Assessing, predicting and designing peptide ligands for proteins

Masters Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Kaustubh Amritkar
20151113

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

May, 2020

Supervisor: Dr. M. S. Madhusudhan

Certificate

This is to certify that this dissertation entitled **Assessing, predicting and designing peptide ligands for proteins** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by **Kaustubh Amritkar** at the Indian Institute of Science Education and Research under the supervision of **Dr. M. S. Madhusudhan**, Associate Professor, Department of Biology , during the academic year 2019-2020.

Dr. M. S. Madhusudhan
Associate Professor
IISER Pune

Committee:

Dr. M. S. Madhusudhan

Dr. Siddhesh Kamat

This thesis is dedicated to the five years at IISER.

Declaration

I hereby declare that the matter embodied in the report entitled **Assessing, predicting and designing peptide ligands for proteins** are the results of the work carried out by me at the Department of Biology, Indian Institute of Science Education and Research, Pune, under the supervision of **Dr. M. S. Madhusudhan** and the same has not been submitted elsewhere for any other degree.

Kaustubh Amritkar
20151113
IISER Pune

Acknowledgments

I wish to express my deepest gratitude to my supervisor Dr. M. S. Madhusudhan, for his constant guidance and support in this and the other projects from the past three years. It has been a very erudite and enjoyable experience working with you and thank you for being such an amazing mentor. I want to thank my TAC Dr. Siddhesh Kamat, for reviewing this work and for his valuable inputs in the project.

I am extremely thankful to my lab members Ankit, Neeladri, Sanjana, Tejashree, Golding, Swastik, Neelesh, Gulzar, Yogi, Mukundan and Atreyi for their crucial suggestions in my work and for their continuous help, especially with my presentation skills. You guys made lab a wonderful place to work in. I would also like to thank IISER-Pune Biology department to provide me this opportunity and DST-INSPIRE for their financial support.

I should thank my friends Theja, Neeladri, Raj, Vachan, Suraj, Siddharth, Sreelekha, Anwesha, Shivani, JP and many more that made IISER feel like home during the last year. A special thanks to Sreelakshmi, Santhosh, Ritwik, Sandip, Nida, Sourabh, Muhunden and Prateek for their constant support and encouragement even though they were miles away.

Finally, I would like to pay special regards to my parents and sister, who have been alongside me through all the good and bad times of my life and to whom I shall forever owe all my accomplishments to.

Abstract

Up to 40% of the protein interactions in a cell are mediated by peptides and protein-peptide interactions play a vital role in a cell's functioning. Peptide-mediated protein interactions have been suggested as a potential drug target in many cellular pathways and recently, peptide ligands have attracted a lot of attention as promising drug candidates. Therefore, knowing the structures of such interactions is very essential for their further characterization. In this study, we propose a knowledge-based method for predicting peptide ligands provided a query protein structure with a known binding site. The method first extracts a query structural motif from the binding site of the given protein. We have constructed a library of such structural motifs extracted from the protein structures present in the Protein Data Bank(PDB) against which the query is compared. After finding a structurally similar match from the database, the method extracts the neighbourhood information from the match to predict atoms that will be energetically stable in the query protein's binding site. These predicted atoms will be used to suggest a potential peptide ligand for the given protein. Here, we have developed the framework for this method and performed a set of tests to validate the method's ability to predict an energetically stable partner provided a set of neighbouring atoms. The method, when used to predict a known chemical group when subjected to deletion from a protein structure, was able to correctly predict it back approximately 81% of the time. Since the method focuses on the local packing of atoms in protein structure, it can also be used to predict protein structure stability and to identify missing atoms and residues in protein structures.

Contents

List of Figures

List of Tables

Chapter 1

Background & Introduction

Proteins, often referred to as a "cell's workforce", participate in almost all functions and processes within the cell. But proteins rarely carry out these functions in isolation and more than 80% of proteins interact with other proteins or other biomolecules present in a cell [Berggård et al. 2007, Dhawanjewar et al., 2019]. An important class of these is the interaction between protein and peptide molecules, 15-40% of all interactions in a cell are estimated to be mediated by peptides. Peptide mediated protein interactions are of significant importance in cellular processes like signal transduction, immune responses and transcriptional regulation [Berggård et al. 2007, Yan et al., 2017]. Some classic examples of peptide-protein interactions are the binding of tyrosyl-phosphorylated peptides to proteins containing Src homology domain 2 (SH2) or phosphotyrosyl binding domain (PTB) domain [Bradshaw and Waksman, 2002, Yaffe, 2002]. Many diseases like cancer, amyloidosis, cardiovascular and neurodegenerative disease have been associated with protein-peptide interactions [Johansson-Åkhe et al., 2019].

Multiple studies have shown that peptide mediated protein interactions can be targeted by small molecules [Hammoudeh et al., 2009, Metallo, 2010], making peptide binding sites and protein-peptide interaction potential drug targets. Recently, because of their pharmacological and intrinsic properties, peptides have shown great potential as promising drug candidates and multiple approaches for peptide design have been developed [Ciemny et al. 2018, Bruzzoni-Giovanelli et al. 2018, Fosgerau and Hoffmann, 2015]. Hence, understanding the molecular and structural details for these protein-peptide interactions is very crucial to un-

derstand the functioning of cellular processes and diseases. Understanding this will be very helpful in designing drugs targeting protein-peptide interactions or in designing of peptides as potential drug molecules.

A variety of experimental methods like X-ray crystallography, Cryo-EM and NMR are used [Crystallogr. Made Cryst. Clear, 2006] to obtain the molecular details for a protein-peptide complex. But due to the technical difficulties, time consumption and expenses required to resolve the complex structures, there are very few experimentally determined protein-peptide complex structures present in the Protein Data Bank [Berman et al. 2000] (PDB) as compared to the number of possible complexes that exist in nature. Hence, there is a need for computational methods to build protein-peptide complexes.

Multiple computational techniques to predict protein-peptide interactions and build protein-peptide complexes have been developed [Shoemaker and Panchenko, 2007, Watkins et al., 2017]. There are three major ways for computational prediction of a protein-peptide complex: *De novo*, knowledge-based and docking. *De novo* methods like VitAl generate a peptide sequence by docking amino acid residues pair by pair along the binding site on the query protein [Besray Unal et al. 2010]. *De novo* methods become computationally very expensive as the peptide-length and the number of interactions increase. Knowledge-based methods like SPOT-peptide searches for a homologous protein with a known protein-peptide complex that is similar to the query protein and uses this complex as a template to build a peptide binder for the provided query protein [Litfin et al., 2019]. Knowledge-based methods are highly dependent on the homologous protein-peptide complex that is used as the template, and fail to build a model when a template cannot be found. Docking tools essentially map the peptide(s) at a single or multiple(depending on the method) binding sites on a protein and compare the binding energies for different conformations to predict the most stable protein-peptide complex. A lot of development has happened in the past couple of decades for the docking approach compared to the knowledge-based or *de novo* approach as multiple software and tools have been developed to build a complex by protein-peptide docking [Diller et al., 2015, Watkins et al., 2017] (see [Ciemny et al. 2018] for a comprehensive review of different docking methods).

A limitation with the docking tools is that it needs to sample a very large number of protein-peptide conformations before predicting an optimal binding pose. This sampling step is time consuming and computationally expensive. Another limitation with most of the

docking methods is the requirement of both the protein structure and the peptide sequence(if not structure), hence the inputs information about the peptide is very crucial for accuracy of the method. In many cases, for instance when peptide binders are to be predicted or designed for a novel protein of interest, the sequence or structure information for a peptide binder to the protein is unknown.

In this study, we have shifted from the conventional knowledge-based approaches to develop a method for building protein-peptide complexes that address the above-mentioned problem. The study is based on one main assumption that the local packing of residues in a protein corresponds to a low(or even the lowest) free energy that the atoms in that packing could attain when the whole protein has attained a global free energy minimum [Chen and Kihara, 2011]. It has been reported that the total number of unique folds in the PDB has saturated over time [Fernandez-Fuentes et al., 2010] whereas the number of entries in the PDB has increased with time [refer to Figure 1.1]. This lays the foundation for our study where the information about packing of atoms in the PDB is used to predict peptides against a query protein structure.

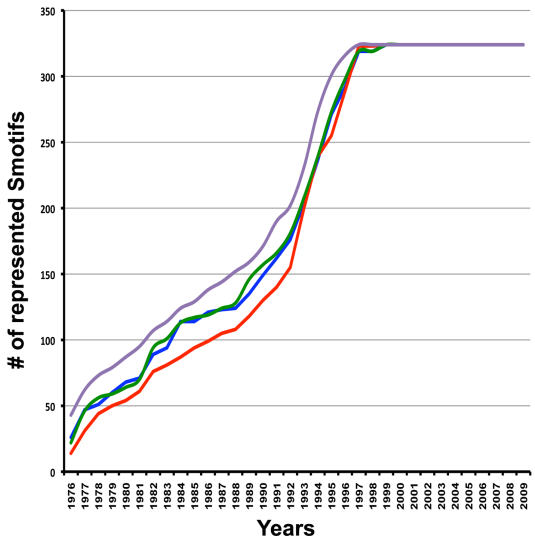


Figure 1.1: Saturation in the unique folds over time in the PDB database. Smotifs are super-secondary structure that represent folds [Fernandez-Fuentes et al., 2010].

Previously, we developed a method to predict alternate binding sites for a given drug molecule. The goal of the study was to predict off-target human proteins that a drug molecule can bind to instead of binding its target protein. The method first extracted the binding sites for the given drug from its target protein and then searched for any alternate binding partners based on the similarity between the drug-bound site and the potential binding pocket of the off-target protein. The method was successfully able to predict alternate binding sites on the off-target proteins and the predictions on the off-target proteins were on a site that was preoccupied by other ligand(s) in the structure [refer to Figure 3.1 in Results]. Based on

the observations of the previous method, for the current project we decided to explore more on the idea of using the specific structural motif to search for structurally similar protein regions.

In this project, the method developed searches for a structural motif similar to that of the binding site in the query protein in all the proteins present in PDB database. Then the method proceeds to predict a peptide ligand for the query protein based on the information retrieved from the matches from the database of protein structures. It has been reported in the literature that protein-peptide interactions observed in nature are similar to and adopt the same structural motifs as present in the monomeric proteins [Vanhee et al., 2009a]. It was also showed that the interaction between these structural motifs present in monomeric proteins can be used to build protein-peptide complexes [Verschueren et al., 2013]. The above mentioned studies provide necessary proof required for justification of the theory used in our approach.

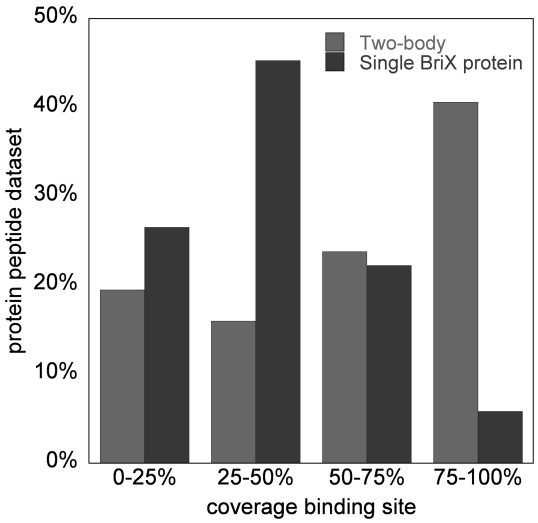


Figure 1.2: Coverage of protein-peptide interfaces.
The two-body(shown in light grey) denote the percentage of protein-peptide interface that can be represented by pair of protein fragments from different proteins [Vanhee et al., 2009b].

For this study, instead of using amino acid residues, we have clustered the heavy atoms of amino acid residues into entities called chemical groups to consider the packing of atoms in a protein and to sample the structural motifs from the binding site. A total of 16 chemical groups are defined for the study(previously based on the work done by Akash Bahai and Swastik Mishra) in such a way that each amino acid residue can be represented as a combination of one or more chemical groups. The chemical groups are used instead of amino acid residues to improve the resolution for observing the local packing in a protein. Although considering atoms to define packing will further improve the resolution, it won't help to focus on the non-covalent interactions in the packing as compared to the chemical groups since atoms usually are involved in strong covalent bonds.

The objectives of this project are to:

(a) Develop a knowledge-based method to design a peptide and model a peptide bound complex for a given query protein structure

(b) Validate working of the developed method by conducting a variety of validation studies

Chapter 2

Methods

2.1 Chemical Groups

In this study, we are using chemical groups instead of amino acid residues to look at the structural motifs in proteins. A chemical group is a group of atoms arranged in a certain way in the three-dimensional space, such that each amino acid in a protein structure can be represented as a combination of these chemical groups. There are a total of 16 chemical groups used in the study [Swastik Mishra Thesis, 2019] as shown in Figure 2.1.

The chemical groups do not consider hydrogen atoms in the proteins since the PDB database, which mostly has X-ray crystallography data that doesn't have information about the hydrogen atoms is used. Each chemical groups is represented by the centroid of all its atoms and distance between two chemical groups is defined as the separation between their centroids in the 3D-space. For this study, the orientation of atoms in a chemical group is not considered.

The r1 chemical group represents the backbone atoms for the residue in a protein. Composition of r1 chemical group for i^{th} residue for a:

(a) Starting residue i.e. N-terminal is N_i, $C_{\alpha,i}$, C_i, O_i and N_{i+1} atoms

(b) Non-terminal residue is $C_{\alpha,i}$, C_i, O_i and N_{i+1} atoms

(c) Ending residue i.e. C-terminal is $C_{\alpha,i}$, C_i, O_i and OXT_i

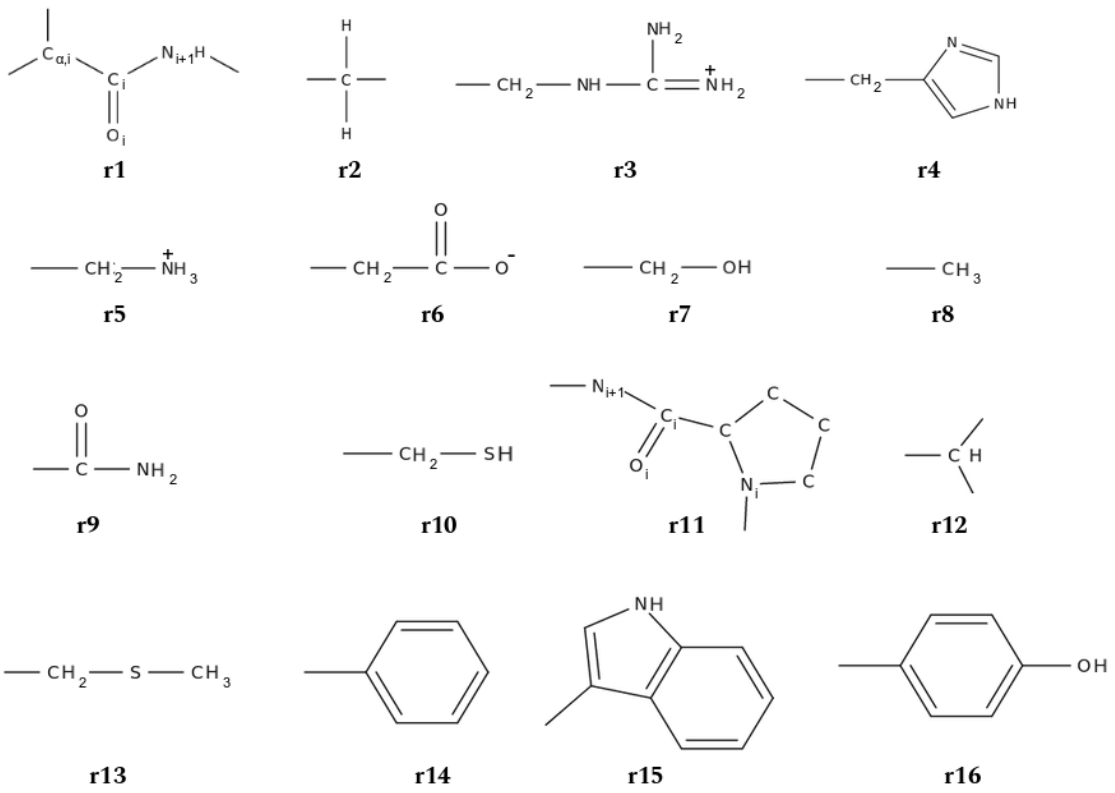


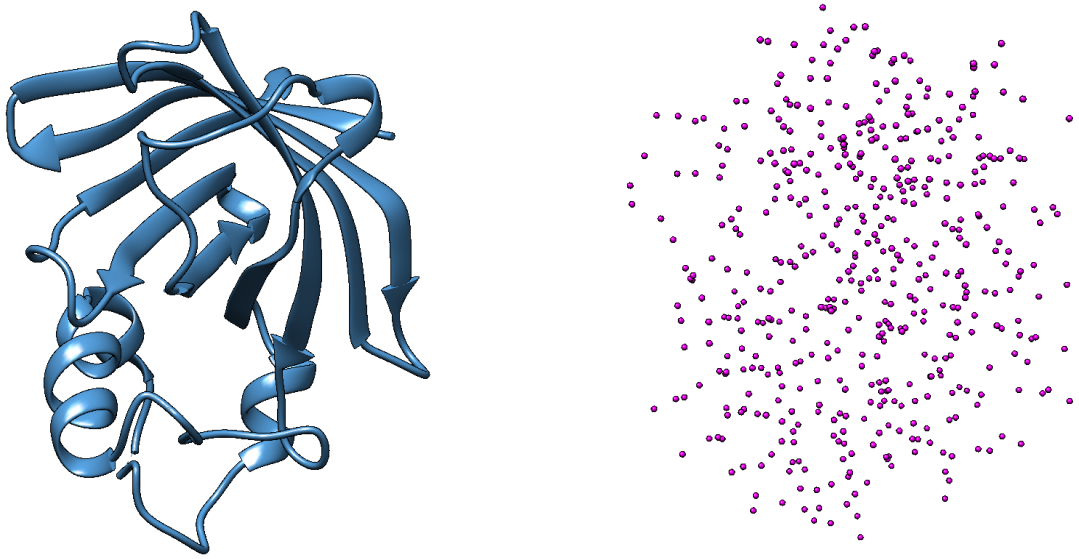
Figure 2.1: The sixteen chemical groups

All amino acids except Proline have one r1 group. The whole Proline amino acid is represented as a separate r11 chemical group since the backbone Nitrogen atom is part of the Proline ring. Atoms with different substitution degrees are considered separately in definition. For ex., r2, r8 and r12 all represent just one Carbon atom but each one has a different number of hydrogen atoms bound to it 2, 3 and 1 respectively in this case. This is done to consider the primary, secondary and tertiary Carbon atoms in a protein separately. All 20 amino acids in nature are a composition of these 16 chemical groups as shown in Table 2.1.

All protein structures are represented in *.pdb* format in the PDB database. These PDB files are converted into their respective *.gpdb* (group-pdb) files which represent the protein structure in the form of chemical groups [refer to Appendix for more details]. See Figure 2.2 for an illustration of a protein and its chemical group representation.

Amino Acid	Chemical Groups	Amino Acid	Chemical Groups
Alanine (A)	r1 + r8	Leucine (L)	r1 + r2 + r12 + r8 + r8
Arginine (R)	r1 + r2 + r2 + r3	Lysine (K)	r1 + r2 + r2 + r2 + r5
Asparagine (N)	r1 + r2 + r9	Methionine (M)	r1 + r2 + r13
Aspartic Acid (D)	r1 + r6	Phenylalanine (F)	r1 + r2 + r14
Cysteine (C)	r1 + r10	Proline (P)	r11
Glutamic Acid (E)	r1 + r2 + r6	Serine (S)	r1 + r7
Glutamine (Q)	r1 + r2 + r2 + r9	Threonine (T)	r1 + r7 + r8
Glycine (G)	r1	Tryptophan (W)	r1 + r2 + r15
Histidine (H)	r1 + r4	Tyrosine (Y)	r1 + r2 + r16
Isoleucine (I)	r1 + r12 + r2 + r8 + r8	Valine (V)	r1 + r12 + r8 + r8

Table 2.1: Chemical group composition of Amino acids



(a) (b)

Figure 2.2: Chemical group representation of a protein (PDB_id: 6CCU).
(a) Ribbon representation of the protein. (b) Chemical group representation of the protein.

2.2 Using stars to represent the structural motif

To search in the database for potential chemical groups that can occupy the binding site cavity in the protein, the method needs to consider the local neighbourhood of the chemical groups in the binding site. The method will then go on to look in the database for a similar "neighbourhood" to gain more knowledge about that specific packing of atoms. For the purpose of this method, a structural motif called **star** is defined to consider the packing of atoms to represent the local neighbourhood of a given site. The star S is defined by two parameters: total number of elements/chemical groups in the star(k) and a maximum threshold distance(d_{thr}).

Let Tn be a set of all n chemical groups in a protein structure. For the i^{th} chemical group $A_i \in T_n$ at the center, a star S_i of size k is defined as the set of k-1 nearest-neighbours $A_j \in T_n$ from A_i such the Euclidean distance between A_i and A_j, $D[A_i, A_j] < d_{thr}$, where d_{thr} is a pre-defined optimal distance cut-off. Figure 2.3 shows an illustration of a typical 9-body star. According to this definition, each chemical group in a protein will have a corresponding star with that chemical group at its center.

Several other definitions for a star were also considered over the course of this project. In one definition, the chemical groups from the same residue as the center were not part of the star. Problem with this definition would be the inability to match query with the database star because the query comprises the same residue chemical groups. Another definition was having a fixed-distance star instead of fixed-body star, where all chemical groups within a certain defined distance would be part of the star, irrespective of the number. Problem with this definition is more on the practical side, because in this case, there will be multiple hits compositions in the database for a query composition since stars with larger number of chemical groups in the database will be increased, increasing the total computational time required for searching in the database. [This paragraph will be more comprehensible to the reader after going through section 2.5]

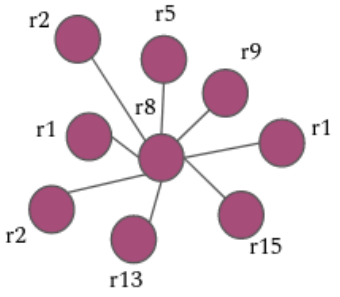


Figure 2.3: Illustration of a 9-body star.
The central chemical group is r8 and the rest are the nearest neighbours.

2.3 Stars Database

In this study, we are observing the structure features in all the proteins present in nature. To achieve this, a database called **stars database** is made based on the stars definition from the previous section. For the creation of the stars database, we created one star each for all the chemical groups present in each of the protein in the PDB database. These stars are then grouped together based on their composition i.e., stars with the same central and neighbouring chemical groups irrespective of their distance-based order from the central group are combined together. The database essentially has the *.cliqs* files for each unique (center + neighbouring) chemical groups composition. These files include the PDB_id and chemical group ids for all stars from the database that have the specific composition. The *.cliqs* filename has a specific nomenclature associated with it, in which the chemical groups separated by "_" are arranged in an alpha-numerical way such that the first one is always the central chemical group. For example, the star in Figure 2.3 would be present in the *r8_r1_r1_r13_r15_r2_r2_r5_r9.cliqs* file in the database.
Refer to Appendix to know in more detail about the database format and the nomenclature followed to store the star composition information in the database.

For this study, proteins from a nr30 version of the PDB database (list of proteins previously curated by Swastik Mishra) is used because using the whole database compared to the nr30 is computationally expensive [Wang and Dunbrack, 2003, Swastik Mishra Thesis, 2019]. Total number of proteins in the nr30 PDB database used as downloaded on 20^{th} May 2019 is 25318, whereas it is 122936 proteins in the complete PDB.

2.4 Input for the method

The problem for designing peptide or small-molecule ligands computationally, can be majorly classified in two steps:

1. Prediction of a potential binding site on the given protein
2. Prediction of a ligand molecule complementary to that binding site

For this study, the method is not addressing the first step of this problem. Input for

the method will be the query protein structure with an user-specified binding site i.e., the list of residues present in the binding site that is to be targeted. There are multiple software and webservers in the field like ACCLUSTER [Yan et al., 2017] and PeptiMap [Bohnuud et al. 2017, Lavi et al., 2013] which perform this specific function of predicting potential **peptide** binding sites for a given protein structure with a significant accuracy. Output from these tools can be used directly as input for this method.
Once the method is complete and validated, we plan to device a strategy to address the problem of predicting potential binding sites and incorporate it with the peptide ligand prediction part.

