285_4
thr378_1	True	True	True	True
thr378_2	True	True	True	True
thr378_3	True	True	True	True

	leu136_1	leu136_2	leu136_3
thr378_1	True	True	True
thr378_2	True	True	True
thr378_3	True	True	True

	his377_1	his377_2
thr378_1	True	True
thr378_2	True	True
thr378_3	True	True

Figure B.6: Check for steric clashes between thr378 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with "True" and the ones that clash with "False". The pairs already shown do not appear again.[6]

	asp283_1	asp283_2	asp283_3	asp283_4
glu88_1	True	True	True	True
glu88_2	True	True	True	True

	leu136_1	leu136_2	leu136_3
glu88_1	True	True	True
glu88_2	True	True	True

	phe285_1	phe285_2	phe285_3	phe285_4
glu88_1	True	True	True	True
glu88_2	True	True	True	True

	his377_1	his377_2
glu88_1	True	True
glu88_2	True	True

Figure B.7: Check for steric clashes between glu88 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with "True" and the ones that clash with "False". The pairs already shown do not appear again.[7]

	phe285_1	phe285_2	phe285_3	phe285_4
asp283_1	True	True	True	True
asp283_2	True	True	True	True
asp283_3	True	True	True	True
asp283_4	True	True	True	True

	leu136_1	leu136_2	leu136_3
asp283_1	True	True	True
asp283_2	True	True	True
asp283_3	True	True	True
asp283_4	True	True	True

	his377_1	his377_2
asp283_1	True	True
asp283_2	True	True
asp283_3	True	True
asp283_4	True	True

Figure B.8: Check for steric clashes between asp283 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with "True" and the ones that clash with "False". The pairs already shown do not appear again.[8]

	leu136_1	leu136_2	leu136_3
phe285_1	True	True	True
phe285_2	True	True	True
phe285_3	True	True	True
phe285_4	True	True	True

	his377_1	his377_2
phe285_1	True	True
phe285_2	True	True
phe285_3	True	True
phe285_4	True	True

Figure B.9: Check for steric clashes between phe285 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with ”True” and the ones that clash with ”False”. The pairs already shown do not appear again.[9]

	his377_1	his377_2
leu136_1	True	True
leu136_2	True	True
leu136_3	True	True

Figure B.10: Check for steric clashes between leu136 rotamers and rotamers of the remaining amino acids. The ones that will be used together are marked with ”True” and the ones that clash with ”False”. The pairs already shown do not appear again.[10]

BIBLIOGRAPHY

[1] Dihedral angles. https://bit.ly/3jd3WNn. Accessed: 18/10/2021.

[2] Ligand. https://en.wikipedia.org/wiki/Ligand_(biochemistry). Accessed: 2/09/2021.

[3] Proteins and enzymes. https://bit.ly/2TqJAqg. Accessed: 20/07/2021.

[4] Chen C.R. & Makhatadze G.I. ProteinVolume: calculating molecular van der Waals and void volumes in proteins. *BMC Bioinformatics*, 16: 101, 2015.

[5] E D Chrysina. The prototype of glycogen phosphorylase. *Mini-Reviews in Medicinal Chemistry*, 10, 2010.

[6] Edelsbrunner & Herbert and Mücke & Ernst P. three-dimensional alpha shapes. *ACM Trans. Graph.*, 13: 43–72, 1994.

[7] John T. Moore & Richard H. Langley. *Biochemistry for dummies*.

[8] Mamais et al. A New Potent Inhibitor of Glycogen Phosphorylase Reveals the Basicity of the Catalytic Site. *Chemistry—A European Journal*, 23: 8800–8805, 2017.

[9] M.D. Winn et al. Overview of the CCP4 suite and current developments. *Acta Crystallographica*, 67: (D):235–242, 2011.

[10] Pettersen EF & Goddard TD & Huang CC & Couch GS & Greenblatt DM & Meng EC & Ferrin TE. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem.*, 25: 1605–12, 2004.

[11] S.C. Lovell & J.M.Word & J.S. Richardson & and D.C. Richardson. the penultimate rotamer library. *PROTEINS: Structure, Function, and Genetics*, 40: 389–408, 2000.

[12] Schrödinger, LLC. the PyMOL molecular graphics system, version 2.0.

[13] Tian W & Chen C & Lei X & Zhao J & Liang J. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Research*, 46: (W1):W363–W367, 2018.