2.5 Protocol

Since the binding site details are provided by the user, the method initiates by grouping the chemical groups present in the binding site to form a query composition. This query composition of chemical groups is then used to search for a star from the fixed size stars database to find a hit star with the same chemical group composition. These hit stars are then structurally superimposed onto the query composition to find for stars that have significant similarity with the query. Chemical groups from the hit star that don't correspond to any chemical group from the query composition is part of the potential prediction of the peptide ligand. Figure 2.4 summarizes the method. Note that the size of the hit-star will always be larger than the size of the query composition.

Step-wise details of the method are provided in the following subsections:

2.5.1 Extracting query chemical group composition

The user specified binding-site amino acids are converted to their respective chemical groups. The query composition is supposed to be comprised of this set of chemical groups from the binding site. But since the number of binding site chemical groups can be large and the method requires it to be less than the database star size, the binding site chemical groups are distributed into clusters of fixed size as follows.

(a) Consider the chemical group closest to the centroid of all binding site chemical

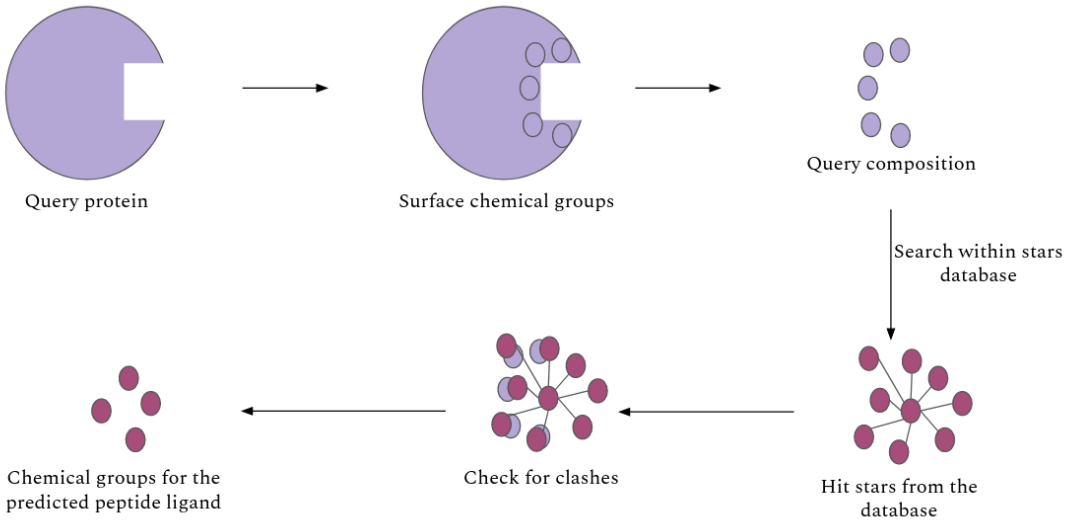


Figure 2.4: Summary of the method

groups

(b) Make a fixed body star around with the chosen chemical group at center, this is the first query composition

(c) Consider the chemical group closest to the previous center but it should not be part of any query composition

(d) Repeat until all chemical groups are incorporated

All the query chemical group compositions built like this are considered separately and each will have its corresponding predictions.

2.5.2 Search in the database

The method aims to search for a star from the database that has **100% Structural Overlap** with the query composition i.e., the database star should have at least one chemical group corresponding to each of the chemical groups from the query. The query composition is small in size compared to stars from the database. Therefore, based on the chemical group composition of the query, multiple **search strings** are built to find all possible stars from the database.

For instance, if the query composition is r1, r2, r2, r5, r13 the following would be the all possible search strings:

(a) *r1_*r13_*r2_*r2_*r5*

(b) r13_*r1_*r2_*r2_*r5*

(c) r2_*r1_*r13_*r2_*r5*

(d) r5_*r1_*r13_*r2_*r2*

Here, these search strings are regular expressions and the asterisk(*) represents zero or more occurrences of any possible characters in the regular expression. The search strings are built such that all unique chemical groups from the query are considered as the center and the rest are arranged in an alpha-numerical order with a "*" between them.
This step provides the list of all possible *.cliqs* files of the stars from the stars database that can potentially overlap with the query.

2.5.3 Finding hit stars for the query composition

The previous step provides a set of all stars from the PDB database that are compositionally similar to the query and can potentially have an 100% overlap with the query. Filtering of these obtained stars has to be done to find for the stars that are structurally similar to the query. To do this, the query is structurally superimposed onto each of the obtained database stars and a RMSD(Root mean square distance) is calculated to assess the quality of the superimposition.

The most straight-forward way to perform superimposition is to carry out superimposition for all possible one-to-one permutations between the query and the database star. But, it is computationally very expensive since the method is only interested in identical mapping.

Following are the steps by which the method chooses a structurally similar star for the query:

(a) Enlist the chemical groups from the database star with no mapping onto the query, these are called "no_maps". Delete the set of all no-maps from the database star.

Note that all no_maps will belong to the database star since the star size is larger than the query

(b) Get all chemical groups that occur exactly once in both query and the database star, these are called "single_maps" and delete them from the database star. This is done because the method requires 100% Structural overlap and single occurring chemical groups between the two structures should correspond to each other

(c) After the previous step, the chemical groups present occur more than once in at least one of the two structures. Now, all possible one-on-one permutations between the chemical groups of the two structures are carried out. The previous two steps are performed to reduce the permutations possible, so as to reduce the computational time required

(d) For each set of permutations along with the single_maps, if any non-identical pairing is not found, the set is discarded. And for the remaining sets with all identical pairings, structural superimposition between the structures using a 3d-least square fit is carried out

(e) During superimposition, the two structures are transformed onto each other such that their centroids are positioned together. Then one structure is rotated with other stationary and RMSD for the corresponding pairs is calculated for each rotation. Rotation with the lowest RMSD value is considered the best superimposition for this set of permutations. And the permutation set with the least RMSD value(RMSD_best) is considered as the optimal correspondence between the chemical groups from the two structures

$$RMSD = \sqrt{\sum_{x,y} [(x_i - y_i)^2 + (x_j - y_j)^2 + (x_k - y_k)^2]} \;\; \forall\, (x, y) \in identical\, pairs$$

(f) The above-mentioned steps are followed for all the database stars(output of 2.5.2). The database stars with the RMSD_best lower than a set threshold are considered as the hit stars for the query with significant structural similarity

After obtaining the hit stars for the query, the method needs to advance towards prediction of a potential peptide, which is ideally to be constructed from the non-superimposed chemical groups from the hit stars. Location of these potential predictions for the peptide could be obtained by transforming the hit stars onto the query protein which gives relative

orientation of the predicted chemical groups w.r.t the query protein. But before that, it is required to perform a positional assessment of the predictions to ensure if they are occupying the binding site. Next section(2.5.4) discusses it in detail.

2.5.4 Checking for Clashes

To ensure if the potential predictions i.e. the non-superimposed chemical groups from the hit stars are occupying the binding site and not interior of the protein, the method checks for clashes between the prediction and the query protein from whose binding site the query composition is extracted. In computational protein modeling or energy minimization protocols, the built structures are often checked for "clashes" to assess the packing and stability of the structure(more the clashes, lesser a structure's stability). Two atoms are considered clashing when they are closer in a 3D-space than the sum of their respective Van der Waals radius. But there is no pre-defined criteria for deciding clashes in case of chemical groups, so we developed a measure very similar to the Van der Waals radii for chemical groups.

For each chemical group, the **clash distance** is defined as the distance from the centroid of the chemical group to its farthest atom along with that atom's Van der Waals radius. Therefore, two chemical groups in a structure would be clashing if the distance between them is less than the sum of their respective clashing distances. The chemical groups' structure in nature won't be completely rigid and would differ with different occurrences of the chemical groups. For this study, we have extracted an occurrence of the chemical groups from one specific protein(pdb_id: 6CCU) and considered the clash distance for that chemical group. Table 2.2 shows the calculated clash distances for all chemical groups. Note that r2, r8 and r12 have the same clash distance and is much smaller compared to other chemical groups. In future, we aim to use a more complete set of all possible conformations of each amino acid in nature like the Dunbrack rotamer library [Shapovalov and Dunbrack, 2011] to calculate the clash distances.

The clash distances for the chemical groups were considered along with a tolerance value. Because for small chemical groups like r2, r8, and r12 the clash distances are very small and having the same cut-off for tolerance value for these small and large chemical groups like r11 and r15 is not fair since small chemical groups tend to have a smaller clash distance, thus significantly reducing their clashes, resulting into a higher frequency of small(r2, r8, r12)

Chemical Groups	Clash Distance(in Å)	Chemical Groups	Clash Distance(in Å)
r1	3.39	r9	2.7
r2	1.7	r10	2.6
r3	3.7	r11	4.3
r4	3.9	r12	1.7
r5	2.4	r13	3.1
r6	3.2	r14	3.1
r7	2.4	r15	4.1
r8	1.7	r16	3.9
r9	2.7	r1	3.4

Table 2.2: Clash Distances for each chemical group.

chemical groups in the predictions. Therefore, we classified chemical groups into 4 classes of tolerance values such that, small chemical groups have lower tolerance and comparatively higher tolerance values for the larger chemical groups. Table 2.3 depicts this classification. This classification is performed on a knowledge basis and has not been optimized.

Chemical Groups	Tolerance Values
r2, r8, r12	10%
r14, r13, r10, r7, r5	20%
r16, r9	30%
r15, r11, r6, r4, r3, r1	40%

Table 2.3: Tolerance values for all chemical groups.

2.5.5 Prediction of chemical groups in the peptide ligand

The non-clashing non-superimposed chemical groups from the hit stars are the predictions for the query compositions. All these steps(2.5.2 - 2.5.4) are carried out for all the query compositions obtained in 2.5.1.

The next step is combining the chemical group predictions from all the query compositions and designing a peptide parsing through this group of predicted chemical groups. We haven't decided on a procedure to perform this step because a set of validation tests are being conducted first to assess the working of the method. One potential way to do this is by making a database of all peptides in nature and then searching for a peptide that best fits orientation of the predicted chemical groups.

According to the method's assumption, the interaction between a peptide binder and a receptor is similar to the interactions observed within protein structure. To validate this assumption and the working of this method, we plan to do validation tests.

The next section will describe the first validation test and its various aspects in detail.

2.6 Validation by Missing chemical group case

In this section of validations, the aim is to delete a known chemical group from the defined structural motif and then using the developed method to identify the deleted chemical group. The advantage of this validation is knowledge of the deletion, and hence the ability to assess the prediction accuracy.

Any star from a protein structure is considered and the central chemical group from this star is deleted, making it the query shell(shell is defined as a star without its central chemical group). For a query shell, the stars database is searched for a star that has the shell chemical group composition similar to that of the query shell. The stars database should have stars of the same size as the query star. Since in the stars database, the star compositions are named in an alpha-numerical manner with the central chemical group as the first(refer to section 2.3), there is a maximum of 16 different chemical group compositions in the database that can match with the query shell. For instance, consider the star shown in Figure 2.3, the query star size is 9, the query shell composition will be *"r1_r1_r13_r15_r2_r2_r5_r9"* and the search string will be *"*r1_r1_r13_r15_r2_r2_r5_r9.cliqs"*, which will have at most 16 matches in the database. After getting this list of all possible stars composition from the database, all the stars are stripped of their central chemical groups and the shells are superimposed structurally onto the query shell as described in section 2.5.3. For all the shells from the above extracted stars with the RMSD_best lower than a threshold RMSD, the central chemical group is considered as the prediction.

This analysis is performed for a whole protein(pdb_id: 1Z7K) structure, such that individual deletion of each chemical group from the structure is carried out and the predictions are considered to assess working of the method.

This missing chemical group validation is used to test for multiple aspects as follows:

2.6.1 Comparison with CLICK

CLICK [Nguyen et al., n.d., Nguyen et al., 2011] is a topology independent software to compare biomolecular 3D structures. It is a tool capable of performing 3D-structural superimposition between the two given structures of any biomolecules(protein, DNA, RNA etc.). This tool is used to compare against our method in predicting the deleted chemical groups in the missing chemical group validation test. CLICK is used because its working is similar to the method developed in this study since it considers for a 'clique' of points (usually 3-7 amino acids) between the two structures and tries to find a structurally similar ones to superimpose the two structures.

The main difference between CLICK and the our method is that CLICK does not have a fixed size for the clique as compared to the our method where the star sizes are fixed. Also in the our method, only superimpositions with a 100% structural overlap are considered whereas, CLICK can have superimpositions without 100% structural overlap.
For this analysis, the same query shells are used to make predictions for the central chemical group for both methods. Note that predictions from CLICK superimpositions with a 100% structural overlap are only considered.

2.6.2 Correlation with conservation profiles of the deletions

This part of the study is to check for a relation between the predictions made by our method and the evolutionary aspect of the deletions. After performing the missing chemical group validation test, the deletions with incorrect predictions were analyzed. We compared the incorrect predictions with their corresponding Amino Acid conservation profile obtained via the protein's Multiple Sequence Alignment.

2.6.3 Correlation with the DEPTH

Residue DEPTH [Chakravarty and Varadarajan, 1999, Pern Tan et al., 2013] is a software that calculates the depth of a residue from the protein surface and can be used as a measure for protein structure stability. In this study, the missing chemical group test is performed on all chemical groups present in a protein, for which a sequential deletion of each chemical

group was performed individually. The position-wise prediction accuracy of the method is compared with the DEPTH values for each of the chemical groups subjected to deletion. The goal here is to check if the prediction quality for the deleted chemical groups is correlated with the respective stability in the protein. For instance, if a chemical group with a low depth value(present on surface) has a poor prediction accuracy compared to one which is buried inside the protein and vice-versa.

2.6.4 Comparison between using the database and a smaller random set

In the missing chemical group validation, there is a total of 16 different possible predictions i.e. the 16 chemical groups. To obtain the predictions, all stars(without their centre) from 16 star compositions from the stars database need to be compared with the query shell. This is a very large number of comparisons and becomes computationally very expensive. Therefore, a set of random stars from each of this stars composition is used for comparison with the query shell, reducing the computation time significantly. This analysis is done to check if a random small subset instead of all stars from the database can be used to obtain the correct predictions and if there is a significant difference in predictions by this small subset and the whole database.

2.7 Alternate Approach

Before deciding on the method explained in the previous sections, we had tried using a different approach to address the problem of designing peptide ligands for protein structure. In comparison to the above-mentioned method(Section 2.5), in this approach the query chemical group compositions were extracted in a different manner and the steps involved in predicting chemical groups occupying the binding site were different.

Following subsections explain the steps involved in working of the method in more detail:

2.7.1 Protocol

After obtaining the chemical groups present in the binding site(as provided by user), a fixed size n-body star is created for each of the binding site chemical groups as the centre from the query protein. These stars are considered as the 'query' in this method.

A stars database with the star size same as the query stars is used. For each of the query stars, there is only one unique *.cliqs* composition in the database. Each of the database star from this *.cliqs* file is structurally superimposed onto the query star to obtain hit stars with RMSD lower than a threshold(refer to section 2.5.3).

After obtaining hit stars from the stars database, the chemical groups surrounding the hit stars are considered as predictions for the query star. To extract the surrounding chemical groups from the hit star, the hit star is extended. This is done by creating a separate star for each one of the chemical groups as a center. All the chemical groups present in the newly created stars that were not part of the hit star are the extension of the hit star and the chemical groups present in the extended hit stars are the potential prediction in this method. Figure. 2.5 for a schematic of extension of hit star from the database.

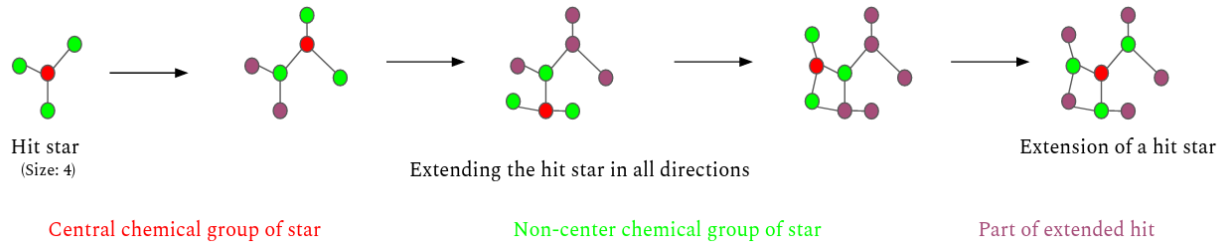


Figure 2.5: Schematic for extension of hit stars from the database.
First sub-figure represent a 4-body hit star, with center colored red and green for non-central chemical groups. In each step, the hit star is extended by considering neighbours of the non-central chemical groups of the initial hit star. Extended chemical groups are represented in purple.

The extended hit star chemical groups are then transformed onto the query protein, based on the transformation matrix obtained during superimposition of the query and hit star. These potential predictions are checked for clashes(refer to section 2.5.4) with the query protein structure and the non-clashing chemical groups are the predictions to be part of the peptide ligand.

2.7.2 Validation by missing Amino Acid Case

For the method described in section 2.7.1, we had performed a missing amino acid validation test. In this test, an amino acid is deleted from the protein structure and using the method described, we predicted the chemical groups that fill up the cavity left behind by deleting the amino acid. The validation is carried out to assess performance of the method by checking if the predicted chemical groups are the ones that belong to the deleted amino acid.

Chapter 3

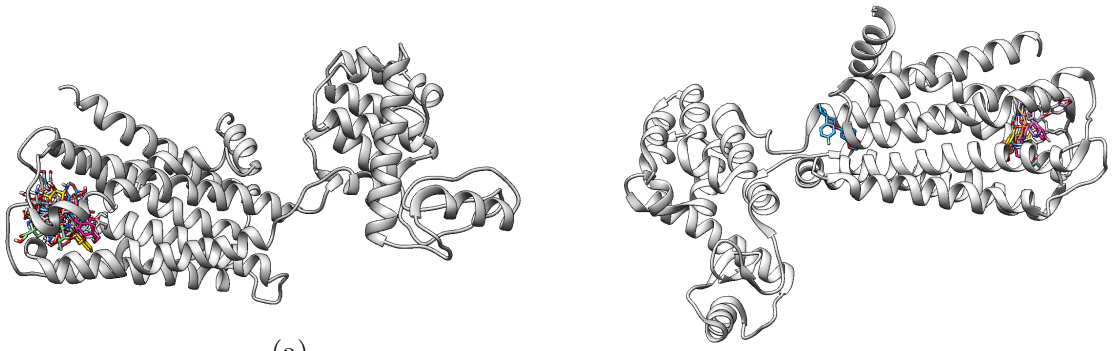
Results & Discussion

3.1 Prediction of binding site for drugs on off-target proteins

Earlier we had developed a method to predict binding sites on off-target proteins for a given drug. The method extracted binding site for the given drug molecule from its drug-bound complex from the PDB. Residue DEPTH is then used to extract potential ligand binding sites from the off-target protein structures. These two extracted binding sites from the drug-bound complex and the off-target protein are structurally superimposed onto each other using CLICK. This superimposed structures if have a good structural overlap and a low RMSD, are considered as a match and the drug molecule is then superimposed onto the off-target protein binding site to predict the protein-drug complex. This method was tested on an experimentally validated dataset of known drug molecules and their off-target proteins [Campillos et al. 2008]. Figure 3.1 shows the results for the binding pose of two drug molecules Doxorubicin(DM2) and Paroxetine(8PR) onto their off-target proteins HRH1 and DRD3 respectively. Both DRD3 and HRH1 are membrane receptors.

Predictions made using all 10 sites from DM2 bound protein complexes were on the same binding site on the off-target HRH1 protein(Figure 3.1(a)). Similarly out of 9 sites from 8PR bound proteins, 7 superimposed onto one binding site and the remaining 2 onto another binding site on the off-target DRD3 protein(Figure 3.2(b)). These predicted binding

sites on the off-target proteins of both HRH1 and DRD3(the one with 7 grouped) already had a ligand bound to it in their crystal structure. Also, both predicted binding sites were present on the extracellular side of the membrane proteins increasing confidence in the predictions.



(a) (b)

Figure 3.1: Binding predictions on off-target proteins for drug molecules
(a) Predicted binding poses for Doxorubicin on its off-target protein HRH1. (b) Predicted binding poses for Paroxetine on its off-target protein DRD3.

While searching off-target binding sites for a drug molecule in the human proteome, we obtained a large number of false positives, as approximately 38% of all human proteins were predicted as off-targets for the drug molecule. The method only used structural features to search for off-target binding sites and was not considering the chemical information like interaction details. In this study, we expand more on predicting binding sites and binders based on the already present information about these interactions in the PDB. We are addressing the problem of incorporating chemical features by considering the packing of atoms in protein structures. We are trying to predict peptide binders based on configuration of the binding site and the neighbourhood that is energetically stable with it. This study is focused specifically towards prediction of peptide ligands because the study looks at interactions within a protein structure, hence the interacting partners(Amino acids) can be represented in a peptide ligand, which essentially is an amino acid sequence, on a binding site.

3.2 Details about Stars Database

For a stars database of 7-body stars, there are a total of 100,797 unique star compositions as defined in section 3.3. Figure 3.2 shows the frequency of occurrence of the 16 chemical groups in the nr_30 PDB database. r1 is the most abundant chemical group in the database, which makes sense since it is present in all amino acid residues except Proline. Apart from r1, the single carbon atom chemical groups r2, r8 and r12 are present in large number in the database compared to the other chemical groups.

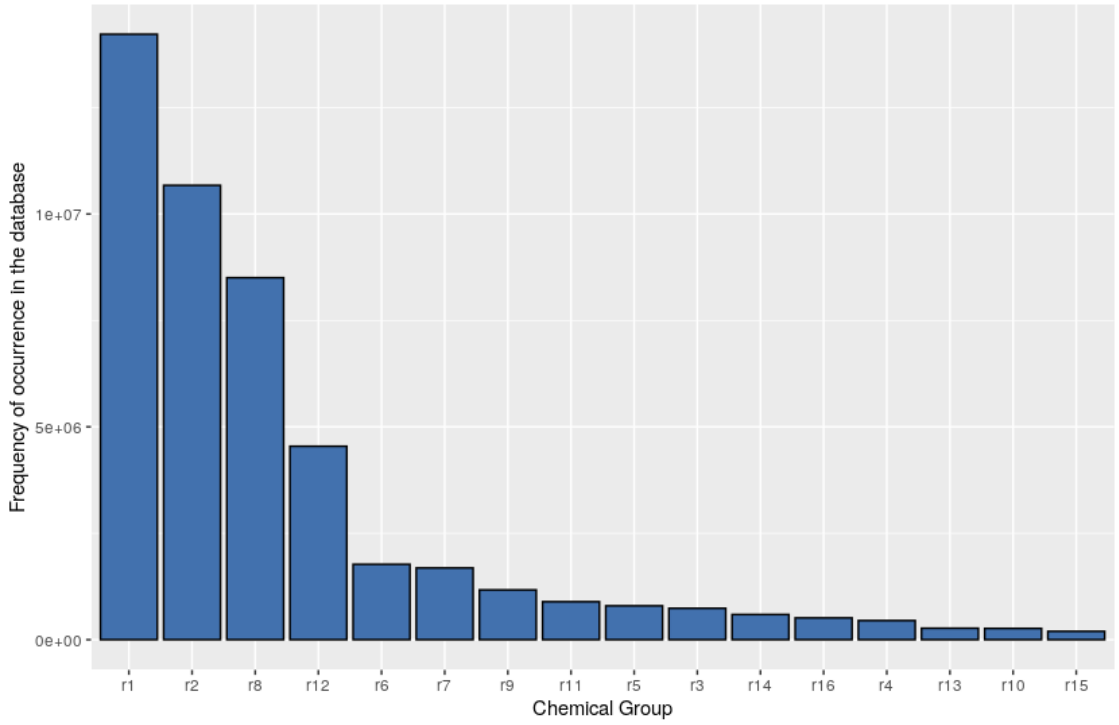


Figure 3.2: Frequency of all chemical groups in the protein gpdb database

Following section has results for the Missing Chemical Group validation test.

3.3 Missing Chemical Group Validation

This section describes a validation test of our method. This is a Missing chemical group validation study, where predictions are done for individual deletions of known chemical groups from a protein structure. Note that all the predictions made by our method for the deleted

chemical groups is completely independent of their respective Amino acid information, i.e., no prior information about the amino acid identity of the deleted chemical group is provided while making the predictions.

This section is divided into following subsections: comparison of our method to CLICK for predicting deleted chemical groups, validation of method by deleting chemical groups sequentially and studying the predictions in structural and evolutionary context.

3.3.1 Samples more stars compared to CLICK

In this analysis, the predictions for a set of missing chemical group cases are performed by our method and compared to the predictions done by CLICK. A set of 5 different chemical groups (refer to Table 3.1) from a human Angiogenin protein (PDB_id: 1ANG) were deleted individually. 7-body stars were used for carrying out this part of the analysis, that makes the shell(star without its center) size of 6 chemical groups and the RMSD(RMSD_best) cut-off of 1Å was set. Table 3.1 shows the details of the chemical groups that are deleted for this analysis.

Chemical Group	**Chemical Group Number**	**Amino Acid**	**Star Composition**
r1	101	Gly(34)	r1_r1_r1_r12_r2_r7_r8
r3	100	Arg(33)	r3_r1_r1_r2_r2_r2_r8
r8	246	Thr(79)	r8_r1_r1_r1_r2_r7_r8
r11	53	Pro(18)	r11_r1_r1_r2_r2_r2_r6
r15	271	Trp(89)	r15_r1_r1_r11_r11_r2_r7

Table 3.1: Details about the chemical group deletions from 1ANG

For the shells corresponding to these chemical groups in the structure, central chemical groups were predicted. In Table 3.2, Total shells in db represents the total number of stars from the database that can have a 100% structural overlap with query shell, i.e. the total number of superimpositions that were carried out to find for database shells with same chemical group composition as query shell and the number in brackets represents the number of central chemical groups for that shell composition in the stars database. Total Predictions is the number of shells from the database that have an RMSD of superimposition lower than the threshold and Correct Predictions are the number of predictions where the predicted chemical group is same as the deleted one.

Out of the 5 chemical group deletions, the r15 chemical group in the 271_th Tryptophan residue is the only deletion without any predictions, because both the methods were unable to find a shell from the database that is structurally similar to the query shell corresponding to the r15 deletion. This suggests no or a very limited occurrence of that specific structural motif in the protein database. In most cases, the prediction accuracy is better with CLICK but our method is able to sample more stars for comparison than CLICK.

Chemical Group	**Correct Predictions**		**Total Predictions**		**Correct Prediction %**		**Total Shells**
	CLICK	**Our Method**	**CLICK**	**Our Method**	**CLICK**	**Our Method**	
r1_101	1603	5875	1811	5967	88.51%	98.46%	524285 (16)
r3_100	1	10	24	314	4.16%	3.18%	560126 (16)
r8_246	300	223	302	293	99.34%	76.11%	315627 (16)
r11_53	23	9	44	20	52.27%	45%	349950 (16)
r15_271	0	0	0	0	-	-	1862 (15)

Table 3.2: Details about the predictions with CLICK and our method for the 5 chemical group deletions from protein 1ANG.

Over the deletions performed, it can be observed that using our method we are able to sample many more database shells with 100% structural overlap as compared to CLICK. This increases the confidence in using our method over CLICK, since we are able to sample more structurally similar shells and not missing out on potential shells as with CLICK. It might also increases the number of incorrect shells but more structurally similar shells would benefit in the generation of large number of conformers for a prediction.

3.3.2 Individual deletion of all chemical groups in a protein

In this analysis, all chemical groups from the protein(pdb_id:1Z7K) were deleted sequentially and predictions are made for each one of them. The structure used to conduct these deletions is a protein-peptide complex with a total of 842 chemical groups. Figure 3.3 represents the

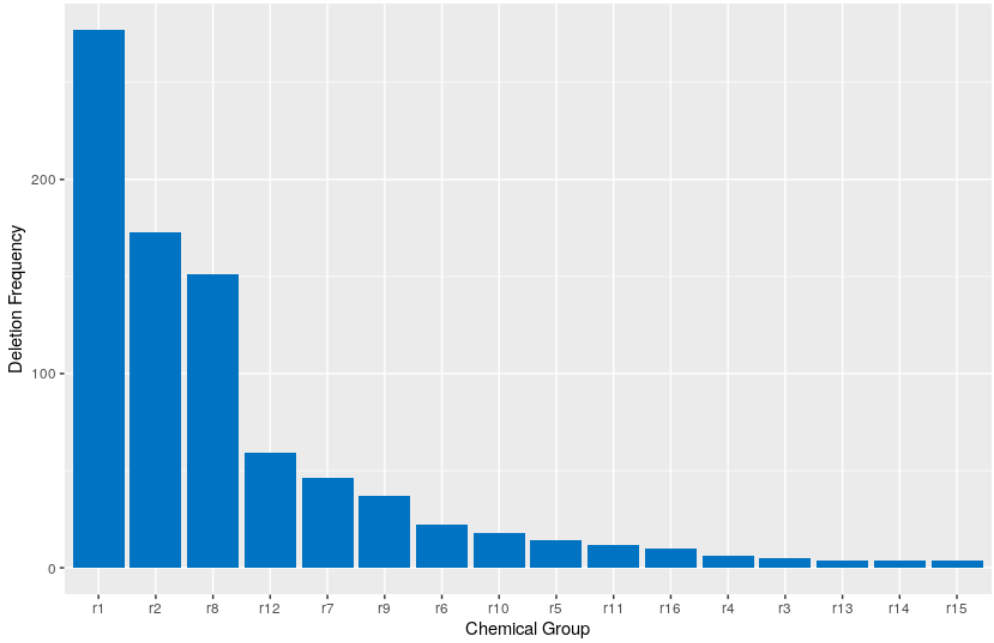


Figure 3.3: Frequency of chemical groups in the protein(PDB_id: 1Z7K)

frequency of chemical groups present in the protein. The protein structure comprises a large number of r1, r2 and r8 chemical groups compared to rest of the chemical groups.

Figure 3.4 shows a histogram for the correct prediction percentage over all 842 chemical group deletions. This plot shows that for most of the deletions, the prediction accuracy for deleted chemical groups is very high(90-100%). Percentage Correct Prediction(x-axis) is a percentage of total predictions where the predicted chemical group is identical to the deleted chemical group(same as column 4 in Table 3.2).

$$Correct\ Prediction\ Percentage\ =\ \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}\ \times\ 100$$

Figure 3.5 shows the average correct prediction percentage for each one of the chemical group over all 842 deletions. Histogram in Figure 3.4 suggests that for the majority of deletions, the predictions made were correct. This does not effectively validate the method's working because there is a disparity in chemical groups frequency in the protein and the plot in Figure 3.5 suggests that the prediction accuracy for different chemical groups is very different. The average prediction accuracy is much higher for the more frequently occurring

chemical groups like r1, r2, r8 and r12 compared to other chemical groups, therefore, making the overall correct prediction percentage(Figure 3.4) biased.

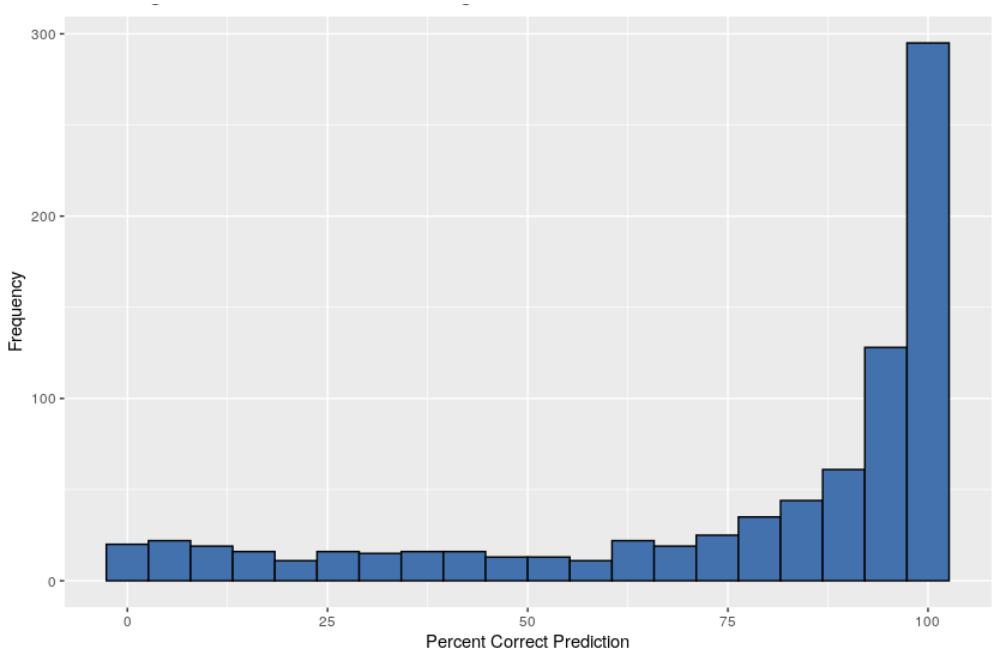


Figure 3.4: Correct prediction percentage over all chemical group deletions

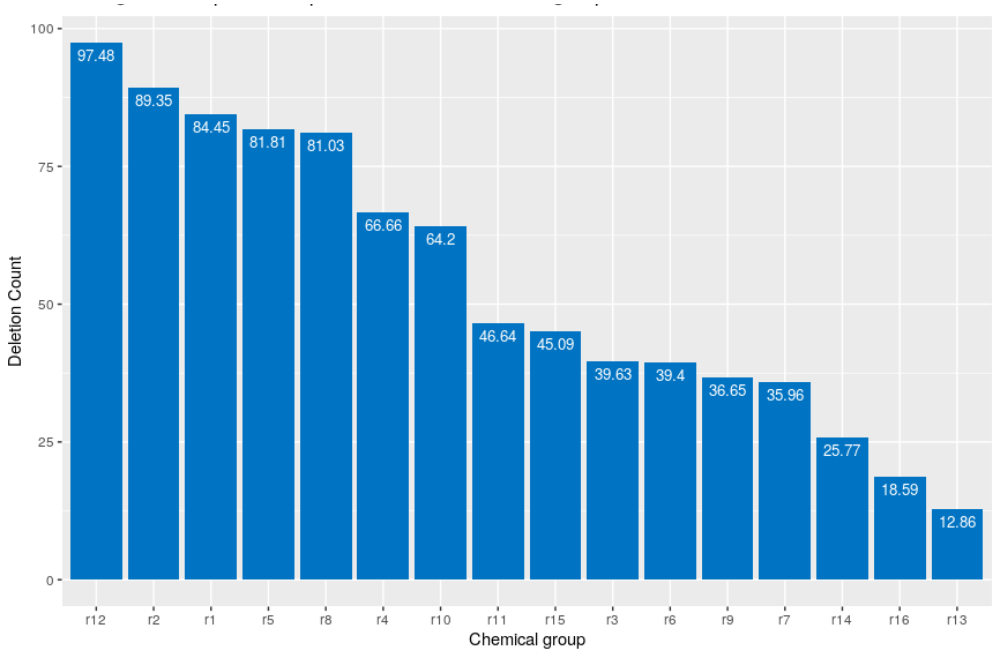
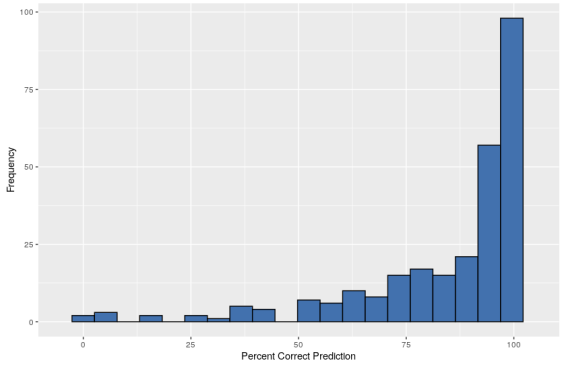
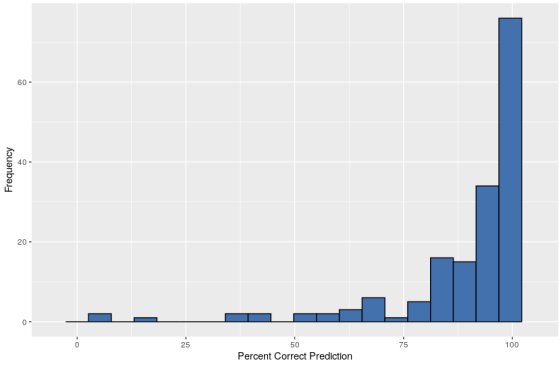


Figure 3.5: Average prediction percentage over all chemical group deletions

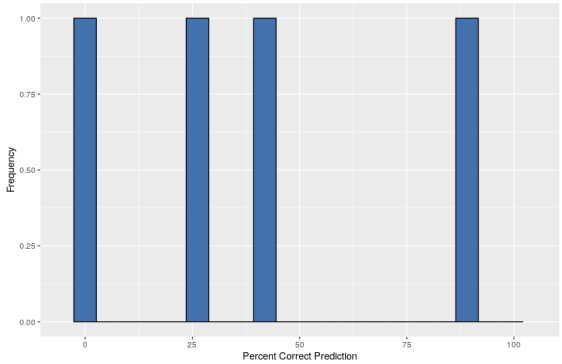
Figure 3.6 has the collection of frequency plots for correct prediction percentage for all 16 chemical groups in the protein. Values represented in Figure 3.5 are the average for each chemical group shown in Figure 3.6. It can be observed from the plots in Figure 3.6 that the prediction accuracy for the frequently occurring chemical groups like r1, r2, r8 and r12 deletions is much higher compared to the chemical groups like r11, r14, r16 etc. that are not present abundantly in the protein.



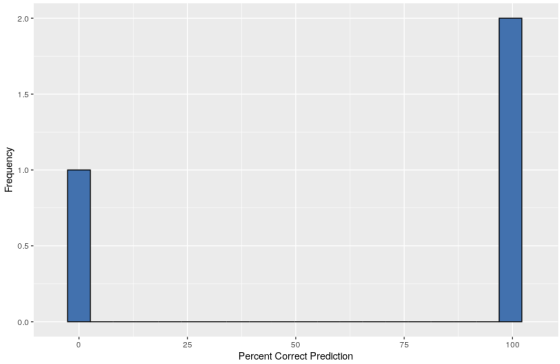
(a) r1 deletions



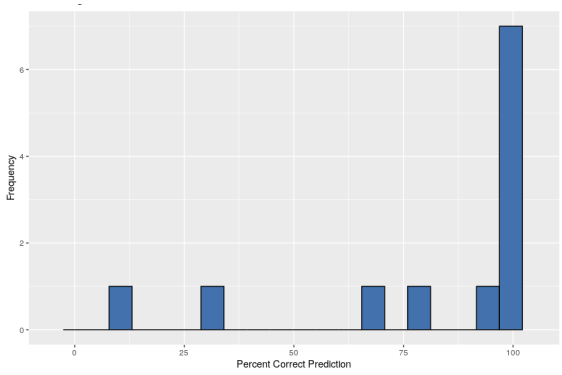
(b) r2 deletions



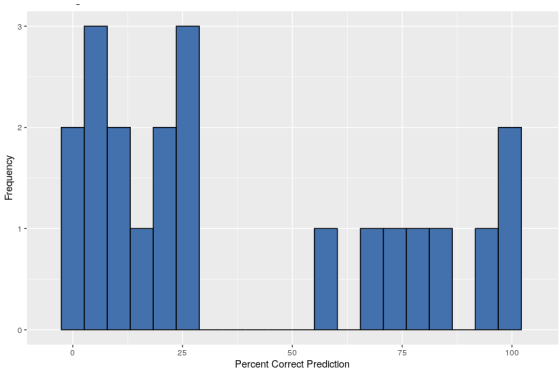
(c) r3 deletions



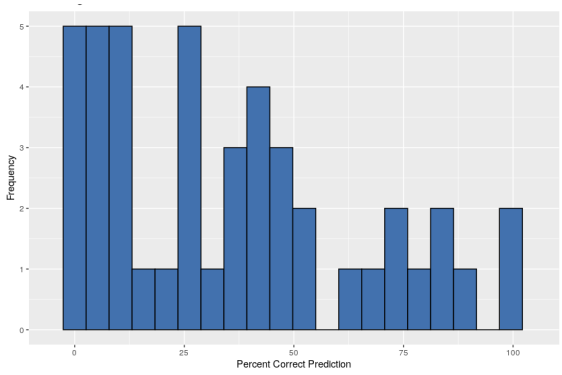
(d) r4 deletions



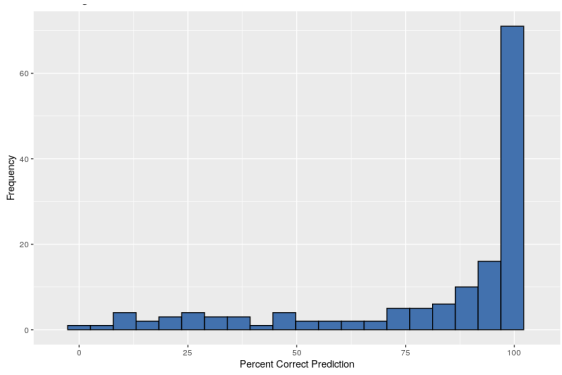
(e) r5 deletions



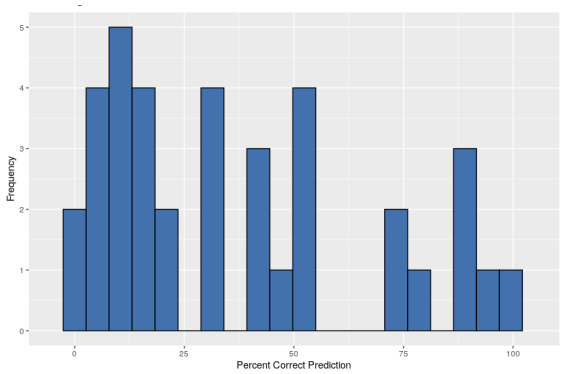
(f) r6 deletions



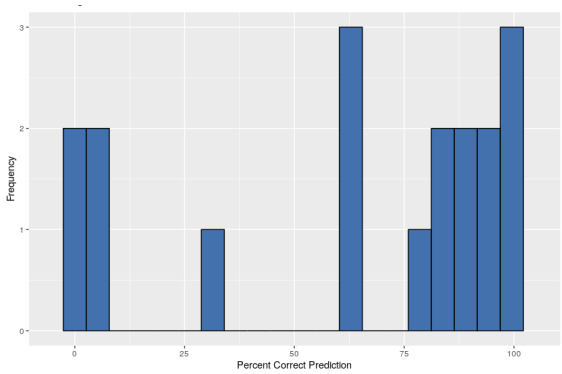
(g) r7 deletions



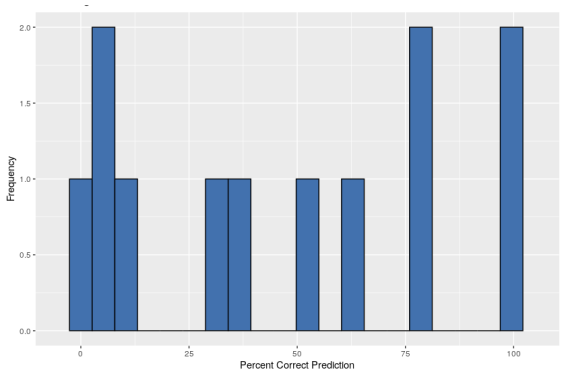
(h) r8 deletions



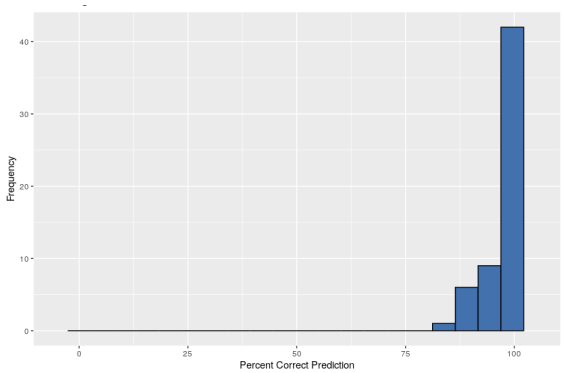
(i) r9 deletions



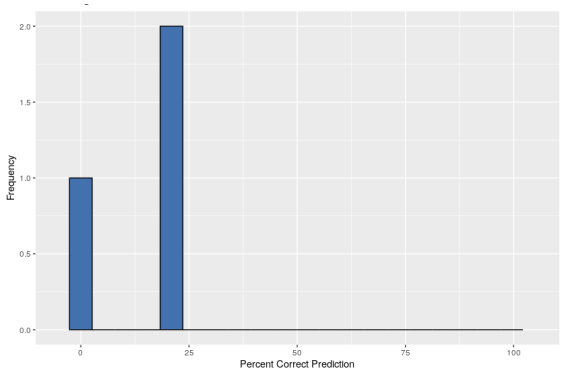
(j) r10 deletions



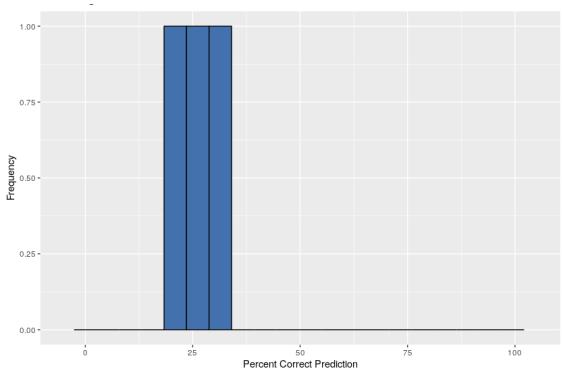
(k) r11 deletions



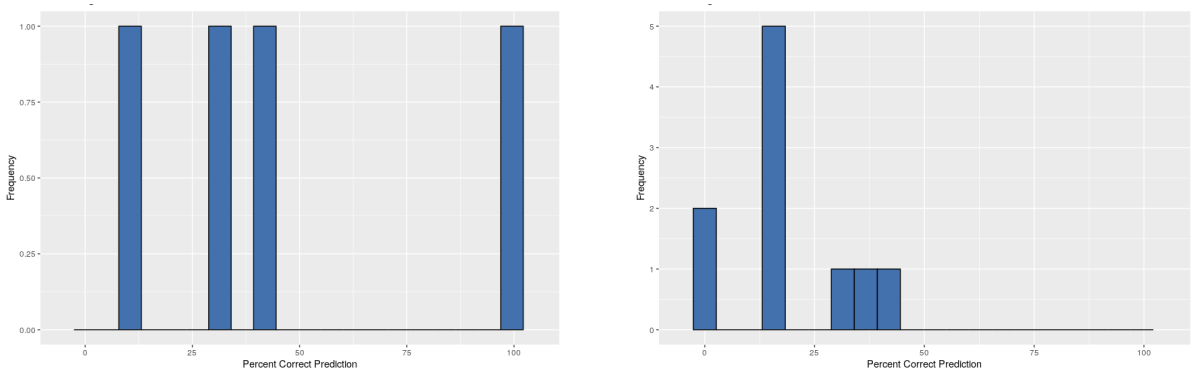
(l) r12 deletions



(m) r13 deletions



(n) r14 deletions



(o) r15 deletions (p) r16 deletions

Figure 3.6: Correct prediction percentage for individual chemical groups

When all the predictions are considered for a specific chemical group deletion, the chemical group that is predicted with majority, is considered as that deletion's top prediction. Histogram in Figure 3.7 shows the correct top prediction percent i.e. the percentage of times the predicted top chemical group is identical to the deleted one over all 16 chemical groups.

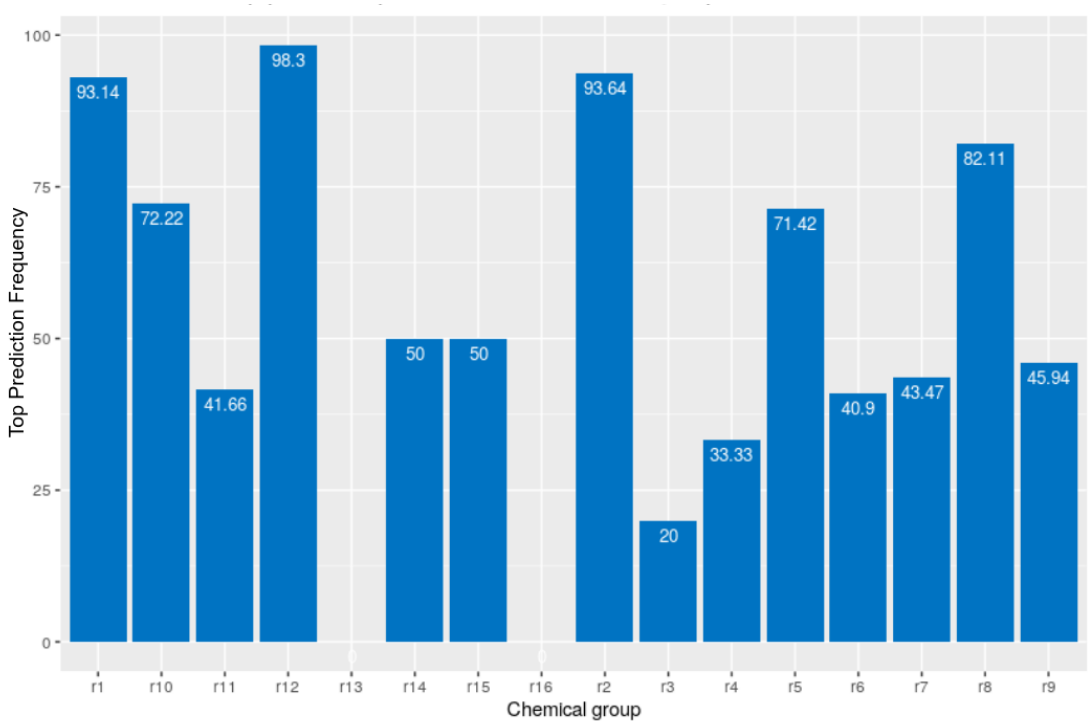


Figure 3.7: Percentage of correct Top predictions for all chemical group deletions

Out of the 16 chemical groups, r13 and r16 were predicted as the top prediction in none

of their individual deletions in the protein. Similar to the previous analysis, r1, r2, r8 and r12 chemical groups had much higher correct top predictions compared to other chemical groups. Out of the 842 deletions, 683(81.12%) had correct and 159(18.88%) had incorrect top prediction. A remarkable result for these predictions is the high accuracy in prediction for small chemical group deletions like r2, r8 and r12. This shows the sensitivity of the method to accurately differentiate between chemical groups that are inherently very similar to each other to give a correct prediction.

Chemical Group	**Correct Top Predictions Percentage**	
	with common chemical groups	**without common chemical groups**
r3	20.00	20.00
r4	33.33	33.33
r5	71.42	71.42
r6	40.90	50.00
r7	43.47	78.26
r9	45.94	78.37
r10	72.22	88.88
r11	41.66	75.00
r13	0	50.00
r14	50.00	50.00
r15	50.00	50.00
r16	0	0

Table 3.3: Variation in correct Top prediction accuracy with and without common chemical group predictions

But the method is failing to make correct predictions for deletions of chemical groups that are relatively larger in size compared to r2, r8 and r12. A potential reason for this can be the cavity size left behind after a deletion, since the small chemical groups can fit into these big cavities but the other way round is not possible reducing the prediction accuracy for chemical groups other than r2, r8 and r12. To validate this theory, we removed the r1, r2, r8 and r12 predictions for the other chemical group(r3, r4, r5, r6, r7, r9, r10, r11, r13, r14, r15, r16) deletions from the protein and then examined the changes in top prediction accuracy for these deletions. Table 3.3 depicts the difference in top chemical group prediction accuracy for the other chemical groups[1]. We are considering r1 as an exception here because all amino acids except Proline will have an r1 chemical group and due to it's abundant nature in the database, it will be predicted correctly more often compared to the other large

[1]In the table, 'common chemical group' refers to r1, r2, r8 and r12 chemical groups.

chemical groups.

Out of the 12 chemical group deletion types, 6 observed an increase in their correct top prediction percentage depicting an increase in deletion cases where top prediction after removal of r1, r2, r8 and r12 chemical groups was identical to the deleted chemical group. This shows that a heuristic method like the one we used here is required for correctly predicting deletions that would leave a larger cavity behind.

3.3.3 More incorrect predictions on the surface than the core

In this analysis, we have tried to check for a correlation between the prediction accuracy for the developed method and the DEPTH of the deleted chemical group. Figure 3.8 shows the plot for correct prediction percentage and chemical group depth for deleted chemical groups from the protein sequentially. The r-squared value for the Pearson's correlation coefficient between the correct prediction percentage and the DEPTH for each chemical group is 0.012. This study infers that there is a very poor or no correlation between the prediction accuracy of a chemical group deletion at a specific position in a protein structure and it's distance from the surface of the protein.

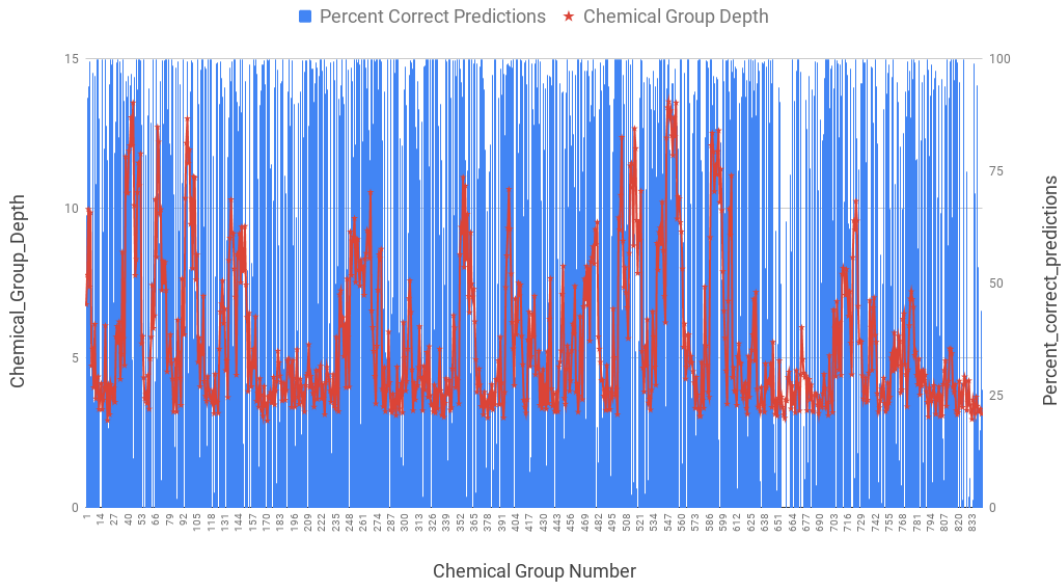


Figure 3.8: Sequential Correct Predication Percentage and Chemical Group DEPTH

We also compared the distribution for chemical group DEPTH for the deletions with correct and incorrect top predictions. Figure 3.9 shows this chemical group DEPTH distribution for both type of predictions. The total number of correct top predictions(683) in the protein is much higher compared to the total number of incorrect top predictions(159). Hence, we have normalized the y-axis with respect to the total correct and incorrect predictions respectively to compare for the difference in prediction accuracy for both the cases.
The residues present on the surface of the protein are more likely to undergo mutation compared to the residues present in the core of the protein. Because their is a higher penalty associated with mutations at the core as compared to the surface since the stability of protein molecule is much higher at it's core than on the surface.
This results suggests that, on the surface, where the chemical groups are more prone to undergo mutation as compared to the core, we observe that our method makes more mistakes giving higher fraction of incorrect predictions. Whereas, when we increase the DEPTH, the prediction accuracy improves showing the ability of our method to perform better given a more stable neighbourhood.

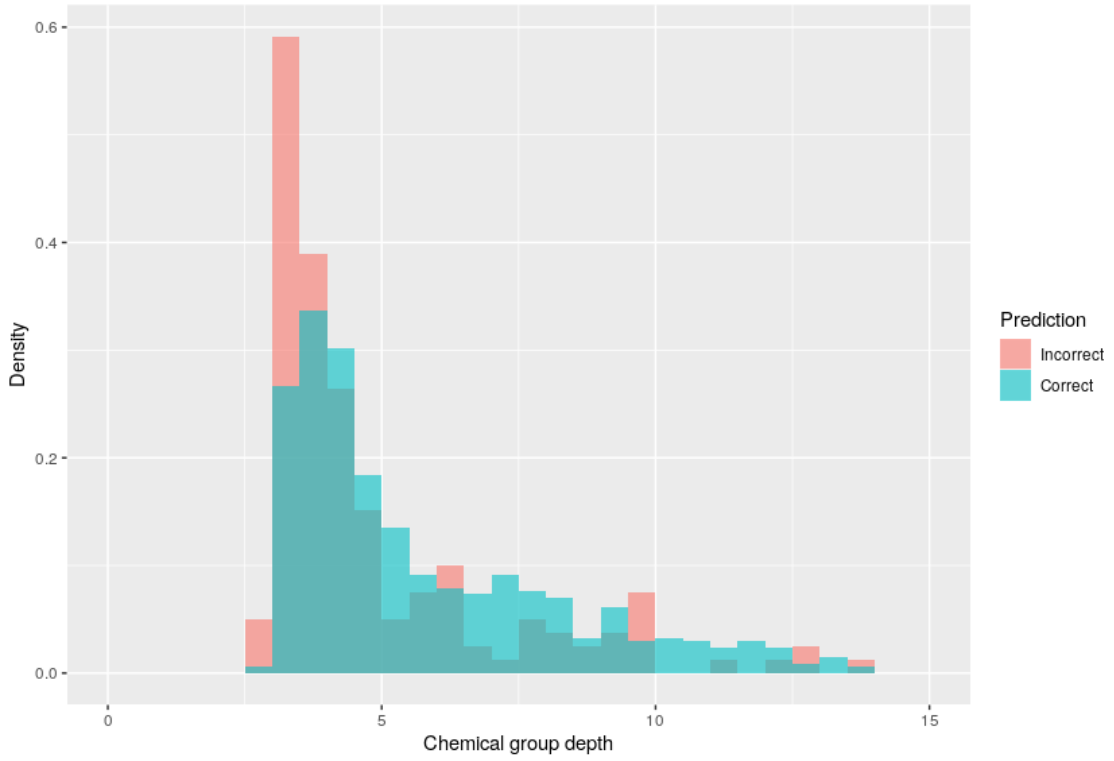


Figure 3.9: Distribution of Chemical group depth for Correct and Incorrect top predictions

3.3.4 Incorrect predictions inconsistent with Conservation Profiles

Multiple Sequence Alignment(MSA) is obtained for the protein 1Z7K using PSI-BLAST [Altschul et al., 1997] on a selection of top 500 protein sequences over 5 iterations. Figure 4.1 in appendix shows the graphical representation of the Multiple Sequence Alignment obtained by Weblogo 3 [Schneider and Stephens, 1990, Crooks et al.,2004].

We performed a total of 842 chemical group deletions for this protein. Out of this 842, 182 were the rare chemical group[2] deletions. Of which, 81 had the correct top prediction and the remaining 101 had an incorrect top prediction. For this analysis, we look at these 101 deletions to check if these predictions are reflected in the proteins conservation profile. Out of these 101, 70 belong to the chain A of the protein(we are only looking at this one chain in this analysis). For the 70 rare chemical group deletions that had incorrect top prediction, only 21 predicted another rare chemical group as their top prediction. Here, we are not looking at the cases with r1, r2, r8 and r12 as the top prediction because these chemical groups are part of multiple Amino acids and can't be effectively used to compare with the Amino acids conservation profiles. For instance, a rare chemical group deletion with r1 as the incorrect top prediction can represent 19 out of the 20 Amino acids in the prediction and will obviously find match with the Amino acid present in the conservation profile. Table 3.4 provides details about the Amino acid predictions and presence in the MSA for those 21 rare chemical group deletions with incorrect rare chemical group prediction.

Out of the 21 observations, almost all the incorrect deletions did not have any common Amino acid in the prediction and the conservation profile. Only one deletion of the chemical group at the 615^{th} position had one Amino acid Phenylalanine common in the prediction and the conservation profile. This analysis shows that the incorrect predictions made by our method are inconsistent with the conserved Amino acids at that position in the protein sequence. It can hence be inferred that our method is unable to suggest potential substitutions in the protein structure when compared to the protein's conservation profile for the individual Amino acids.

[2]The term 'rare chemical groups' here represents the r3, r4, r5, r6, r7, r9, r10, r11, r13, r14, r15 and r16 chemical groups.

Chemical Group Number	Amino Acids present in	
	Prediction	**Conservation Profile**
317	ASP + GLU	SER + ALA + PRO
378	ASP + GLU	SER + PHE
429	ASP + GLU	SER + THR + PRO
485	ASP + GLU	PHE
182	SER + THR	GLU
119	ASN + GLN	LYS
125	ASN + GLN	ARG + HIS
153	ASN + GLN	HIS + TYR
175	ASN + GLN	GLU
189	ASN + GLN	PHE
302	ASN + GLN	ARG + HIS + GLN + TYR
327	ASN + GLN	THR + SER + ARG
389	ASN + GLN	ASP + GLU
586	ASN + GLN	LYS + ARG + ASN
564	MET	TYR + ASP + HIS
344	PHE	GLN + GLU + LYS
37	PHE	TYR
560	PHE	TRP
615	PHE	TYR + PHE
386	TRP	TYR + ASN + GLU
362	TYR	TRP

Table 3.4: Comparison between the predicted and conserved Amino acids for deletions with incorrect top predictions

3.3.5 Smaller random stars dataset can be used instead of the whole stars database

The number of comparisons for superimposition to be performed for finding a hit for a query from the stars database is very large, sometimes even in the order of millions (refer to the last column in Table 4.1 from Appendix). This is computationally very expensive and requires a great deal of time too. To overcome this problem, in this subsection we suggest a way to reduce the number of computations. For this analysis, instead of using all stars in a composition for superimposition, we selected a random of 4000 to perform superimposition and predicted the deleted chemical groups from the same protein and compared it with the results for using all. Figure 3.10 shows the variation in top prediction percentage for

all 16 chemical groups using the whole nr_30 database and a random set of 4000 selected occurrences for each composition for the database.

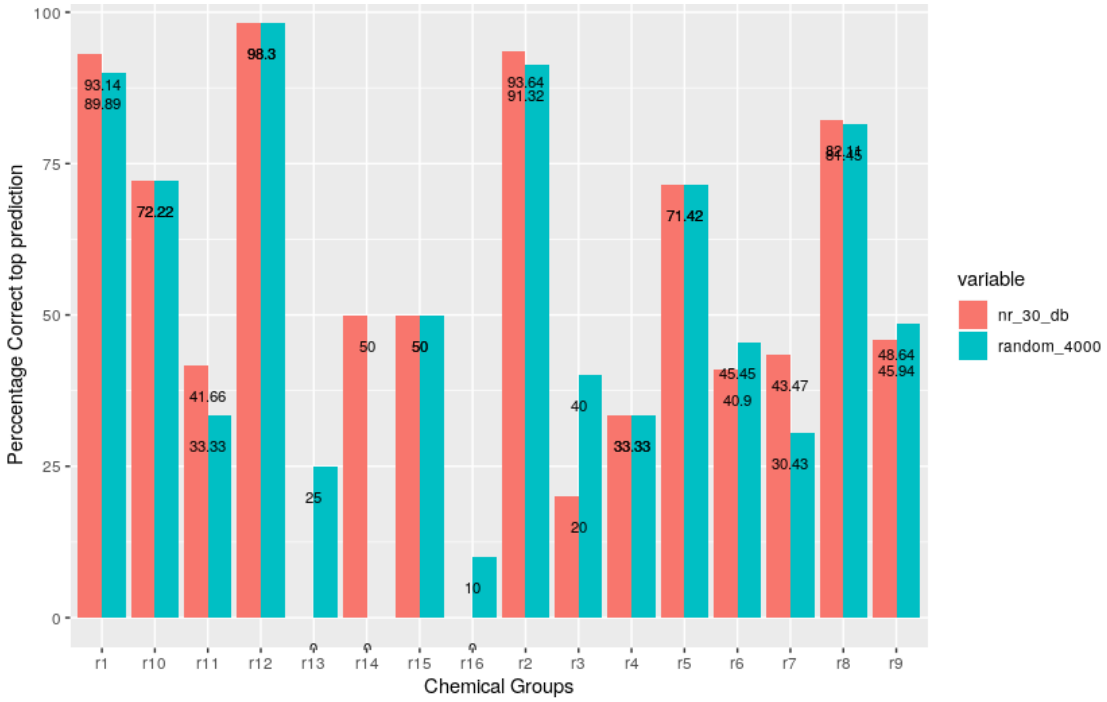


Figure 3.10: Variation in Top prediction accuracy for different datasets

For most of the chemical groups, the correct top prediction percentage is similar in both the cases, with r4, r5, r10, r12 and r15 having the same values in both the variants. For chemical groups like r7, r13, r14 and r16, the difference in prediction was higher compared to other chemical groups. For r14 the predictions with the nr_30 database were correct in 50% of the deletions, but none of the deletions had a correct prediction with the randomly selected small dataset. For r13 and 16 chemical groups, the opposite was observed, where none of the top predictions with the nr_30 database were identical to the deleted chemical group.

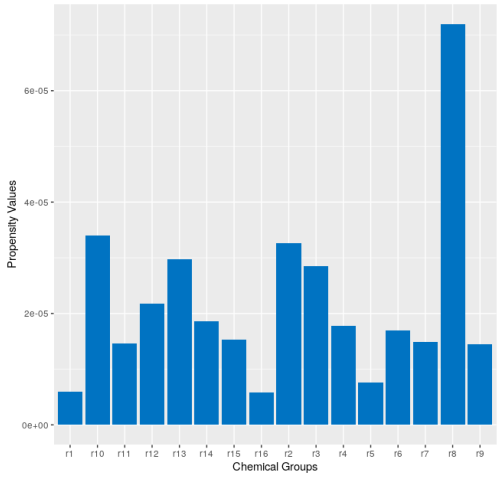
Paired Wilcoxon test was performed to check if the predictions between these variants of the dataset are significantly different. For V=32 and α=0.05, the p-value for the above data (represented in Figure 3.10) is 0.9645, which is higher than α. Hence failing to reject the null hypothesis that the two sets of predictions are similar.
This suggests that a smaller subset of randomly chosen stars can theoretically be used for prediction of the deleted chemical groups in order to save the computational time and power,

since the prediction accuracy is not significantly different for the two cases.

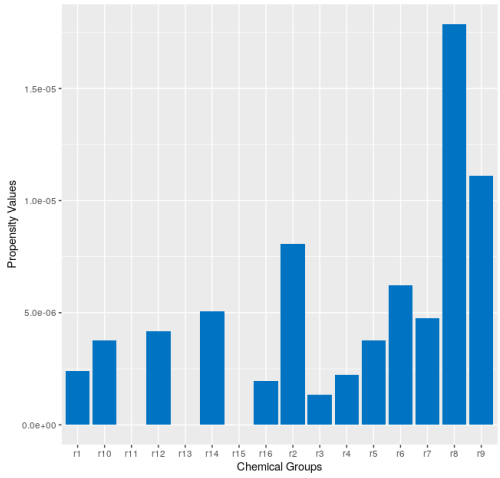
3.4 Individual query star for each chemical group in the binding site

We performed individual deletions for a total of 7 amino acid residues from a human Angiogenin protein(PDB id: 1ANG) and tried to predict them back using the method described in section 2.7. Figure 3.11 shows the plots for propensity values for prediction of all chemical groups for each deletion of the amino acid residues. Propensity value is calculated by normalizing the total prediction count of a chemical group with its frequency in the database. In some cases(as shown in Figure 3.11 (f)), a demarcation can be observed between the expected chemical groups(r7 and r8) and the others, whereas, in some cases(as shown in Figure 3.11(c)), the method was not able to predict the deleted chemical groups(r4).
r1 is not considered as part of this analysis because it is a default chemical group for all amino acid residues(except Proline).

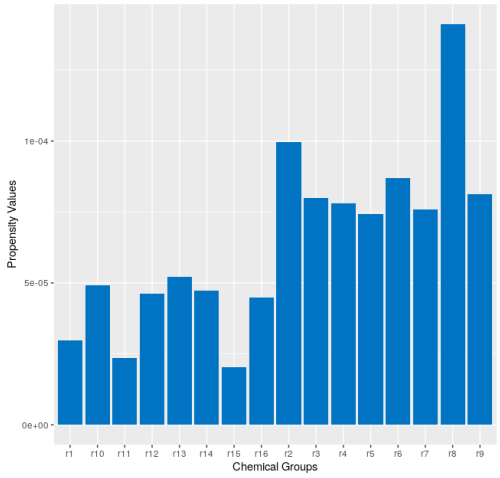
It was after this analysis, we realised that when considering a separate star with each chemical group in the binding site at its centre, we were considering a large number of chemical groups in our query star that are not present in the binding site. To account for this, we decided to go with a simpler approach(section 2.5) where the query star is made completely based on the chemical groups present in the binding site of the given protein structure.



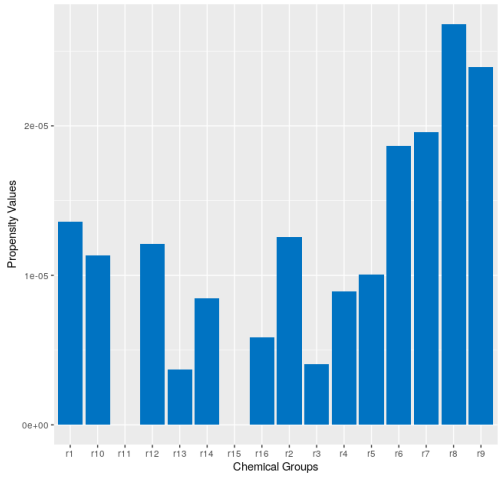
(a) ASN_68



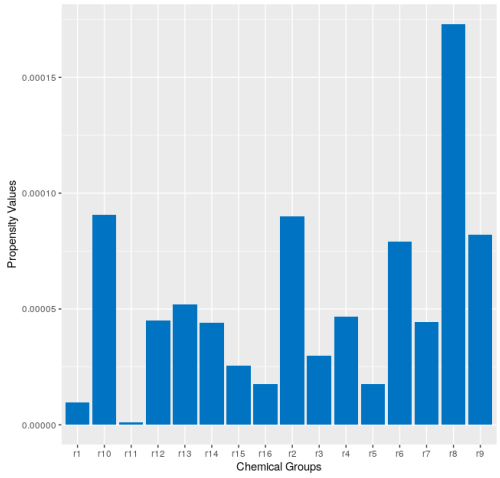
(b) ASN_102



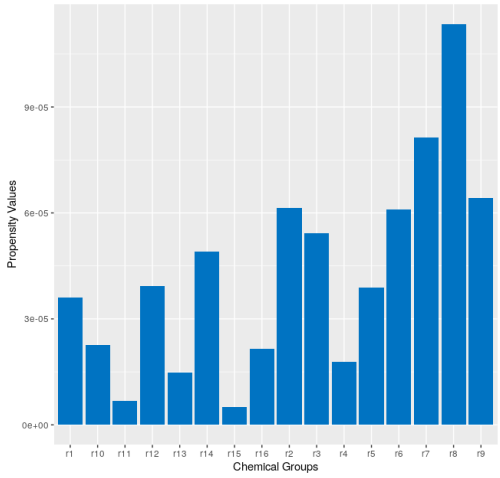
(c) HIS_8



(d) HIS_13



(e) LYS_40



(f) THR_79

Figure 3.11: Propensity values for the predicted chemical groups in the missing residue case

Chapter 4

Conclusion

The main objective of this study is to build a method that will allow us to predict or design peptide ligands for a given query protein structure based on the local packing of atoms in protein structures in the PDB. Usually, docking tools are used to obtain a protein-peptide complex model, which consider a library of peptides and sample all possible conformations of the peptides onto the protein structure to find an optimal binding pose. But this is computationally expensive and time-consuming. Here, we propose a method that predicts the sequence and conformation of the peptide that would bind to the given query protein structure. The proposed method is based on the assumption that a frequently observed structural feature in nature corresponds to a low energy state, hence, a stable conformation. In this study, we have laid the groundwork required for building this method and performed a set of validation tests to assess the working of the proposed method.

The method extracts a binding site structural motif defined as a query star in terms of chemical groups and searches the PDB database for another motif that is structurally similar to this. Initially, we were creating an individual query star for all chemical groups present in the binding site and searching the database for stars that were of the same size as the query star and predictions were made by extending the hit stars from the database. The problem with this approach was that the inclusion of non-surface chemical groups in the query stars. Therefore, we shifted to a simpler approach where the query stars only have the chemical groups that are part of the query protein's binding site. The peptide that would bind the query protein structure was predicted by extracting the neighbours for the hit stars from the

database.

To examine the credibility of our proposed method, we performed the "Missing chemical group" validation test where a chemical group is deleted from a protein structure and predictions are made using our method. The method was able to predict the correct chemical group as the top prediction in approximately 81% of the total deletions. All these predictions were performed without providing any amino acid information for the deleted chemical group, this shows the method's ability to accurately utilise the deletion's surrounding to search for a similar structural motif in the protein database. The prediction accuracy was very high for chemical groups like r1, r2, r8 and r12 that are present abundantly in the database, but the method was also able to efficiently differentiate between r2, r8 and r12 chemical group deletions, which are very similar to each other. This increases our confidence in the method for correctly predicting small chemical group deletions. For larger chemical groups, applying a heuristic method improved the prediction accuracy.
No proper correlation was observed between the individual prediction accuracy of a deletion and its DEPTH. But when the correct and incorrect top predictions were compared based on the distribution of their chemical group DEPTH values, it was observed that the method was making more mistakes i.e., higher incorrect top predictions for the surface chemical group deletions and had better accuracy with higher DEPTH i.e., the more stable chemical groups in the protein structure. No pattern was observed when incorrect predictions were compared against the protein's Multiple Sequence Alignment for checking potential substitutions.

The proposed method can be of significant importance for determining the quality of protein structures by assessing the packing of atoms in the structure, which is very essential in protein structure modeling. It can also be used for the completion of protein structures with missing atoms. We expect the method to have higher accuracy in predicting missing details from a protein structure because a larger packing of chemical groups will provide more details about the neighbourhood refining the resulting predictions.

4.1 Future Perspectives

In the future, we plan to step-wise increase the complexity of the validation tests for the assessment of the method. The Missing chemical group validation will be followed by the "Missing residue test" where an amino acid residue is deleted from the protein structure,

which is followed by the deletion of a group of amino acids and predicting them back. The final validation test would be deleting peptides from know protein-peptide complexes and comparing predictions with the known peptide ligand.

Once the method is validated, we plan to perform the following:

(a) Benchmark the results on the dataset of all experimentally determined protein-peptide complexes and calculating the accuracy of the method

(b) Use the database of known peptide binders in their unbound(*apo*) state to check for the difference, if any in predictions using our method

(c) Comparing the developed method against the other pre-existing protein-peptide complex prediction software in the field to check the effectiveness and efficiency of the method

This method can then be extended to prediction of other small molecule ligands by categorizing them based on their similarity to the amino acid chemical groups.

Bibliography

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25(17), 3389–3402. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9254694

[Berggård et al. 2007] Berggård, T., Linse, S., & James, P. (2007, August 1). Methods for the detection and analysis of protein-protein interactions. Proteomics, Vol. 7, pp. 2833–2842. https://doi.org/10.1002/pmic.200700131

[Berman et al. 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The Protein Data Bank. In Nucleic Acids Research (Vol. 28). Retrieved from http://www.rcsb.org/pdb/status.html

[Besray Unal et al. 2010] Besray Unal, E., Gursoy, A., & Erman, B. (2010). Vital: Viterbi algorithm for de novo peptide design. PLoS ONE, 5(6), e10926. https://doi.org/10.1371/journal.pone.0010926

[Bohnuud et al. 2017] Bohnuud, T., Jones, G., Schueler-Furman, O., & Kozakov, D. (2017). Detection of peptide-binding sites on protein surfaces using the peptimap server. In Methods in Molecular Biology (Vol. 1561, pp. 11–20). https://doi.org/10.1007/978-1-4939-6798-8_2

[Bradshaw and Waksman, 2002] Bradshaw, J. M., & Waksman, G. (2002, January 1). Molecular recognition by SH2 domains. Advances in Protein Chemistry, Vol. 61, pp. 161–210. https://doi.org/10.1016/S0065-3233(02)61005-8

[Bruzzoni-Giovanelli et al. 2018] Bruzzoni-Giovanelli, H., Alezra, V., Wolff, N., Dong, C. Z., Tuffery, P., & Rebollo, A. (2018, February 1). Interfering peptides targeting protein–protein interactions: the next generation of drugs? Drug Discovery Today, Vol. 23, pp. 272–285. https://doi.org/10.1016/j.drudis.2017.10.016

[Campillos et al. 2008] Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. Science, 321(5886), 263–266. https://doi.org/10.1126/science.1158140

[Chakravarty and Varadarajan, 1999] Chakravarty, S., & Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. Structure (London, England: 1993), 7(7), 723–732. https://doi.org/10.1016/s0969-2126(99)80097-5

[Chen and Kihara, 2011] Chen, H., & Kihara, D. (2011). Effect of using suboptimal alignments in template-based protein structure prediction. Proteins: Structure, Function and Bioinformatics, 79(1), 315–334. https://doi.org/10.1002/prot.22885

[Ciemny et al. 2018] Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O., & Kmiecik, S. (2018, August 1). Protein–peptide docking: opportunities and challenges. Drug Discovery Today, Vol. 23, pp. 1530–1537. https://doi.org/10.1016/j.drudis.2018.05.006

[Crooks et al.,2004] Crooks, G. E., Hon, G., Chandonia, J. M., Brenner, S. E. (2004). WebLogo: A sequence logo generator. Genome Research, 14(6), 1188–1190. https://doi.org/10.1101/gr.849004

[Crystallogr. Made Cryst. Clear, 2006] Crystallography Made Crystal Clear. (2006). In Crystallography Made Crystal Clear. https://doi.org/10.1016/b978-0-12-587073-3.x5000-4

[Dhawanjewar et al., 2019] Dhawanjewar, A. S., Roy, A. A., & Madhusudhan, M. S. (2019). A knowledge-based scoring function to assess the stability of quaternary protein assemblies. BioRxiv, 562520. https://doi.org/10.1101/562520

[Diller et al., 2015] Diller, D. J., Swanson, J., Bayden, A. S., Jarosinski, M., & Audie, J. (2015, October 1). Rational, computer-enabled peptide drug design: Principles, methods, applications and future directions. Future Medicinal Chemistry, Vol. 7, pp. 2173–2193. https://doi.org/10.4155/fmc.15.142

[Fernandez-Fuentes et al., 2010] Fernandez-Fuentes, N., Dybas, J. M., & Fiser, A. (2010). Structural Characteristics of Novel Protein Folds. PLoS Computational Biology, 6(4), e1000750. https://doi.org/10.1371/journal.pcbi.1000750

[Fosgerau and Hoffmann, 2015] Fosgerau, K., & Hoffmann, T. (2015, January 1). Peptide therapeutics: Current status and future directions. Drug Discovery Today, Vol. 20, pp. 122–128. https://doi.org/10.1016/j.drudis.2014.10.003

[Hammoudeh et al., 2009] Hammoudeh, D. I., Follis, A. V., Prochownik, E. V., & Metallo, S. J. (2009). Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. Journal of the American Chemical Society, 131(21), 7390–7401. https://doi.org/10.1021/ja900616b

[Johansson-Åkhe et al., 2019] Johansson-Åkhe, I., Mirabello, C., & Wallner, B. (2019). Predicting protein-peptide interaction sites using distant protein complexes as structural templates. Scientific Reports, 9(1), 1–13. https://doi.org/10.1038/s41598-019-38498-7

[Lavi et al., 2013] Lavi, A., Ngan, C. H., Movshovitz-Attias, D., Bohnuud, T., Yueh, C., Beglov, D., ... Kozakov, D. (2013). Detection of peptide-binding sites on protein surfaces: The first step toward the modeling and targeting of peptide-mediated interactions. Proteins: Structure, Function and Bioinformatics, 81(12), 2096–2105. https://doi.org/10.1002/prot.24422

[Litfin et al., 2019] Litfin, T., Yang, Y., & Zhou, Y. (2019). SPOT-Peptide: Template-Based Prediction of Peptide-Binding Proteins and Peptide-Binding Sites. Journal of Chemical Information and Modeling, 59(2), 924–930. https://doi.org/10.1021/acs.jcim.8b00777

[Metallo, 2010] Metallo, S. J. (2010, August 1). Intrinsically disordered proteins are potential drug targets. Current Opinion in Chemical Biology, Vol. 14, pp. 481–488. https://doi.org/10.1016/j.cbpa.2010.06.169

[Nguyen et al., 2011] Nguyen, M. N., Tan, K. P., & Madhusudhan, M. S. (2011). CLICK - Topology-independent comparison of biomolecular 3D structures. Nucleic Acids Research, 39(SUPPL. 2), W24–W28. https://doi.org/10.1093/nar/gkr393

[Nguyen et al., n.d.] Nguyen, Minh N, & Madhusudhan, M. S. (n.d.). Biological insights from topology independent comparison of protein 3D structures. https://doi.org/10.1093/nar/gkr348

[Pern Tan et al., 2013] Pern Tan, K., Varadarajan, R., & Madhusudhan, M. S. (2013). DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. https://doi.org/10.1093/nar/gkr356

[Schneider and Stephens, 1990] Schneider, T. D., Stephens, R. M. (1990). Sequence logos: A new way to display consensus sequences. Nucleic Acids Research, 18(20), 6097–6100. https://doi.org/10.1093/nar/18.20.6097

[Shapovalov and Dunbrack, 2011] Shapovalov, M. V., & Dunbrack, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure, 19(6), 844–858. https://doi.org/10.1016/j.str.2011.03.019

[Shoemaker and Panchenko, 2007] Shoemaker, B. A., & Panchenko, A. R. (2007). Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. PLoS Computational Biology, 3(4), e43. https://doi.org/10.1371/journal.pcbi.0030043

[Swastik Mishra Thesis, 2019] Swastik Mishra, MS thesis, 2019

[Vanhee et al., 2009a] Vanhee, P., Stricher, F., Baeten, L., Verschueren, E., Lenaerts, T., Serrano, L., ... Schymkowitz, J. (2009a). Protein-Peptide Interactions Adopt the Same Structural Motifs as Monomeric Protein Folds. Structure, 17(8), 1128–1136. https://doi.org/10.1016/j.str.2009.06.013

[Vanhee et al., 2009b] Vanhee, P., Stricher, F., Baeten, L., Verschueren, E., Lenaerts, T., Serrano, L., ... Schymkowitz, J. (2009b). Protein-Peptide Interactions Adopt the Same Structural Motifs as Monomeric Protein Folds. Structure, 17(8), 1128–1136. https://doi.org/10.1016/j.str.2009.06.013

[Verschueren et al., 2013] Verschueren, E., Vanhee, P., Rousseau, F., Schymkowitz, J., & Serrano, L. (2013). Protein-peptide complex prediction through fragment interaction patterns. Structure, 21(5), 789–797. https://doi.org/10.1016/j.str.2013.02.023

[Wang and Dunbrack, 2003] Wang, G., & Dunbrack, R. L. (2003). PISCES: A protein sequence culling server. Bioinformatics, 19(12), 1589–1591. https://doi.org/10.1093/bioinformatics/btg224

[Watkins et al., 2017] Watkins, A. M., Bonneau, R., & Arora, P. S. (2017). Modeling Peptide-Protein Interactions. Methods in Molecular Biology, 1561, 109–138. https://doi.org/10.1007/978-1-4939-6798-8

[Yaffe, 2002] Yaffe, M. B. (2002, March 1). Phosphotyrosine-binding domains in signal transduction. Nature Reviews Molecular Cell Biology, Vol. 3, pp. 177–186. https://doi.org/10.1038/nrm759

[Yan et al., 2017] Yan, C., Xu, X., & Zou, X. (2017). The usage of ACCLUSTER for peptide binding site prediction. In Methods in Molecular Biology (Vol. 1561, pp. 3–9). https://doi.org/10.1007/978-1-4939-6798-8_1

Appendix

Framework of the files

Following are the details of file formats used in this project and were previously developed by Swastik Mishra.

The gpdb file format:

The protein structure information in the PDB database is stored in a *.pdb* file. To represent protein structures in terms of chemical groups, a *.gpdb* format is made. It is very similar to the standard *.pdb* format, except the atom names and numbers are replaced by the chemical group names and numbers in the protein. Following is an example for illustration of the *.pdb* and the *.gpdb* format.

Sample *.pdb* lines:

```
ATOM 1 N ILE A 16 27.760 -32.484 36.747 1.00 20.06 N
ATOM 2 CA ILE A 16 27.185 -31.137 36.675 1.00 21.41 C
ATOM 3 C ILE A 16 26.315 -30.887 35.380 1.00 24.91 C
ATOM 4 O ILE A 16 26.832 -31.067 34.276 1.00 24.68 O
ATOM 5 CB ILE A 16 28.251 -30.113 36.742 1.00 24.83 C
ATOM 6 CG1 ILE A 16 29.158 -30.226 38.005 1.00 24.14 C
ATOM 7 CG2 ILE A 16 27.749 -28.753 36.523 1.00 26.08 C
ATOM 8 CD1 ILE A 16 30.236 -29.192 38.147 1.00 33.96 C
```

Sample *.gpdb* lines:

```
ATOM 1 r1 ILE A 16 26.627 -31.218 35.725
ATOM 2 r12 ILE A 16 28.251 -30.113 36.742
ATOM 3 r2 ILE A 16 29.158 -30.226 38.005
ATOM 4 r8 ILE A 16 27.749 -28.753 36.523
ATOM 5 r8 ILE A 16 30.236 -29.192 38.147
```

The cliqs file format:

The stars database is stored in *.cliqs* format, where each *.cliqs* file represent a unique star composition and has stars of the same composition from all *.gpdb* files from the database.
Sample *.cliqs* lines:

```
3b63 12202 12203 12205 12204 12200 12208 12206 12209 12210 12171
3b63 14420 14421 14423 14418 14422 14428 14426 14424 14391 14427
3bj5 0 1 2 8 5 6 3 4 143 7
3bjq 26 27 28 749 747 30 39 36 748 29
3bjq 5356 5357 7201 7107 5351 7106 5372 7202 7105 7199
3boq 258 259 260 262 263 264 291 261 247 265
3ikb 485 486 490 491 487 449 492 493 489 378
```

The above lines are from *r1_r1_r12_r12_r2_r8_r8_r8_r8_r8.cliqs* file, where r1(on the first position) is the centre of the star and the rest are arranged in alphanuerical order. Each line shows the star details which include pdb id of the protein it is extracted from and index of the chemical groups that are involved in the star, in order of their distance from the central chemical group.

Conservation Profile

The following figure is the sequence logo for the Multiple Sequence Alignment for chain A of the protein 1Z7K. This represents the Amino acid conservation profile at each position in the protein sequence.

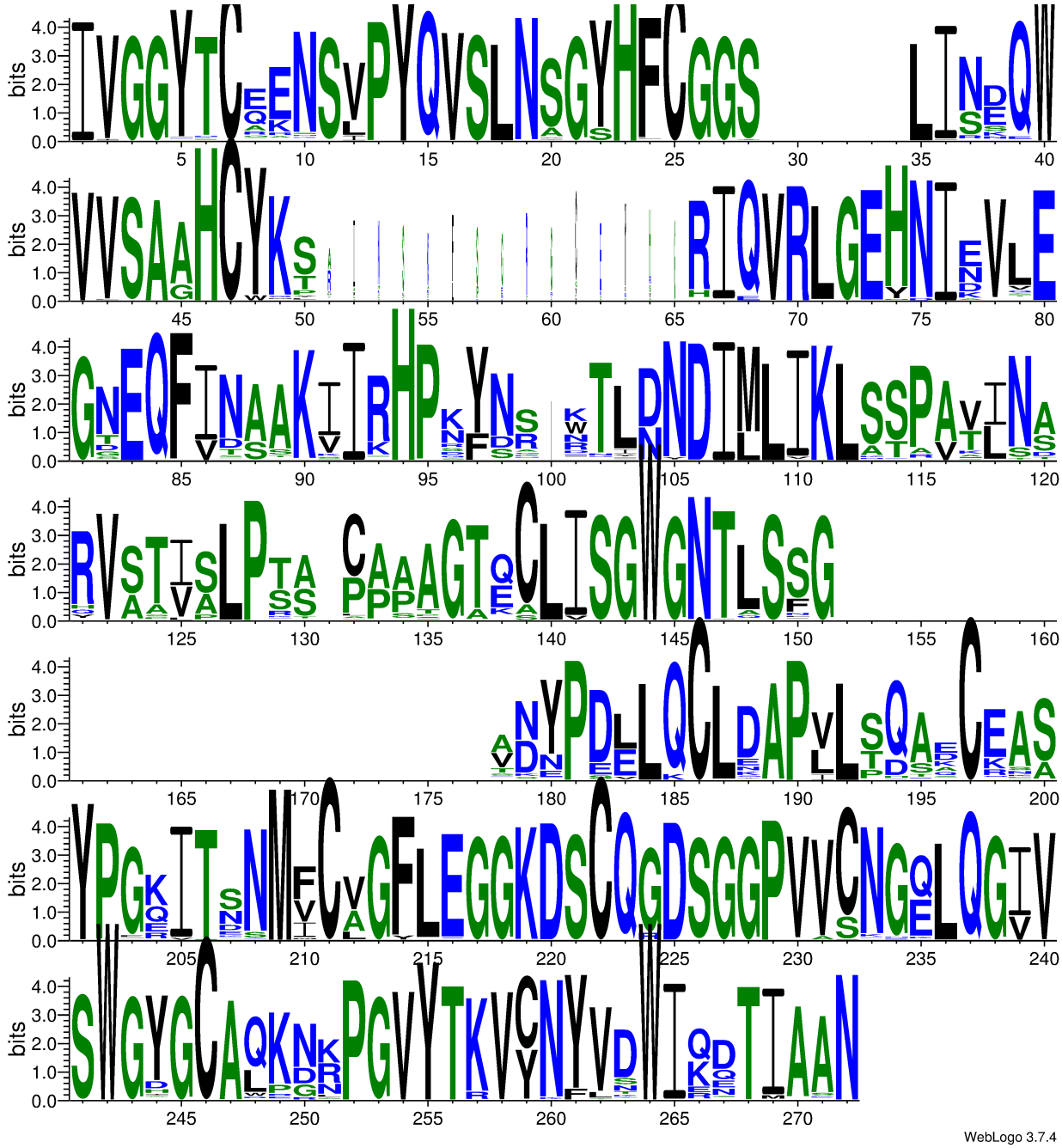


Figure 4.1: Multiple Sequence Alignment of the protein 1Z7K

Prediction details

The following table represents the prediction details for individual deletions from the protein 1Z7K using the nr30 stars database.

Table 4.1: Prediction details for all individual chemical group deletions from protein 1Z7K.

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	1	r1_r1_r12_r12_r2_r8_r8	6593	4301	65.23	219140
r12	2	r12_r1_r2_r2_r2_r8_r8	4208	3844	91.34	88685
r2	3	r2_r1_r12_r2_r6_r8_r8	228	214	93.85	63509
r8	4	r8_r1_r12_r2_r2_r2_r8	2797	2781	99.42	207466
r8	5	r8_r1_r1_r12_r2_r2_r8	24505	7198	29.37	1661263
r1	6	r1_r1_r1_r12_r8_r8_r8	110522	39407	35.65	1304116
r12	7	r12_r1_r1_r1_r8_r8_r8	12686	12281	96.80	311591
r8	8	r8_r1_r1_r12_r7_r8_r8	3522	826	23.45	401410
r8	9	r8_r1_r1_r1_r1_r12_r8	2202	2117	96.13	143285
r1	10	r1_r1_r1_r2_r2_r2_r8	1019	735	72.12	560126
r1	11	r1_r1_r1_r1_r16_r2_r8	25	25	100.00	69525
r1	12	r1_r1_r1_r16_r2_r7_r9	56	56	100.00	8171
r2	13	r2_r1_r1_r1_r1_r16_r7	60	52	86.66	13347
r16	14	r16_r1_r1_r10_r2_r2_r5	1	0	0	9651
r1	15	r1_r1_r1_r10_r2_r7_r8	174	174	100.00	16508
r7	16	r7_r1_r1_r2_r8_r8_r9	13	8	61.53	155735
r8	17	r8_r1_r1_r12_r2_r7_r8	345	276	80.00	524285
r1	18	r1_r1_r1_r10_r7_r8_r8	36	36	100.00	12496
r10	19	r10_r1_r1_r1_r1_r10_r2	81	69	85.18	18944
r1	20	r1_r1_r1_r1_r7_r8_r8	1344	1042	77.52	184875
r8	21	r8_r1_r1_r1_r1_r7_r8	198	35	17.67	147919
r1	22	r1_r1_r1_r2_r7_r8_r9	636	600	94.33	130533
r8	23	r8_r1_r1_r2_r4_r8_r9	1	1	100.00	10475
r1	24	r1_r1_r1_r1_r2_r7_r9	393	388	98.72	119735
r2	25	r2_r1_r1_r1_r2_r8_r9	24	24	100.00	247808
r9	26	r9_r1_r1_r2_r2_r3_r8	70	32	45.71	112512
r1	27	r1_r1_r1_r12_r2_r7_r8	6708	5301	79.02	524285

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r7	28	r7_r1_r1_r1_r1_r1_r8	985	847	85.98	113500
r1	29	r1_r1_r11_r12_r2_r8_r8	5614	5564	99.10	119620
r12	30	r12_r1_r1_r2_r8_r8_r8	159913	159858	99.96	871220
r2	31	r2_r1_r1_r10_r12_r8_r8	1651	1361	82.43	47724
r8	32	r8_r1_r12_r16_r2_r2_r8	416	414	99.51	31151
r8	33	r8_r1_r12_r2_r7_r8_r8	440	426	96.81	126449
r11	34	r11_r1_r1_r1_r1_r2_r8	42	15	35.71	343173
r1	35	r1_r1_r11_r16_r2_r2_r2	79	79	100.00	5761
r2	36	r2_r1_r1_r11_r12_r16_r8	32	29	90.62	3556
r16	37	r16_r1_r12_r2_r8_r8_r8	222	78	35.13	836752
r1	38	r1_r1_r1_r12_r2_r2_r8	102235	97080	94.95	1661263
r2	39	r2_r1_r1_r2_r8_r8_r9	4565	4182	91.61	155735
r2	40	r2_r1_r2_r7_r8_r8_r9	27	26	96.29	6970
r9	41	r9_r1_r1_r2_r2_r7_r8	193	151	78.23	402973
r1	42	r1_r1_r1_r12_r7_r8_r8	48898	45997	94.06	401410
r12	43	r12_r1_r1_r1_r8_r8_r8	60997	58988	96.70	311591
r8	44	r8_r1_r1_r12_r12_r8_r8	64051	21208	33.11	278939
r8	45	r8_r1_r1_r12_r8_r8_r8	181364	19882	10.96	1304116
r1	46	r1_r1_r1_r12_r2_r7_r8	24408	23413	95.92	524285
r7	47	r7_r1_r1_r1_r2_r4_r8	31	27	87.09	36545
r1	48	r1_r1_r1_r12_r14_r2_r2	1457	1457	100.00	37876
r2	49	r2_r1_r1_r1_r12_r8_r8	409773	382096	93.24	887582
r12	50	r12_r1_r1_r1_r2_r8_r8	214752	214646	99.95	866629
r8	51	r8_r10_r12_r2_r7_r8_r8	23	23	100.00	983
r8	52	r8_r1_r12_r2_r8_r8_r8	27834	27772	99.77	836752
r1	53	r1_r1_r1_r1_r2_r7_r9	471	402	85.35	119735
r2	54	r2_r1_r1_r1_r2_r3_r9	46	46	100.00	45325
r9	55	r9_r1_r1_r1_r1_r2_r3	30998	33	.10	243463
r1	56	r1_r1_r1_r1_r2_r7_r9	122	96	78.68	119735
r7	57	r7_r1_r1_r14_r2_r2_r5	91	21	23.07	36967
r1	58	r1_r1_r1_r1_r2_r7_r9	126	110	87.30	119735
r1	59	r1_r1_r1_r1_r2_r7_r9	107	100	93.45	119735

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r7	60	r7_r1_r1_r1_r1_r2_r9	53	15	28.30	267543
r1	61	r1_r1_r1_r2_r4_r7_r8	7	6	85.71	18853
r4	62	r4_r1_r1_r1_r2_r7_r8	2	0	0	315627
r1	63	r1_r1_r1_r10_r14_r2_r2	23	23	100.00	4697
r2	64	r2_r1_r1_r1_r14_r5_r9	0	0		228
r14	65	r14_r1_r1_r2_r2_r2_r5	203	57	28.07	300752
r1	66	r1_r1_r1_r10_r12_r2_r4	27	27	100.00	495
r10	67	r10_r1_r1_r1_r10_r14_r8	3	1	33.33	3590
r1	68	r1_r1_r1_r1_r12_r7_r8	10	7	70.00	91753
r1	69	r1_r1_r1_r12_r7_r8_r8	258	152	58.91	401410
r1	70	r1_r1_r1_r12_r2_r7_r8	1082	1029	95.10	524285
r7	71	r7_r1_r1_r12_r8_r8_r8	9433	576	6.10	1304116
r1	72	r1_r1_r1_r12_r12_r2_r8	37090	36940	99.59	272697
r2	73	r2_r1_r1_r12_r8_r8_r8	277675	237588	85.56	1304116
r12	74	r12_r1_r1_r2_r8_r8_r8	46786	46650	99.70	871220
r8	75	r8_r1_r12_r2_r8_r8_r8	56965	56914	99.91	836752
r8	76	r8_r12_r2_r8_r8_r8_r8	17168	17168	100.00	291541
r1	77	r1_r1_r1_r12_r2_r2_r8	77633	68964	88.83	1661263
r12	78	r12_r1_r1_r2_r8_r8_r8	69573	69549	99.96	871220
r2	79	r2_r1_r1_r12_r8_r8_r8	85874	69610	81.06	1304116
r8	80	r8_r1_r1_r12_r2_r8_r8	52700	41434	78.62	2027886
r8	81	r8_r1_r12_r2_r8_r8_r8	10407	10397	99.90	836752
r1	82	r1_r1_r1_r2_r7_r8_r9	380	374	98.42	130533
r2	83	r2_r1_r1_r1_r2_r8_r9	99	69	69.69	247808
r9	84	r9_r1_r1_r1_r2_r2_r2	18707	2549	13.62	537457
r1	85	r1_r1_r1_r2_r2_r7_r8	1388	1360	97.98	402973
r7	86	r7_r1_r1_r11_r2_r2_r9	145	3	2.06	39171
r1	87	r1_r1_r1_r12_r2_r2_r2	14783	14713	99.52	364903
r2	88	r2_r1_r1_r2_r2_r7_r9	198	122	61.61	116008
r2	89	r2_r1_r1_r2_r2_r7_r9	141	133	94.32	116008
r9	90	r9_r1_r1_r2_r2_r5_r7	33	14	42.42	125071
r1	91	r1_r1_r1_r12_r2_r2_r8	13496	9763	72.33	1661263

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	92	r2_r1_r1_r15_r2_r8_r9	26	26	100.00	5701
r15	93	r15_r12_r2_r2_r2_r2_r8	479	194	40.50	24397
r1	94	r1_r1_r1_r12_r12_r8_r8	71136	51321	72.14	278939
r12	95	r12_r1_r1_r12_r8_r8_r8	16722	16618	99.37	1304116
r8	96	r8_r1_r1_r1_r12_r8_r8	53629	16036	29.90	887582
r8	97	r8_r1_r1_r12_r8_r8_r8	102599	54634	53.25	1304116
r1	98	r1_r1_r1_r12_r7_r8_r8	54421	50469	92.73	401410
r12	99	r12_r1_r1_r8_r8_r8_r8	17297	17227	99.59	159241
r8	100	r8_r1_r12_r2_r8_r8_r8	4734	4693	99.13	836752
r8	101	r8_r1_r1_r12_r8_r8_r8	151550	56991	37.60	1304116
r1	102	r1_r1_r1_r2_r7_r8_r8	290	194	66.89	333145
r7	103	r7_r1_r1_r1_r1_r1_r8	153	74	48.36	113500
r1	104	r1_r1_r1_r1_r2_r8_r8	4293	1658	38.62	866629
r8	105	r8_r1_r1_r10_r4_r6_r8	93	93	100.00	494
r1	106	r1_r1_r1_r1_r14_r2_r8	30	27	90.00	79345
r8	107	r8_r1_r1_r1_r1_r14_r8	3178	361	11.35	65209
r1	108	r1_r1_r1_r1_r10_r2_r4	40	39	97.50	3547
r4	109	r4_r1_r2_r6_r7_r8_r8	2	2	100.00	4941
r1	110	r1_r1_r1_r1_r10_r13_r2	46	46	100.00	2263
r10	111	r10_r1_r1_r1_r10_r7_r8	118	74	62.71	15957
r1	112	r1_r1_r1_r16_r2_r2_r2	8814	8661	98.26	87190
r2	113	r2_r1_r1_r1_r13_r16_r8	178	174	97.75	5086
r16	114	r16_r1_r1_r12_r2_r8_r8	245	34	13.87	2027886
r1	115	r1_r1_r1_r16_r2_r2_r8	27	27	100.00	56395
r2	116	r2_r1_r1_r2_r2_r2_r8	1932	697	36.07	560126
r2	117	r2_r1_r1_r14_r2_r2_r5	2317	2317	100.00	36967
r2	118	r2_r1_r14_r2_r2_r5_r7	59	59	100.00	1725
r5	119	r5_r14_r2_r2_r2_r7_r9	0	0		261
r1	120	r1_r1_r1_r2_r7_r8_r8	343	269	78.42	333145
r7	121	r7_r1_r1_r1_r2_r2_r2	18505	636	3.43	537457
r1	122	r1_r1_r12_r2_r2_r2_r8	2015	1040	51.61	207466
r2	123	r2_r1_r1_r1_r2_r3_r7	434	430	99.07	65399

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	124	r2_r1_r1_r1_r2_r3_r7	511	507	99.21	65399
r3	125	r3_r1_r1_r1_r2_r2_r9	0	0		286355
r1	126	r1_r1_r1_r12_r2_r2_r8	209991	189039	90.02	1661263
r12	127	r12_r1_r1_r2_r8_r8_r8	154879	154152	99.53	871220
r2	128	r2_r1_r1_r12_r2_r8_r8	55895	53556	95.81	2027886
r8	129	r8_r1_r1_r12_r2_r2_r8	13879	12716	91.62	1661263
r8	130	r8_r1_r1_r12_r2_r8_r8	39423	18686	47.39	2027886
r1	131	r1_r1_r1_r12_r2_r2_r8	105447	98768	93.66	1661263
r2	132	r2_r1_r1_r1_r2_r2_r9	16900	10081	59.65	286355
r2	133	r2_r1_r1_r1_r2_r2_r9	3922	3064	78.12	286355
r9	134	r9_r1_r2_r2_r2_r2_r3	54	47	87.03	26937
r1	135	r1_r1_r1_r12_r2_r2_r8	93922	91991	97.94	1661263
r12	136	r12_r1_r1_r1_r8_r8_r8	76089	72255	94.96	311591
r8	137	r8_r1_r12_r12_r8_r8_r8	8016	7978	99.52	99927
r8	138	r8_r1_r1_r12_r8_r8_r8	105242	37910	36.02	1304116
r1	139	r1_r1_r1_r12_r2_r2_r2	51138	51089	99.90	364903
r2	140	r2_r1_r1_r2_r3_r6_r7	617	616	99.83	23680
r2	141	r2_r1_r1_r14_r2_r2_r3	55	51	92.72	18011
r3	142	r3_r14_r2_r2_r2_r9_r9	31	28	90.32	226
r1	143	r1_r1_r1_r12_r2_r8_r8	28119	23573	83.83	2027886
r2	144	r2_r1_r1_r12_r8_r8_r8	268459	240448	89.56	1304116
r12	145	r12_r1_r1_r2_r8_r8_r8	189061	189016	99.97	871220
r8	146	r8_r1_r1_r12_r2_r8_r8	31343	27593	88.03	2027886
r8	147	r8_r12_r2_r8_r8_r8_r8	31407	31406	99.99	291541
r1	148	r1_r1_r1_r2_r2_r2_r3	112	16	14.28	107568
r1	149	r1_r1_r1_r1_r2_r3_r6	1	0	0	96234
r2	150	r2_r1_r1_r1_r1_r6_r8	285	271	95.08	137169
r6	151	r6_r1_r2_r2_r2_r2_r8	47	13	27.65	65951
r1	152	r1_r1_r1_r2_r2_r4_r8	17	9	52.94	31267
r4	153	r4_r1_r1_r2_r6_r8_r8	0	0		139365
r1	154	r1_r1_r1_r12_r2_r8_r9	14745	14561	98.75	184439
r2	155	r2_r1_r1_r12_r4_r8_r9	0	0		1742

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r9	156	r9_r1_r12_r2_r7_r8_r8	33	18	54.54	126449
r1	157	r1_r1_r1_r12_r2_r8_r8	61813	31811	51.46	2027886
r12	158	r12_r1_r1_r15_r2_r8_r8	283	283	100.00	26796
r2	159	r2_r1_r1_r12_r6_r8_r8	5648	5567	98.56	165713
r8	160	r8_r1_r1_r12_r15_r2_r8	678	678	100.00	26523
r8	161	r8_r1_r12_r2_r6_r7_r8	90	87	96.66	10600
r1	162	r1_r1_r1_r12_r6_r8_r9	101	74	73.26	9105
r6	163	r6_r1_r1_r2_r7_r8_r9	11	1	9.09	130533
r1	164	r1_r1_r1_r12_r2_r8_r8	210794	189120	89.71	2027886
r12	165	r12_r1_r1_r2_r8_r8_r9	4327	4024	92.99	155735
r8	166	r8_r1_r1_r1_r12_r6_r8	1995	1970	98.74	73460
r8	167	r8_r1_r1_r12_r2_r8_r9	2740	2707	98.79	184439
r1	168	r1_r1_r1_r12_r2_r2_r8	13050	12998	99.60	1661263
r2	169	r2_r1_r1_r12_r6_r8_r8	15963	15056	94.31	165713
r12	170	r12_r1_r1_r2_r6_r8_r8	13570	13550	99.85	139365
r8	171	r8_r1_r1_r12_r14_r2_r8	921	919	99.78	88596
r8	172	r8_r1_r12_r14_r2_r6_r8	74	74	100.00	3079
r1	173	r1_r1_r1_r2_r6_r6_r9	41	31	75.60	10755
r2	174	r2_r1_r1_r6_r6_r8_r9	17	13	76.47	1618
r6	175	r6_r1_r1_r2_r4_r8_r9	0	0		10475
r1	176	r1_r1_r1_r2_r2_r6_r9	386	264	68.39	113200
r1	177	r1_r1_r1_r2_r2_r6_r9	1005	941	93.63	113200
r2	178	r2_r1_r1_r1_r2_r3_r9	19	19	100.00	45325
r9	179	r9_r1_r1_r1_r2_r2_r3	1699	93	5.47	173650
r1	180	r1_r1_r1_r2_r2_r2_r6	5563	5429	97.59	349950
r2	181	r2_r1_r1_r1_r1_r1_r6	497	291	58.55	76976
r6	182	r6_r1_r1_r1_r1_r1_r2	2012	386	19.18	199396
r1	183	r1_r1_r1_r2_r2_r2_r8	15639	14778	94.49	560126
r2	184	r2_r1_r1_r2_r2_r8_r9	4451	4234	95.12	231501
r2	185	r2_r1_r1_r2_r8_r8_r9	107	107	100.00	155735
r9	186	r9_r1_r1_r2_r2_r7_r8	232	72	31.03	402973
r1	187	r1_r1_r1_r12_r2_r2_r8	51735	38212	73.86	1661263

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	188	r2_r1_r1_r14_r2_r2_r9	87	61	70.11	32287
r14	189	r14_r1_r2_r2_r3_r8_r9	0	0		6523
r1	190	r1_r1_r1_r12_r2_r2_r8	332272	308339	92.79	1661263
r12	191	r12_r1_r1_r2_r8_r8_r8	71243	71186	99.91	871220
r2	192	r2_r1_r1_r12_r8_r8_r8	133354	113584	85.17	1304116
r8	193	r8_r1_r1_r12_r2_r2_r8	14907	13622	91.37	1661263
r8	194	r8_r1_r1_r12_r2_r7_r8	462	461	99.78	524285
r1	195	r1_r1_r1_r2_r8_r8_r9	1633	953	58.35	155735
r2	196	r2_r1_r1_r1_r7_r8_r9	30	27	90.00	35747
r9	197	r9_r1_r1_r2_r2_r2_r9	6736	923	13.70	337570
r1	198	r1_r1_r1_r1_r2_r8_r8	1108	716	64.62	866629
r8	199	r8_r1_r1_r12_r2_r8_r8	33	13	39.39	2027886
r1	200	r1_r1_r1_r2_r2_r8_r8	14572	13153	90.26	607919
r8	201	r8_r1_r1_r1_r1_r2_r7	969	246	25.38	157473
r1	202	r1_r1_r1_r1_r12_r2_r2	1183	1098	92.81	127788
r2	203	r2_r1_r1_r1_r2_r2_r2	15062	2454	16.29	537457
r2	204	r2_r1_r1_r2_r2_r2_r5	38963	38820	99.63	300752
r2	205	r2_r2_r2_r2_r2_r2_r5	46	38	82.60	11001
r5	206	r5_r2_r2_r2_r2_r2_r9	244	237	97.13	7899
r1	207	r1_r1_r1_r12_r12_r2_r8	35182	35107	99.78	272697
r12	208	r12_r1_r1_r1_r2_r8_r8	120455	120050	99.66	866629
r2	209	r2_r1_r1_r12_r8_r8_r8	244528	207883	85.01	1304116
r8	210	r8_r1_r12_r13_r2_r2_r8	686	685	99.85	21633
r8	211	r8_r1_r12_r2_r8_r8_r8	2900	2888	99.58	836752
r1	212	r1_r1_r1_r12_r2_r7_r8	50493	49413	97.86	524285
r12	213	r12_r1_r1_r2_r2_r8_r8	96215	94032	97.73	607919
r2	214	r2_r1_r1_r1_r12_r8_r8	148382	134186	90.43	887582
r8	215	r8_r1_r1_r12_r2_r8_r8	37087	35070	94.56	2027886
r8	216	r8_r12_r2_r2_r8_r8_r9	134	133	99.25	27224
r1	217	r1_r1_r1_r12_r7_r8_r8	8569	8402	98.05	401410
r7	218	r7_r1_r1_r1_r2_r8_r8	1061	482	45.42	866629
r8	219	r8_r1_r1_r2_r2_r7_r8	108	104	96.29	402973

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	220	r1_r1_r1_r11_r2_r4_r7	23	23	100.00	2826
r4	221	r4_r1_r11_r15_r2_r8_r9	23	23	100.00	409
r11	222	r11_r1_r1_r1_r2_r4_r9	76	74	97.36	21012
r1	223	r1_r1_r2_r2_r2_r9_r9	113	30	26.54	9230
r2	224	r2_r1_r11_r2_r4_r9_r9	14	14	100.00	179
r9	225	r9_r1_r1_r11_r2_r4_r9	13	12	92.30	2982
r1	226	r1_r1_r1_r14_r2_r2_r2	3542	3477	98.16	98724
r2	227	r2_r1_r1_r1_r1_r14_r8	234	218	93.16	65209
r14	228	r14_r1_r1_r1_r2_r8_r8	410	120	29.26	866629
r1	229	r1_r1_r1_r1_r14_r2_r9	34	14	41.17	28921
r2	230	r2_r1_r1_r1_r7_r8_r9	125	124	99.20	35747
r9	231	r9_r1_r1_r2_r2_r7_r8	5118	621	12.13	402973
r1	232	r1_r1_r1_r2_r8_r9_r9	92	27	29.34	24677
r1	233	r1_r1_r1_r2_r7_r8_r9	3717	3693	99.35	130533
r2	234	r2_r1_r1_r1_r8_r9_r9	269	256	95.16	6408
r9	235	r9_r1_r1_r1_r2_r2_r9	44	22	50.00	286355
r1	236	r1_r1_r1_r12_r2_r7_r8	6917	6916	99.98	524285
r7	237	r7_r1_r1_r1_r2_r8_r9	1500	807	53.80	247808
r8	238	r8_r1_r1_r2_r2_r7_r9	841	767	91.20	116008
r1	239	r1_r1_r1_r12_r2_r7_r8	1118	978	87.47	524285
r2	240	r2_r1_r1_r12_r6_r8_r8	11712	11017	94.06	165713
r12	241	r12_r1_r1_r2_r6_r8_r8	12035	12029	99.95	139365
r8	242	r8_r12_r2_r2_r2_r4_r8	12	11	91.66	1826
r8	243	r8_r1_r1_r1_r12_r2_r8	18361	16511	89.92	533479
r1	244	r1_r1_r1_r1_r2_r6_r9	503	207	41.15	103659
r6	245	r6_r1_r2_r7_r8_r9_r9	5	5	100.00	1035
r1	246	r1_r1_r1_r2_r2_r8_r9	301	160	53.15	231501
r2	247	r2_r1_r1_r1_r2_r4_r9	41	40	97.56	21012
r9	248	r9_r1_r1_r2_r2_r6_r9	7	3	42.85	113200
r1	249	r1_r1_r1_r12_r6_r7_r8	207	203	98.06	29476
r6	250	r6_r1_r1_r4_r7_r8_r8	41	39	95.12	12669
r1	251	r1_r1_r1_r12_r2_r2_r8	50620	39557	78.14	1661263

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r12	252	r12_r1_r1_r2_r8_r8_r8	164508	164434	99.95	871220
r2	253	r2_r1_r12_r8_r8_r8_r8	13141	10641	80.97	164919
r8	254	r8_r1_r1_r12_r16_r2_r8	2896	2893	99.89	76312
r8	255	r8_r12_r2_r2_r8_r8_r8	12932	12842	99.30	247346
r1	256	r1_r1_r1_r12_r13_r2_r2	1570	1570	100.00	15117
r2	257	r2_r1_r1_r1_r1_r13_r8	8209	8104	98.72	37227
r13	258	r13_r1_r1_r2_r2_r8_r8	455	85	18.68	607919
r1	259	r1_r1_r1_r12_r12_r2_r2	38562	38515	99.87	71819
r2	260	r2_r1_r1_r12_r12_r8_r8	26811	21858	81.52	278939
r12	261	r12_r1_r1_r2_r8_r8_r8	53574	53568	99.98	871220
r8	262	r8_r12_r2_r7_r8_r8_r8	852	852	100.00	35244
r8	263	r8_r1_r12_r2_r2_r8_r8	37862	37119	98.03	611159
r1	264	r1_r1_r1_r12_r2_r2_r8	321734	300640	93.44	1661263
r12	265	r12_r1_r1_r12_r2_r8_r8	22615	22606	99.96	2027886
r2	266	r2_r1_r1_r12_r13_r8_r8	4851	4078	84.06	36448
r8	267	r8_r1_r1_r12_r2_r8_r8	64077	49666	77.50	2027886
r8	268	r8_r12_r2_r8_r8_r8_r8	20598	20595	99.98	291541
r1	269	r1_r1_r1_r12_r2_r2_r2	61480	61391	99.85	364903
r2	270	r2_r1_r1_r15_r2_r2_r2	214	183	85.51	29214
r2	271	r2_r1_r1_r15_r2_r2_r5	1535	1535	100.00	19351
r2	272	r2_r1_r2_r2_r2_r5_r5	1200	1194	99.50	10342
r5	273	r5_r2_r2_r2_r2_r5_r9	32	32	100.00	4268
r1	274	r1_r1_r1_r12_r2_r7_r8	9218	9214	99.95	524285
r2	275	r2_r1_r1_r12_r8_r8_r8	350061	325859	93.08	1304116
r12	276	r12_r1_r1_r1_r2_r8_r8	198085	197913	99.91	866629
r8	277	r8_r1_r12_r2_r8_r8_r8	56787	56738	99.91	836752
r8	278	r8_r12_r12_r2_r8_r8_r8	32049	31897	99.52	148587
r1	279	r1_r1_r1_r2_r7_r7_r8	132	108	81.81	95144
r7	280	r7_r1_r1_r1_r2_r8_r8	2102	934	44.43	866629
r1	281	r1_r1_r1_r11_r2_r7_r8	41	38	92.68	41738
r7	282	r7_r1_r1_r11_r7_r8_r8	26	9	34.61	23026
r11	283	r11_r1_r1_r1_r2_r7_r7	15	5	33.33	58983

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	284	r1_r1_r11_r7_r8_r8_r9	17	17	100.00	1441
r8	285	r8_r1_r1_r8_r8_r8_r8	667	337	50.52	159241
r1	286	r1_r1_r1_r12_r2_r7_r8	7191	7184	99.90	524285
r7	287	r7_r1_r1_r1_r8_r9_r9	44	19	43.18	6408
r8	288	r8_r1_r1_r2_r2_r7_r9	64	30	46.87	116008
r1	289	r1_r1_r1_r12_r2_r2_r9	10757	10738	99.82	85760
r2	290	r2_r1_r1_r1_r12_r8_r8	262799	254815	96.96	887582
r12	291	r12_r1_r1_r1_r2_r8_r8	55959	55869	99.83	866629
r8	292	r8_r1_r1_r12_r2_r8_r8	6228	6223	99.91	2027886
r8	293	r8_r1_r12_r2_r8_r8_r8	38878	38827	99.86	836752
r1	294	r1_r1_r1_r1_r2_r7_r9	318	124	38.99	119735
r2	295	r2_r1_r1_r1_r1_r7_r9	146	120	82.19	27680
r9	296	r9_r1_r1_r12_r2_r2_r8	2046	228	11.14	1661263
r1	297	r1_r1_r1_r2_r2_r7_r9	497	401	80.68	116008
r7	298	r7_r1_r1_r2_r2_r2_r9	309	29	9.38	337570
r1	299	r1_r1_r1_r12_r2_r2_r8	57248	53294	93.09	1661263
r2	300	r2_r1_r1_r2_r3_r8_r9	171	168	98.24	23037
r2	301	r2_r1_r1_r2_r3_r7_r9	68	61	89.70	10623
r3	302	r3_r1_r2_r2_r2_r9_r9	7	2	28.57	9230
r1	303	r1_r1_r1_r12_r8_r8_r8	121608	74466	61.23	1304116
r12	304	r12_r1_r1_r8_r8_r8_r9	1293	1246	96.36	18547
r8	305	r8_r1_r12_r8_r8_r8_r8	13470	13428	99.68	164919
r8	306	r8_r1_r1_r1_r12_r8_r9	136	135	99.26	33018
r1	307	r1_r1_r1_r2_r7_r8_r8	1221	1195	97.87	333145
r8	308	r8_r1_r1_r1_r1_r11_r7	1	1	100.00	11398
r1	309	r1_r1_r1_r12_r2_r7_r8	3139	3139	100.00	524285
r7	310	r7_r1_r1_r1_r1_r1_r8	2133	1810	84.85	113500
r8	311	r8_r1_r1_r1_r1_r12_r7	10	8	80.00	8589
r1	312	r1_r1_r1_r12_r7_r8_r8	25714	24458	95.11	401410
r12	313	r12_r1_r1_r16_r2_r8_r8	222	197	88.73	73310
r8	314	r8_r1_r1_r12_r2_r8_r8	2136	1560	73.03	2027886
r8	315	r8_r1_r1_r12_r8_r8_r8	26303	7023	26.70	1304116

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	316	r1_r1_r1_r2_r7_r8_r8	271	233	85.97	333145
r7	317	r7_r1_r1_r1_r2_r2_r2	4632	108	2.33	537457
r1	318	r1_r1_r11_r12_r2_r7_r8	1107	1089	98.37	23911
r2	319	r2_r1_r1_r11_r12_r8_r8	21489	21163	98.48	79472
r12	320	r12_r1_r1_r2_r8_r8_r8	42998	42997	99.99	871220
r8	321	r8_r1_r12_r2_r8_r8_r8	38389	38384	99.98	836752
r8	322	r8_r12_r2_r8_r8_r8_r8	40753	40750	99.99	291541
r11	323	r11_r1_r1_r2_r2_r8_r9	25	19	76.00	231501
r1	324	r1_r1_r11_r2_r2_r7_r9	11	11	100.00	6155
r2	325	r2_r1_r11_r2_r3_r7_r9	7	7	100.00	614
r2	326	r2_r1_r11_r2_r2_r3_r9	21	21	100.00	3019
r3	327	r3_r1_r2_r2_r2_r2_r9	444	176	39.63	50950
r1	328	r1_r1_r1_r10_r10_r7_r9	29	29	100.00	198
r7	329	r7_r1_r1_r1_r10_r2_r9	5	5	100.00	11326
r1	330	r1_r1_r1_r10_r7_r8_r9	33	33	100.00	1619
r10	331	r10_r1_r1_r1_r10_r2_r5	20	17	85.00	1357
r1	332	r1_r1_r1_r7_r8_r8_r8	212	129	60.84	116575
r8	333	r8_r1_r1_r2_r7_r8_r9	30	27	90.00	130533
r1	334	r1_r1_r1_r7_r8_r8_r8	1242	961	77.37	116575
r8	335	r8_r1_r1_r1_r7_r8_r8	127	29	22.83	184875
r1	336	r1_r1_r1_r1_r12_r8_r8	368	230	62.50	887582
r8	337	r8_r1_r1_r1_r1_r1_r7	191	95	49.73	44209
r1	338	r1_r1_r1_r1_r12_r7_r8	21	21	100.00	91753
r1	339	r1_r1_r1_r2_r7_r8_r8	6491	6459	99.50	333145
r7	340	r7_r1_r1_r1_r8_r8_r8	38645	4042	10.45	311591
r8	341	r8_r1_r1_r10_r7_r8_r9	8	8	100.00	1619
r1	342	r1_r1_r1_r10_r2_r6_r8	242	242	100.00	6975
r2	343	r2_r1_r1_r1_r1_r6_r8	269	246	91.44	137169
r6	344	r6_r1_r1_r1_r11_r2_r2	218	41	18.80	53861
r1	345	r1_r1_r1_r1_r10_r2_r8	406	405	99.75	28917
r10	346	r10_r1_r1_r1_r10_r8_r8	172	112	65.11	18733
r1	347	r1_r1_r1_r12_r12_r2_r8	34026	33941	99.75	272697

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	348	r2_r1_r1_r1_r12_r8_r8	319922	307601	96.14	887582
r12	349	r12_r1_r1_r2_r8_r8_r8	43714	43646	99.84	871220
r8	350	r8_r1_r1_r10_r12_r2_r8	2113	1194	56.50	59010
r8	351	r8_r1_r10_r12_r2_r8_r8	533	533	100.00	24725
r1	352	r1_r1_r1_r12_r2_r7_r8	52567	51431	97.83	524285
r12	353	r12_r1_r1_r2_r2_r8_r8	105838	100397	94.85	607919
r2	354	r2_r1_r1_r12_r8_r8_r8	204954	187651	91.55	1304116
r8	355	r8_r1_r12_r2_r8_r8_r8	38924	38921	99.99	836752
r8	356	r8_r1_r12_r2_r8_r8_r8	23111	23099	99.94	836752
r1	357	r1_r1_r1_r2_r7_r8_r8	706	546	77.33	333145
r7	358	r7_r1_r1_r2_r8_r8_r9	492	12	2.43	155735
r1	359	r1_r1_r1_r2_r2_r8_r8	5827	4375	75.08	607919
r1	360	r1_r1_r1_r1_r11_r2_r6	28	28	100.00	33428
r2	361	r2_r1_r1_r15_r2_r4_r8	8	8	100.00	1150
r15	362	r15_r12_r12_r2_r2_r8_r8	1065	320	30.04	59797
r1	363	r1_r1_r1_r1_r16_r2_r6	44	44	100.00	27066
r1	364	r1_r1_r1_r2_r7_r8_r9	4462	4421	99.08	130533
r2	365	r2_r1_r1_r1_r10_r16_r9	15	15	100.00	378
r9	366	r9_r1_r1_r1_r1_r2_r2	12953	429	3.31	380927
r1	367	r1_r1_r1_r2_r2_r7_r8	22532	22291	98.93	402973
r7	368	r7_r1_r1_r1_r2_r7_r8	598	438	73.24	315627
r8	369	r8_r1_r1_r1_r2_r7_r9	78	59	75.64	119735
r1	370	r1_r1_r1_r2_r2_r7_r8	6872	6760	98.37	402973
r2	371	r2_r1_r1_r2_r2_r5_r9	1569	1569	100.00	97129
r2	372	r2_r1_r1_r2_r2_r5_r7	2399	2396	99.87	125071
r2	373	r2_r1_r2_r2_r5_r7_r7	72	69	95.83	2882
r5	374	r5_r1_r2_r2_r2_r7_r7	102	95	93.13	3950
r1	375	r1_r1_r1_r1_r2_r7_r7	448	357	79.68	58983
r7	376	r7_r1_r1_r1_r8_r8_r8	24661	2202	8.92	311591
r1	377	r1_r1_r1_r1_r2_r2_r7	2340	1545	66.02	250211
r7	378	r7_r1_r1_r1_r1_r2_r7	306	78	25.49	157473
r1	379	r1_r1_r1_r1_r2_r7_r9	8	6	75.00	119735

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	380	r1_r1_r1_r2_r7_r7_r9	42	34	80.95	19158
r7	381	r7_r1_r1_r1_r16_r6_r9	57	0	0	2498
r1	382	r1_r1_r1_r16_r2_r7_r7	226	218	96.46	4555
r7	383	r7_r1_r1_r1_r2_r2_r7	212	18	8.49	250211
r1	384	r1_r1_r11_r16_r2_r7_r8	18	18	100.00	2366
r2	385	r2_r1_r1_r1_r11_r16_r8	5	5	100.00	5952
r16	386	r16_r1_r12_r2_r2_r8_r9	7	1	14.28	81133
r11	387	r11_r1_r1_r1_r1_r2_r7	6	3	50.00	157473
r1	388	r1_r1_r11_r12_r2_r7_r8	605	553	91.40	23911
r7	389	r7_r1_r11_r2_r6_r8_r9	0	0		3561
r1	390	r1_r1_r1_r12_r2_r2_r8	263239	257706	97.89	1661263
r2	391	r2_r1_r1_r12_r8_r8_r8	358164	329710	92.05	1304116
r12	392	r12_r1_r1_r2_r8_r8_r8	92444	92425	99.97	871220
r8	393	r8_r1_r12_r2_r7_r8_r8	2153	2082	96.70	126449
r8	394	r8_r1_r1_r12_r2_r2_r8	38440	31946	83.10	1661263
r1	395	r1_r1_r1_r12_r2_r2_r8	446256	429828	96.31	1661263
r2	396	r2_r1_r1_r12_r15_r8_r8	7883	7782	98.71	32449
r12	397	r12_r1_r1_r10_r2_r8_r8	677	670	98.96	30972
r8	398	r8_r1_r1_r12_r15_r2_r8	1108	1092	98.55	26523
r8	399	r8_r12_r12_r2_r2_r2_r8	546	536	98.16	3105
r1	400	r1_r1_r1_r10_r2_r2_r8	772	769	99.61	29042
r2	401	r2_r1_r1_r12_r2_r8_r9	805	790	98.13	184439
r2	402	r2_r1_r1_r12_r2_r8_r9	94	94	100.00	184439
r9	403	r9_r1_r1_r1_r2_r2_r7	1904	554	29.09	250211
r1	404	r1_r1_r1_r10_r16_r2_r8	42	42	100.00	2662
r10	405	r10_r1_r1_r10_r12_r16_r8	15	15	100.00	977
r1	406	r1_r1_r1_r12_r2_r8_r8	15841	12525	79.06	2027886
r2	407	r2_r1_r1_r12_r8_r8_r8	310435	265631	85.56	1304116
r12	408	r12_r1_r1_r2_r8_r8_r8	24076	24076	100.00	871220
r8	409	r8_r1_r12_r2_r8_r8_r8	18257	18205	99.71	836752
r8	410	r8_r1_r12_r2_r2_r2_r8	6246	6011	96.23	207466
r1	411	r1_r1_r1_r2_r2_r8_r8	13944	9068	65.03	607919

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	412	r2_r1_r1_r1_r2_r2_r5	7499	7436	99.15	606334
r2	413	r2_r1_r1_r2_r2_r5_r8	3600	3584	99.55	348399
r2	414	r2_r1_r1_r16_r2_r2_r5	90	85	94.44	66544
r5	415	r5_r1_r1_r2_r2_r2_r8	3362	2311	68.73	560126
r1	416	r1_r1_r11_r6_r8_r8_r8	1	1	100.00	2067
r8	417	r8_r1_r1_r11_r8_r8_r8	446	36	8.07	25136
r11	418	r11_r1_r1_r1_r12_r8_r8	376	42	11.17	887582
r1	419	r1_r1_r12_r12_r2_r8_r8	2565	2536	98.86	219140
r12	420	r12_r1_r1_r1_r8_r8_r8	1130	1002	88.67	311591
r8	421	r8_r1_r1_r1_r12_r8_r8	1848	1578	85.38	887582
r8	422	r8_r1_r10_r12_r8_r8_r8	97	96	98.96	5360
r1	423	r1_r1_r1_r12_r2_r7_r8	6543	6540	99.95	524285
r2	424	r2_r1_r1_r1_r12_r8_r8	418521	392084	93.68	887582
r12	425	r12_r1_r1_r1_r2_r8_r8	192237	192154	99.95	866629
r8	426	r8_r1_r12_r12_r2_r8_r8	3870	3315	85.65	219140
r8	427	r8_r1_r1_r1_r12_r2_r8	8422	8382	99.52	533479
r1	428	r1_r1_r1_r1_r2_r7_r7	685	468	68.32	58983
r7	429	r7_r1_r1_r1_r1_r7_r7	1742	607	34.84	37641
r1	430	r1_r1_r1_r1_r2_r6_r7	300	284	94.66	136980
r6	431	r6_r1_r1_r2_r2_r2_r5	5216	4157	79.69	300752
r1	432	r1_r1_r1_r1_r2_r7_r7	1881	1759	93.51	58983
r7	433	r7_r1_r1_r1_r1_r7_r7	3048	287	9.41	37641
r1	434	r1_r1_r1_r1_r10_r7_r7	82	79	96.34	3064
r7	435	r7_r1_r1_r1_r2_r7_r8	162	74	45.67	315627
r1	436	r1_r1_r1_r10_r2_r2_r2	513	494	96.29	25135
r10	437	r10_r1_r1_r10_r2_r8_r8	25	24	96.00	30972
r1	438	r1_r1_r1_r1_r2_r2_r7	3940	3705	94.03	250211
r2	439	r2_r1_r1_r1_r2_r2_r5	228763	228695	99.97	606334
r2	440	r2_r1_r1_r12_r2_r2_r5	3431	3426	99.85	22822
r2	441	r2_r1_r1_r2_r2_r5_r6	8486	8471	99.82	234219
r5	442	r5_r1_r2_r2_r2_r6_r8	369	368	99.72	64140
r1	443	r1_r1_r1_r1_r1_r7_r7	4475	2480	55.41	37641

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r7	444	r7_r1_r1_r1_r1_r1_r7	4690	1116	23.79	44209
r1	445	r1_r1_r1_r1_r16_r2_r7	562	531	94.48	29052
r7	446	r7_r1_r1_r1_r1_r1_r8	1807	556	30.76	113500
r1	447	r1_r1_r1_r11_r2_r2_r2	212	196	92.45	57457
r2	448	r2_r1_r1_r16_r2_r2_r8	438	406	92.69	56395
r16	449	r16_r1_r1_r2_r2_r2_r8	1022	157	15.36	560126
r11	450	r11_r1_r1_r1_r1_r2_r8	5	4	80.00	343173
r1	451	r1_r1_r1_r11_r2_r2_r9	332	307	92.46	39171
r1	452	r1_r1_r1_r12_r2_r2_r2	83182	83013	99.79	364903
r2	453	r2_r1_r1_r1_r2_r2_r9	13802	9262	67.10	286355
r2	454	r2_r1_r1_r2_r2_r8_r9	559	544	97.31	231501
r9	455	r9_r1_r1_r2_r2_r7_r8	64	56	87.50	402973
r1	456	r1_r1_r1_r12_r2_r8_r8	85870	69281	80.68	2027886
r12	457	r12_r1_r1_r2_r2_r8_r8	83300	81139	97.40	607919
r2	458	r2_r1_r1_r12_r2_r8_r8	56955	47435	83.28	2027886
r8	459	r8_r1_r1_r12_r2_r2_r8	18719	13370	71.42	1661263
r8	460	r8_r1_r12_r2_r2_r8_r8	11834	11421	96.51	611159
r1	461	r1_r1_r1_r1_r7_r8_r8	2002	1066	53.24	184875
r7	462	r7_r1_r1_r1_r13_r6_r8	6	6	100.00	4934
r8	463	r8_r1_r1_r1_r13_r6_r7	16	16	100.00	2540
r1	464	r1_r1_r1_r2_r2_r7_r9	109	72	66.05	116008
r1	465	r1_r1_r1_r2_r2_r2_r9	24146	23943	99.15	337570
r2	466	r2_r1_r1_r6_r7_r9_r9	2	2	100.00	893
r9	467	r9_r1_r1_r2_r6_r7_r9	4	3	75.00	32663
r1	468	r1_r1_r1_r12_r2_r2_r8	116801	96513	82.63	1661263
r2	469	r2_r1_r1_r1_r13_r8_r8	3817	3718	97.40	31195
r13	470	r13_r1_r1_r2_r7_r8_r8	231	46	19.91	333145
r1	471	r1_r1_r1_r10_r12_r2_r8	2074	2018	97.29	59010
r12	472	r12_r1_r1_r1_r2_r8_r8	31050	30964	99.72	866629
r2	473	r2_r1_r1_r12_r2_r8_r8	102775	98786	96.11	2027886
r8	474	r8_r1_r12_r2_r2_r2_r8	15095	12979	85.98	207466
r8	475	r8_r1_r12_r2_r2_r8_r8	9638	8649	89.73	611159

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	476	r1_r1_r1_r10_r12_r2_r8	311	303	97.42	59010
r10	477	r10_r1_r1_r1_r10_r11_r8	1	1	100.00	2557
r1	478	r1_r1_r1_r12_r12_r8_r8	52907	46153	87.23	278939
r12	479	r12_r1_r1_r16_r8_r8_r8	906	887	97.90	24021
r8	480	r8_r1_r1_r12_r8_r8_r8	37752	6820	18.06	1304116
r8	481	r8_r1_r1_r12_r2_r8_r8	124724	12540	10.05	2027886
r1	482	r1_r1_r1_r14_r2_r8_r8	162	144	88.88	89429
r1	483	r1_r1_r1_r12_r2_r2_r6	951	943	99.15	83964
r2	484	r2_r1_r1_r1_r1_r14_r6	71	66	92.95	12438
r14	485	r14_r1_r2_r2_r2_r5_r8	180	36	20.00	69145
r1	486	r1_r1_r1_r12_r2_r2_r6	1451	1430	98.55	83964
r2	487	r2_r1_r1_r12_r6_r8_r8	9140	8627	94.38	165713
r12	488	r12_r1_r1_r2_r8_r8_r8	170466	170437	99.98	871220
r8	489	r8_r1_r12_r2_r7_r8_r8	1130	1130	100.00	126449
r8	490	r8_r1_r12_r2_r7_r8_r8	2596	2592	99.84	126449
r1	491	r1_r1_r1_r1_r2_r2_r6	247	232	93.92	312636
r2	492	r2_r1_r1_r1_r1_r2_r6	1639	1469	89.62	190786
r6	493	r6_r1_r1_r1_r2_r2_r2	48532	3408	7.02	537457
r1	494	r1_r1_r1_r1_r2_r2_r8	33	20	60.60	484108
r1	495	r1_r1_r1_r2_r2_r8_r8	11270	8944	79.36	607919
r1	496	r1_r1_r1_r2_r2_r6_r8	801	730	91.13	220713
r2	497	r2_r1_r1_r2_r2_r5_r8	50511	50502	99.98	348399
r2	498	r2_r1_r1_r2_r2_r5_r8	35810	35791	99.94	348399
r2	499	r2_r1_r14_r2_r2_r5_r8	37	37	100.00	9191
r5	500	r5_r1_r14_r2_r2_r2_r8	57	57	100.00	13260
r1	501	r1_r1_r1_r1_r12_r6_r7	5	5	100.00	7171
r6	502	r6_r1_r1_r1_r1_r5_r7	5	5	100.00	4052
r1	503	r1_r1_r1_r10_r2_r5_r7	13	13	100.00	577
r7	504	r7_r1_r1_r2_r5_r8_r8	34	15	44.11	20044
r1	505	r1_r1_r1_r10_r2_r2_r2	634	611	96.37	25135
r10	506	r10_r1_r1_r1_r10_r2_r8	129	116	89.92	28917
r1	507	r1_r1_r1_r1_r1_r2_r2	1647	1018	61.80	380927

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	508	r2_r1_r1_r1_r2_r6_r9	236	235	99.57	103659
r2	509	r2_r1_r1_r2_r8_r9_r9	127	80	62.99	24677
r9	510	r9_r1_r10_r2_r2_r2_r8	4	4	100.00	2752
r1	511	r1_r1_r1_r1_r1_r2_r6	4	2	50.00	190786
r1	512	r1_r1_r1_r12_r6_r7_r8	165	163	98.78	29476
r6	513	r6_r1_r1_r1_r1_r1_r2	461	13	2.81	199396
r1	514	r1_r1_r1_r10_r12_r7_r8	52	52	100.00	5763
r7	515	r7_r1_r1_r1_r2_r4_r8	69	4	5.79	36545
r1	516	r1_r1_r1_r1_r1_r7_r8	146	89	60.95	147919
r1	517	r1_r1_r1_r11_r12_r8_r8	335	117	34.92	79472
r11	518	r11_r1_r1_r7_r8_r8_r8	189	8	4.23	116575
r1	519	r1_r1_r12_r12_r8_r8_r8	12253	4569	37.28	99927
r12	520	r12_r1_r1_r1_r2_r8_r8	2391	2183	91.30	866629
r8	521	r8_r1_r10_r12_r2_r8_r8	53	53	100.00	24725
r8	522	r8_r1_r1_r12_r16_r2_r8	261	260	99.61	76312
r1	523	r1_r1_r1_r10_r12_r8_r8	8244	6984	84.71	47724
r12	524	r12_r1_r1_r1_r8_r8_r8	67323	63834	94.81	311591
r8	525	r8_r1_r1_r1_r12_r8_r8	338111	10723	3.17	887582
r8	526	r8_r1_r1_r12_r16_r8_r8	2333	1655	70.93	79460
r1	527	r1_r1_r1_r1_r10_r2_r9	51	49	96.07	11326
r10	528	r10_r1_r1_r10_r2_r2_r8	622	38	6.10	29042
r1	529	r1_r1_r1_r2_r2_r2_r9	15299	14654	95.78	337570
r2	530	r2_r1_r1_r1_r2_r9_r9	524	494	94.27	43555
r9	531	r9_r1_r1_r10_r2_r7_r8	9	8	88.88	16508
r1	532	r1_r1_r1_r2_r2_r8_r8	6405	4849	75.70	607919
r1	533	r1_r1_r1_r2_r2_r2_r8	37379	36376	97.31	560126
r2	534	r2_r1_r1_r1_r2_r2_r9	3775	3339	88.45	286355
r2	535	r2_r1_r1_r1_r2_r7_r9	1210	1136	93.88	119735
r9	536	r9_r11_r2_r2_r2_r2_r3	19	14	73.68	317
r1	537	r1_r1_r1_r12_r2_r2_r8	396333	378998	95.62	1661263
r2	538	r2_r1_r1_r12_r8_r8_r8	254396	218393	85.84	1304116
r12	539	r12_r1_r2_r8_r8_r8_r8	31014	31014	100.00	165098

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r8	540	r8_r1_r12_r2_r7_r8_r8	2610	2604	99.77	126449
r8	541	r8_r12_r12_r2_r7_r8_r8	122	122	100.00	6445
r1	542	r1_r1_r1_r12_r2_r2_r8	1379	1027	74.47	1661263
r2	543	r2_r1_r1_r10_r2_r8_r9	121	121	100.00	8476
r2	544	r2_r1_r1_r10_r2_r8_r9	105	105	100.00	8476
r9	545	r9_r1_r1_r1_r2_r2_r8	799	180	22.52	484108
r1	546	r1_r1_r1_r12_r12_r2_r8	169	158	93.49	272697
r1	547	r1_r1_r1_r12_r12_r2_r8	61994	61278	98.84	272697
r12	548	r12_r1_r1_r2_r7_r8_r8	19546	19517	99.85	333145
r2	549	r2_r1_r1_r12_r8_r8_r8	274629	218781	79.66	1304116
r8	550	r8_r1_r1_r12_r2_r8_r8	63960	43905	68.64	2027886
r8	551	r8_r12_r2_r2_r8_r8_r8	24829	23473	94.53	247346
r1	552	r1_r1_r1_r12_r7_r8_r8	58377	55243	94.63	401410
r12	553	r12_r1_r1_r1_r1_r8_r8	50961	42097	82.60	280823
r8	554	r8_r1_r1_r12_r2_r7_r8	91	89	97.80	524285
r8	555	r8_r1_r1_r1_r12_r8_r8	5928	4303	72.58	887582
r1	556	r1_r1_r1_r2_r2_r2_r7	1754	1493	85.11	225532
r7	557	r7_r1_r1_r1_r2_r6_r7	50	40	80.00	136980
r1	558	r1_r1_r1_r12_r15_r2_r5	1	1	100.00	76
r2	559	r2_r1_r1_r1_r15_r7_r8	46	46	100.00	6612
r15	560	r15_r1_r12_r2_r2_r8_r8	1964	193	9.82	611159
r1	561	r1_r1_r1_r1_r2_r2_r8	125	74	59.20	484108
r1	562	r1_r1_r1_r16_r2_r2_r2	357	346	96.91	87190
r2	563	r2_r1_r1_r16_r2_r2_r8	450	178	39.55	56395
r16	564	r16_r1_r13_r2_r2_r2_r2	2	0	0	5881
r1	565	r1_r1_r1_r1_r10_r2_r5	2	2	100.00	1357
r1	566	r1_r1_r1_r10_r2_r6_r8	31	31	100.00	6975
r10	567	r10_r1_r1_r1_r1_r10_r8	41	37	90.24	15906
r1	568	r1_r1_r1_r1_r2_r2_r8	878	594	67.65	484108
r8	569	r8_r1_r1_r1_r1_r7_r8	824	201	24.39	147919
r1	570	r1_r1_r1_r2_r2_r2_r2	38590	32208	83.46	597149
r2	571	r2_r1_r1_r1_r2_r2_r9	2535	2335	92.11	286355

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	572	r2_r1_r1_r1_r1_r2_r9	2994	2775	92.68	267543
r9	573	r9_r1_r1_r1_r2_r2_r7	1707	274	16.05	250211
r1	574	r1_r1_r1_r2_r2_r2_r9	27490	24982	90.87	337570
r2	575	r2_r1_r1_r1_r2_r2_r5	88289	87994	99.66	606334
r2	576	r2_r1_r1_r2_r2_r2_r5	27751	27709	99.84	300752
r2	577	r2_r1_r2_r2_r2_r5_r9	1623	1620	99.81	34576
r5	578	r5_r1_r1_r2_r2_r2_r9	439	340	77.44	337570
r1	579	r1_r1_r1_r2_r2_r2_r2	25154	18822	74.82	597149
r2	580	r2_r1_r1_r2_r2_r2_r9	1592	580	36.43	337570
r9	581	r9_r1_r1_r2_r2_r2_r2	10284	3002	29.19	597149
r1	582	r1_r1_r1_r11_r2_r2_r2	553	456	82.45	57457
r2	583	r2_r1_r1_r2_r2_r2_r5	7653	7628	99.67	300752
r2	584	r2_r1_r1_r2_r2_r2_r5	6580	6541	99.40	300752
r2	585	r2_r1_r1_r2_r2_r2_r5	26291	25787	98.08	300752
r5	586	r5_r1_r1_r13_r2_r2_r2	9	3	33.33	45691
r11	587	r11_r1_r1_r10_r16_r2_r8	21	21	100.00	2662
r1	588	r1_r1_r11_r12_r16_r8_r8	33	33	100.00	3457
r1	589	r1_r1_r1_r12_r2_r8_r8	218172	195915	89.79	2027886
r12	590	r12_r1_r1_r1_r15_r8_r8	1252	1203	96.08	19710
r8	591	r8_r1_r12_r15_r2_r8_r8	145	137	94.48	19891
r8	592	r8_r1_r1_r12_r15_r8_r8	637	538	84.45	32449
r1	593	r1_r1_r1_r1_r2_r2_r7	3112	2989	96.04	250211
r2	594	r2_r1_r1_r1_r16_r2_r8	305	291	95.40	69525
r16	595	r16_r1_r1_r12_r2_r8_r8	238	36	15.12	2027886
r1	596	r1_r1_r1_r2_r2_r7_r8	5559	4528	81.45	402973
r7	597	r7_r1_r1_r1_r12_r8_r8	264	134	50.75	887582
r8	598	r8_r1_r1_r1_r12_r7_r8	36	36	100.00	91753
r1	599	r1_r1_r1_r12_r2_r2_r8	94182	89449	94.97	1661263
r2	600	r2_r1_r1_r1_r2_r2_r8	6516	4281	65.69	484108
r2	601	r2_r1_r2_r2_r5_r8_r9	677	675	99.70	16887
r2	602	r2_r1_r2_r2_r5_r8_r9	752	750	99.73	16887
r5	603	r5_r10_r2_r2_r2_r8_r9	9	9	100.00	187

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	604	r1_r1_r1_r10_r12_r8_r8	3994	3706	92.78	47724
r12	605	r12_r1_r1_r1_r1_r8_r8	60409	52245	86.48	280823
r8	606	r8_r1_r12_r12_r8_r8_r8	4218	4207	99.73	99927
r8	607	r8_r1_r1_r1_r12_r2_r8	3258	3167	97.20	533479
r1	608	r1_r1_r1_r10_r2_r8_r9	124	124	100.00	8476
r10	609	r10_r1_r1_r1_r1_r10_r2	24	23	95.83	18944
r1	610	r1_r1_r1_r16_r2_r2_r9	1739	1711	98.38	30844
r2	611	r2_r1_r1_r1_r1_r16_r9	858	840	97.90	8613
r9	612	r9_r1_r1_r2_r2_r2_r2	51496	2180	4.23	597149
r1	613	r1_r1_r1_r12_r2_r8_r8	15352	11541	75.17	2027886
r2	614	r2_r1_r1_r1_r16_r8_r8	3458	3122	90.28	53067
r16	615	r16_r1_r12_r2_r2_r8_r8	424	181	42.68	611159
r1	616	r1_r1_r1_r12_r2_r8_r8	237254	217503	91.67	2027886
r12	617	r12_r1_r1_r12_r8_r8_r9	49	48	97.95	83003
r8	618	r8_r1_r12_r2_r8_r8_r9	1142	1107	96.93	66033
r8	619	r8_r1_r1_r1_r11_r12_r8	449	448	99.77	21968
r1	620	r1_r1_r1_r1_r2_r2_r9	5399	4939	91.47	286355
r2	621	r2_r1_r1_r1_r1_r2_r9	475	453	95.36	267543
r9	622	r9_r1_r1_r12_r2_r8_r9	451	72	15.96	184439
r1	623	r1_r1_r1_r12_r15_r2_r7	149	149	100.00	1069
r2	624	r2_r1_r1_r1_r1_r15_r2	262	249	95.03	18907
r15	625	r15_r1_r2_r4_r7_r8_r8	27	27	100.00	1802
r1	626	r1_r1_r1_r12_r2_r8_r8	13645	11854	86.87	2027886
r12	627	r12_r1_r1_r1_r2_r8_r8	224738	224263	99.78	866629
r2	628	r2_r1_r1_r12_r8_r8_r8	63261	58738	92.85	1304116
r8	629	r8_r1_r12_r2_r8_r8_r8	18465	18443	99.88	836752
r8	630	r8_r1_r12_r2_r2_r8_r8	7157	6256	87.41	611159
r1	631	r1_r1_r1_r2_r2_r2_r2	109396	103355	94.47	597149
r2	632	r2_r1_r1_r1_r1_r2_r9	109023	108201	99.24	267543
r2	633	r2_r1_r1_r1_r2_r8_r9	10253	9979	97.32	247808
r9	634	r9_r1_r1_r12_r2_r2_r8	28620	1722	6.01	1661263
r1	635	r1_r1_r1_r2_r2_r7_r8	6522	6477	99.31	402973

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r2	636	r2_r1_r1_r1_r1_r2_r9	105658	104533	98.93	267543
r2	637	r2_r1_r1_r1_r1_r2_r9	91760	91339	99.54	267543
r9	638	r9_r1_r1_r1_r2_r2_r8	846	452	53.42	484108
r1	639	r1_r1_r1_r12_r7_r8_r8	19667	19321	98.24	401410
r7	640	r7_r1_r1_r1_r15_r8_r8	212	89	41.98	19710
r8	641	r8_r1_r7_r8_r8_r8_r8	149	149	100.00	6495
r1	642	r1_r1_r1_r12_r2_r8_r8	82382	70752	85.88	2027886
r12	643	r12_r1_r1_r15_r2_r8_r8	1187	1184	99.74	26796
r2	644	r2_r1_r1_r12_r15_r8_r8	3122	2955	94.65	32449
r8	645	r8_r1_r1_r12_r15_r2_r8	200	200	100.00	26523
r8	646	r8_r1_r1_r12_r2_r8_r8	2191	409	18.66	2027886
r1	647	r1_r1_r1_r1_r1_r8_r8	74158	5804	7.82	280823
r8	648	r8_r1_r1_r1_r1_r1_r8	15634	9635	61.62	113500
r1	649	r1_r1_r1_r1_r2_r8_r9	3850	144	3.74	247808
r8	650	r8_r1_r1_r1_r1_r8_r9	176	64	36.36	61837
r1	651	r1_r1_r1_r12_r2_r2_r2	1010	947	93.76	364903
r2	652	r2_r1_r1_r2_r5_r8_r9	21	19	90.47	8328
r9	653	r9_r1_r2_r2_r5_r8_r8	2	1	50.00	68840
r1	654	r1_r11_r12_r2_r5_r8_r8	0	0		110
r12	655	r12_r1_r11_r2_r5_r8_r8	0	0		999
r8	656	r8_r1_r1_r11_r12_r5_r8	0	0		728
r8	657	r8_r1_r11_r12_r2_r5_r8	0	0		2908
r11	658	r11_r1_r1_r12_r2_r2_r8	821	523	63.70	1661263
r1	659	r1_r1_r10_r11_r13_r2_r6	0	0		52
r2	660	r2_r1_r11_r13_r2_r5_r6	0	0		75
r13	661	r13_r1_r10_r2_r2_r6_r8	0	0		866
r1	662	r1_r1_r1_r1_r10_r2_r6	31	23	74.19	12433
r6	663	r6_r1_r1_r1_r1_r2_r7	1309	748	57.14	157473
r1	664	r1_r1_r1_r1_r10_r6_r7	5	5	100.00	3638
r10	665	r10_r1_r1_r1_r1_r10_r13	1	0	0	1694
r1	666	r1_r1_r1_r1_r2_r6_r7	128	116	90.62	136980
r7	667	r7_r1_r1_r1_r1_r2_r6	2150	603	28.04	190786

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	668	r1_r1_r1_r16_r2_r2_r2	9471	9222	97.37	87190
r2	669	r2_r1_r1_r1_r16_r2_r3	1784	1782	99.88	30692
r2	670	r2_r1_r1_r16_r2_r3_r9	15	15	100.00	3274
r3	671	r3_r1_r1_r16_r2_r2_r9	2	0	0	30844
r1	672	r1_r1_r1_r1_r11_r2_r8	72	70	97.22	47557
r2	673	r2_r1_r1_r10_r11_r16_r8	0	0		305
r16	674	r16_r1_r1_r10_r2_r2_r2	11	2	18.18	25135
r11	675	r11_r1_r1_r12_r2_r2_r8	2945	186	6.31	1661263
r1	676	r1_r1_r1_r2_r7_r8_r9	529	481	90.92	130533
r2	677	r2_r1_r1_r2_r8_r8_r9	83	70	84.33	155735
r9	678	r9_r1_r1_r11_r2_r2_r8	10	1	10.00	53272
r1	679	r1_r1_r1_r1_r7_r7_r8	1533	1474	96.15	89403
r7	680	r7_r1_r1_r12_r8_r8_r9	82	62	75.60	83003
r8	681	r8_r1_r1_r1_r11_r13_r7	0	0		969
r1	682	r1_r1_r1_r7_r7_r8_r8	3468	3228	93.07	90603
r7	683	r7_r1_r1_r1_r12_r8_r8	10297	2463	23.91	887582
r8	684	r8_r1_r1_r1_r1_r1_r7	2090	1920	91.86	44209
r1	685	r1_r1_r1_r1_r2_r7_r8	1208	1200	99.33	315627
r7	686	r7_r1_r1_r1_r2_r2_r2	1977	101	5.10	537457
r1	687	r1_r1_r1_r2_r2_r6_r7	2010	1994	99.20	108233
r2	688	r2_r1_r1_r1_r2_r6_r7	1806	1278	70.76	136980
r6	689	r6_r1_r1_r1_r2_r2_r7	11049	2806	25.39	250211
r1	690	r1_r1_r1_r1_r2_r6_r7	990	857	86.56	136980
r2	691	r2_r1_r1_r1_r2_r2_r6	926	465	50.21	312636
r6	692	r6_r1_r2_r2_r2_r5_r7	216	177	81.94	13414
r1	693	r1_r1_r1_r2_r2_r8_r8	2556	2127	83.21	607919
r1	694	r1_r1_r1_r12_r2_r2_r8	94117	90200	95.83	1661263
r2	695	r2_r1_r1_r2_r2_r5_r7	26921	26899	99.91	125071
r2	696	r2_r1_r1_r2_r2_r5_r6	22303	22280	99.89	234219
r2	697	r2_r1_r2_r2_r5_r6_r7	589	589	100.00	12903
r5	698	r5_r2_r2_r2_r2_r6_r7	227	227	100.00	3821
r1	699	r1_r1_r1_r12_r2_r8_r8	214907	191263	88.99	2027886

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r12	700	r12_r1_r1_r12_r8_r8_r8	19633	19187	97.72	1304116
r8	701	r8_r1_r12_r2_r8_r8_r8	1988	1957	98.44	836752
r8	702	r8_r1_r1_r12_r7_r8_r8	613	392	63.94	401410
r1	703	r1_r1_r1_r12_r2_r2_r8	150171	140234	93.38	1661263
r2	704	r2_r1_r1_r1_r13_r16_r8	170	170	100.00	5086
r13	705	r13_r1_r1_r16_r2_r5_r7	1	0	0	1547
r1	706	r1_r1_r1_r12_r12_r2_r8	44777	44638	99.68	272697
r12	707	r12_r1_r1_r2_r8_r8_r8	188662	186790	99.00	871220
r2	708	r2_r1_r1_r12_r2_r8_r8	68789	56791	82.55	2027886
r8	709	r8_r1_r10_r12_r2_r2_r8	82	81	98.78	9730
r8	710	r8_r1_r12_r12_r2_r2_r8	2954	2712	91.80	32565
r1	711	r1_r1_r1_r10_r12_r2_r8	570	570	100.00	59010
r2	712	r2_r1_r1_r11_r12_r8_r8	8458	5775	68.27	79472
r12	713	r12_r1_r1_r11_r2_r8_r8	3309	3306	99.90	70238
r8	714	r8_r12_r2_r2_r7_r8_r9	133	132	99.24	2363
r8	715	r8_r1_r1_r12_r2_r2_r8	25523	24627	96.48	1661263
r1	716	r1_r1_r1_r10_r2_r2_r9	30	19	63.33	11381
r10	717	r10_r1_r1_r1_r10_r8_r8	29	18	62.06	18733
r1	718	r1_r1_r1_r2_r2_r2_r9	24464	23903	97.70	337570
r2	719	r2_r1_r1_r2_r4_r8_r9	0	0		10475
r9	720	r9_r1_r1_r2_r4_r8_r9	3	0	0	10475
r1	721	r1_r1_r1_r2_r2_r7_r8	500	468	93.60	402973
r2	722	r2_r1_r1_r1_r2_r2_r7	3203	209	6.52	250211
r2	723	r2_r1_r1_r1_r2_r2_r5	78239	78194	99.94	606334
r2	724	r2_r1_r2_r2_r5_r7_r8	85	70	82.35	29600
r5	725	r5_r1_r1_r2_r2_r6_r7	57	7	12.28	108233
r1	726	r1_r1_r1_r1_r2_r2_r8	6820	5135	75.29	484108
r8	727	r8_r1_r1_r10_r7_r9_r9	0	0		55
r1	728	r1_r1_r1_r12_r2_r2_r8	243249	230165	94.62	1661263
r2	729	r2_r1_r1_r1_r12_r8_r8	414950	392511	94.59	887582
r12	730	r12_r1_r1_r16_r2_r8_r8	5839	5838	99.98	73310
r8	731	r8_r1_r1_r12_r16_r2_r8	119	118	99.15	76312

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r8	732	r8_r1_r1_r12_r16_r2_r8	3015	2993	99.27	76312
r1	733	r1_r1_r1_r11_r2_r2_r9	393	303	77.09	39171
r2	734	r2_r1_r1_r1_r11_r2_r9	24	24	100.00	33042
r9	735	r9_r1_r1_r2_r2_r5_r8	43	13	30.23	348399
r11	736	r11_r1_r1_r1_r1_r12_r8	692	13	1.87	143285
r1	737	r1_r1_r1_r1_r12_r8_r8	3589	1573	43.82	887582
r12	738	r12_r1_r1_r2_r2_r8_r8	15409	15131	98.19	607919
r8	739	r8_r1_r1_r12_r2_r2_r8	1268	1243	98.02	1661263
r8	740	r8_r1_r1_r1_r12_r2_r8	523	449	85.85	533479
r1	741	r1_r1_r1_r1_r10_r10_r2	5	2	40.00	5477
r10	742	r10_r1_r1_r1_r10_r2_r4	4	4	100.00	3547
r1	743	r1_r1_r1_r2_r7_r8_r8	654	614	93.88	333145
r1	744	r1_r1_r1_r6_r7_r8_r8	122	99	81.14	51495
r7	745	r7_r1_r1_r1_r1_r8_r8	58248	3246	5.57	280823
r8	746	r8_r1_r1_r1_r1_r7_r8	270	75	27.77	147919
r1	747	r1_r1_r1_r1_r12_r6_r8	467	263	56.31	73460
r6	748	r6_r1_r1_r12_r8_r8_r8	1319	13	.98	1304116
r1	749	r1_r1_r1_r10_r12_r6_r8	10	5	50.00	2640
r1	750	r1_r1_r1_r12_r7_r8_r8	59682	53017	88.83	401410
r12	751	r12_r1_r1_r16_r6_r8_r8	212	212	100.00	4358
r8	752	r8_r1_r1_r1_r12_r6_r8	1502	1385	92.21	73460
r8	753	r8_r1_r1_r12_r16_r6_r8	33	32	96.96	4509
r1	754	r1_r1_r1_r2_r7_r8_r8	13497	13323	98.71	333145
r7	755	r7_r1_r1_r1_r10_r2_r8	30	20	66.66	28917
r8	756	r8_r1_r1_r1_r1_r2_r7	154	76	49.35	157473
r1	757	r1_r1_r1_r2_r6_r8_r8	102	98	96.07	139365
r2	758	r2_r1_r1_r12_r16_r2_r8	10	10	100.00	76312
r16	759	r16_r12_r12_r2_r2_r8_r8	1106	347	31.37	59797
r1	760	r1_r1_r1_r1_r2_r6_r9	126	91	72.22	103659
r6	761	r6_r1_r1_r2_r2_r7_r8	7	1	14.28	402973
r1	762	r1_r1_r1_r2_r2_r8_r9	3430	3248	94.69	231501
r2	763	r2_r1_r1_r1_r1_r2_r9	249	225	90.36	267543

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r9	764	r9_r1_r1_r12_r2_r2_r8	2011	212	10.54	1661263
r1	765	r1_r1_r1_r1_r2_r2_r6	4131	3320	80.36	312636
r2	766	r2_r1_r1_r1_r6_r8_r8	578	29	5.01	91319
r6	767	r6_r1_r13_r2_r2_r2_r5	3	2	66.66	4012
r1	768	r1_r1_r1_r10_r10_r12_r8	229	191	83.40	1707
r10	769	r10_r1_r1_r10_r16_r2_r8	3	0	0	2662
r1	770	r1_r1_r1_r12_r2_r8_r8	226325	209741	92.67	2027886
r12	771	r12_r1_r1_r1_r8_r8_r9	1299	1132	87.14	43222
r8	772	r8_r1_r1_r12_r2_r6_r8	474	471	99.36	287980
r8	773	r8_r1_r12_r2_r8_r9_r9	14	13	92.85	4212
r1	774	r1_r1_r1_r10_r12_r2_r8	7076	6894	97.42	59010
r2	775	r2_r1_r1_r12_r2_r8_r8	59994	52202	87.01	2027886
r12	776	r12_r1_r16_r2_r2_r8_r8	335	335	100.00	11419
r8	777	r8_r1_r12_r13_r2_r8_r8	45	43	95.55	35895
r8	778	r8_r1_r12_r12_r2_r8_r9	54	54	100.00	1928
r1	779	r1_r1_r1_r1_r10_r2_r8	228	226	99.12	28917
r10	780	r10_r1_r1_r1_r1_r1_r10	271	210	77.49	6208
r1	781	r1_r1_r1_r1_r2_r2_r8	1685	1309	77.68	484108
r8	782	r8_r1_r1_r1_r1_r12_r8	5687	3291	57.86	143285
r1	783	r1_r1_r1_r2_r2_r2_r4	713	710	99.57	32987
r4	784	r4_r1_r16_r2_r2_r7_r8	0	0		3767
r1	785	r1_r1_r1_r1_r2_r2_r9	18328	15012	81.90	286355
r2	786	r2_r1_r1_r1_r1_r7_r9	77	73	94.80	27680
r9	787	r9_r1_r1_r1_r12_r2_r8	128	13	10.15	533479
r1	788	r1_r1_r1_r12_r2_r2_r8	437234	421426	96.38	1661263
r2	789	r2_r1_r1_r12_r8_r8_r8	91707	75575	82.40	1304116
r12	790	r12_r1_r2_r8_r8_r8_r8	6802	6802	100.00	165098
r8	791	r8_r1_r12_r12_r2_r8_r8	1027	1002	97.56	219140
r8	792	r8_r1_r1_r12_r2_r8_r8	8495	6649	78.26	2027886
r1	793	r1_r1_r1_r2_r2_r2_r9	7627	7001	91.79	337570
r2	794	r2_r1_r1_r1_r1_r2_r6	21838	21585	98.84	190786
r6	795	r6_r1_r1_r1_r2_r2_r8	11701	675	5.76	484108

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r1	796	r1_r1_r1_r2_r2_r8_r9	361	336	93.07	231501
r2	797	r2_r1_r1_r2_r7_r8_r9	747	715	95.71	130533
r2	798	r2_r1_r1_r1_r2_r4_r9	651	634	97.38	21012
r9	799	r9_r1_r1_r2_r2_r2_r4	312	126	40.38	32987
r1	800	r1_r1_r1_r1_r2_r7_r8	630	436	69.20	315627
r1	801	r1_r1_r1_r7_r7_r8_r9	171	161	94.15	10812
r7	802	r7_r1_r1_r2_r8_r8_r9	60	23	38.33	155735
r8	803	r8_r1_r1_r1_r2_r2_r7	2191	1440	65.72	250211
r1	804	r1_r1_r1_r12_r7_r8_r9	248	237	95.56	11711
r7	805	r7_r1_r1_r1_r2_r7_r9	85	13	15.29	119735
r1	806	r1_r1_r1_r12_r8_r8_r8	166131	10270	6.18	1304116
r12	807	r12_r1_r1_r8_r8_r8_r9	1726	1619	93.80	18547
r8	808	r8_r1_r12_r6_r8_r8_r8	29	26	89.65	9552
r8	809	r8_r1_r1_r12_r6_r7_r8	164	164	100.00	29476
r1	810	r1_r1_r1_r2_r8_r8_r8	338	198	58.57	871220
r1	811	r1_r1_r1_r2_r2_r2_r2	76035	69462	91.35	597149
r2	812	r2_r1_r1_r1_r2_r2_r5	12770	12711	99.53	606334
r2	813	r2_r1_r1_r1_r2_r2_r5	13633	13618	99.88	606334
r2	814	r2_r1_r12_r2_r2_r5_r8	681	441	64.75	60700
r5	815	r5_r2_r2_r2_r2_r6_r6	665	665	100.00	10604
r1	816	r1_r1_r1_r2_r2_r2_r4	622	593	95.33	32987
r2	817	r2_r1_r1_r1_r2_r2_r5	115875	115656	99.81	606334
r2	818	r2_r1_r1_r2_r2_r4_r5	4515	4514	99.97	26621
r2	819	r2_r1_r1_r10_r2_r2_r5	127	126	99.21	9651
r5	820	r5_r10_r10_r2_r2_r2_r4	0	0		5
r1	821	r1_r1_r1_r2_r2_r4_r6	38	36	94.73	16222
r4	822	r4_r1_r1_r1_r10_r2_r5	0	0		1357
r1	823	r1_r1_r1_r1_r12_r6_r8	780	218	27.94	73460
r6	824	r6_r1_r1_r2_r2_r5_r8	8088	143	1.76	348399
r1	825	r1_r1_r1_r1_r11_r2_r4	6	6	100.00	6386
r1	826	r1_r1_r10_r10_r2_r6_r7	0	0		13
r2	827	r2_r1_r1_r11_r6_r7_r8	0	0		14093

Continued on next page

Table 4.1 – continued from previous page

cg	cg_num	Star Composition	Total_preds	Correct_preds	Correct_preds(%)	Total Shells
r6	828	r6_r1_r1_r11_r2_r7_r8	4	3	75.00	41738
r1	829	r1_r1_r12_r2_r2_r8_r8	22165	527	2.37	611159
r10	830	r10_r1_r1_r1_r1_r10_r7	15	1	6.66	7988
r1	831	r1_r1_r2_r2_r7_r8_r9	0	0		12266
r7	832	r7_r1_r1_r2_r2_r8_r9	42	0	0	231501
r8	833	r8_r1_r1_r2_r2_r7_r9	54	1	1.85	116008
r1	834	r1_r1_r1_r2_r2_r6_r9	359	355	98.88	113200
r2	835	r2_r1_r1_r1_r2_r7_r9	10	7	70.00	119735
r9	836	r9_r1_r1_r1_r2_r7_r8	81	16	19.75	315627
r1	837	r1_r1_r1_r2_r2_r6_r6	1034	973	94.10	49314
r2	838	r2_r1_r1_r1_r2_r2_r6	862	462	53.59	312636
r6	839	r6_r1_r1_r2_r2_r2_r6	594	77	12.96	349950
r1	840	r1_r1_r1_r2_r2_r2_r6	594	102	17.17	349950
r2	841	r2_r1_r1_r1_r2_r2_r6	985	432	43.85	312636
r6	842	r6_r1_r1_r1_r2_r2_r6	1090	287	26.33	312636

cg: Chemical Group;

cg_num: Chemical Group Number in the protein gpdb;

Total_preds: Total number of predictions(i.e. superimpositions with RMSD < 1Å);

Correct_preds: Predictions where deleted chemical group is same as predicted one;

Total Shells: Total shells from the stars database that were used for performing the superimposition.