Role of Atomic Packing in Stabilizing Protein Structures

A thesis submitted towards partial fulfilment of
BS-MS Dual Degree Programme

by

Akash Bahai

under the guidance of

Dr. M. S. Madhusudhan

Associate Professor, IISER Pune

Indian Institute of Science Education and Research, Pune

Certificate

This is to certify that this dissertation entitled ''Role of Atomic Packing in Stabilizing Protein Structures" submitted towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research Pune represents original research carried out by ''Akash Bahai" at ''IISER Pune'', under the supervision of ''Dr. M.S. Madhusudhan, Associate Professor, Biology'' during the academic year.

Supervisor

Dr. M.S. Madhusudhan Date: 12/11/2015

Declaration

I hereby declare that the matter embodied in the report entitled " Role of Atomic Packing in Stabilizing Protein Structures" are the results of the investigations 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.

Student

Akash Bahai Date: 12/11/2015

Acknowledgements

I am very grateful to my thesis supervisor, Dr. M. S. Madhusudhan, Associate Professor at the Indian Institute of Science Education and Research, Pune for providing me with the opportunity to work on this project. This project would not have been possible without his invaluable guidance. I learned a lot from you and thanks for being a great mentor.

I would also like to express my sincere thanks to Dr. Chetan Gadgil, from National Chemical Laboratory, Pune for agreeing to be on my Thesis Advisory Committee.

I thank all my lab mates from the Madhusudhan Lab, Neelesh Soni, Abhinash Palai, Sagar Gore, Yogendra Ramtirtha, Parichit Sharma, Shipra Gupta, Neeladri Sen and Sanjana Nair, for the valuable discussions and support with various aspects of the project.

I would also like to thank my friends at IISER Pune, who made my stay at IISER enjoyable for five long years. Thanks for all the memories.

Finally, I would like to thank my parents and sister who have always supported me with all my endeavours. Thanks Mom, Dad and Abha.

Abstract

Proteins are the building blocks of life and vital to the essential functioning of our body. They play very crucial roles in a variety of cellular processes. The three dimensional structure of protein sequence is instrumental to figure out the functioning of that protein. In this study, we have a developed a distance-angle based pairwise statistical potential based on chemical groups for the assessment of computationally determined protein structures. We took previously solved structures from the PDB (Protein Data bank) for the construction of the statistical potential. Depth analysis of the various chemical groups and their depth preferences and neighborhood preferences were found. The performance of the potential was benchmarked using its ability to identify the native structure from a very large set of decoy structures and we compared its performance with six other already existing statistical potentials. Our potential identified the native structures correctly 80% of the time and ranked third best (after DOPE and DFIRE) in its accuracy. We believe that if we increase the number of experimentally determined structures used to build our statistical potential then it should be able to beat DOPE and DFIRE. Our potential was especially successful for very small proteins as well where all the other potentials failed. The most important aspect of our potential is that it can be used to identify local regions within a protein which are incorrectly packed.

This pairwise potential will be further refined by taking more structures in the training set and including DEPTH preferences of the groups in its formulation. Also a multibody potential will be developed in the near future.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Protein Structure Prediction

Proteins are the structures which govern and constitute most of the biological processes. These macromolecules are fundamental to various metabolic processes as they help in catalysis, replication and repair, transportation and so many more. This is called as the central dogma of life. Each protein is made up of amino acid residues, and there are 20 such naturally occurring amino acids. A long chain of amino acid sequence is called a polypeptide sequence, which exists in every protein, and which is decided by the nucleotide sequence of the DNA (Bruce Alberts et al., 2008).

The folded protein with a well-defined three-dimensional stable structure is known as the native state of that protein. The resulting three-dimensional native structure is determined by the amino acid sequence (primary structure of the protein)(Tanaka & Scheraga, 1976)(Anfinsen, 1972). The residues interact among each other to give rise to unique three dimensional structures that are energetically stable, which dictate the activity and the function of the protein. The three dimensional structure of protein sequence is instrumental to figure out various properties of a protein, as well as the role that it plays. It is important also because, it gives us insight about the various metabolic processes that the given protein plays a part in. Also, there are many proteinopathies, diseases caused by abnormally folded proteins like Alzheimer's or Prion diseases (B Alberts et al., 2014); cures for which have not been found yet. Understanding the structures which govern these proteinopathies, may lead us to design drugs, which will help us solve these problems. Therefore studying protein structures, and trying to determine them, is the need of the hour and essential by all means.

Protein structures can be predicted by using both experimental and computational methods. Some experimental methods are discussed below:

1.1.1 Experimental Methods to Determine Protein Structures

X-Ray Crystallography: Crystallised proteins are subjected to X-rays, which in turn produce various diffraction angles, from which the location of the atoms are estimated. Obviously, this method automatically becomes unsuitable for determining the structures of proteins that are difficult to crystallize. Also this process just produces a snapshot, and gives no information about the dynamics of the protein structure. Despite of these limitations it is the most preferred method of calculating protein structures in PDB.

NMR Spectroscopy: NMR spectroscopy looks at transitions of the states of nucleic positions of protein structures, under the influence of a strong magnetic field. NMR spectroscopy of proteins is done when the proteins are in a solution form, which gives a better knowledge of the dynamic behaviour of proteins. But NMR can be only done on smaller protein structures. So about only 10% of the protein structures are based on NMR studies (Shah, Sattar, Benanti, Hollander, & Cheuck, 2006)

Electron Microscopy: This method uses a beam of concentrated electrons to find the structure of the macromolecules. Like X-ray Crystallography, the proteins need to be crystallised prior to the procedure. Also multiple images from many angles can be taken for a better view of the protein structure. (Rudenberg, H Gunther and Rudenberg, Paul G (2010). "Chapter 6 – Origin and Background of the Invention of the Electron Microscope") (Nitta, 2008).

The complexity and the variations of the 3D structure of a protein, and the ability of folding into a stable native state without being induced by any additional genetic mechanisms make it difficult to predict the structure, in case the experimental methods like NMR and X-Ray Spectroscopy fail. Also, the experimental methods are time consuming and costly. So, Computational methods become the most important way to predict protein structures. There are three major ways to predict a structure of a 3-D molecule, computationally (Eswar et al., 2007).

1.1.2 Computational Methods to Determine Protein Structures

Ab initio Modelling: Firstly, there are ab initio methods which use laws of physics to simulate all the energetics of protein-folding, and finding the one which has the lowest free energy and hence closest to the native structure. Ab initio methods aim at developing protein models from scratch, and does not involve previously solved structures. This method tries to calculate the free energy of the protein and find the global minimum of this energy which corresponds to the native state of the protein. These methods require vast computational resources and are only feasible for small protein structures.

Comparative Modelling: It assumes that protein folds into a structure that is similar to the experimentally determined structure of some other homologous protein. This method uses previously solved structures as starting point or templates (Chothia & Lesk, 1986). It involves *homology modelling* which extracts information from known protein structures, and use them to model any similarly sequenced protein; under the principle that evolutionary related proteins often have similar structures. These depend upon the structures available in databases such as the Protein Data Bank (Berman et al., 2000). They also include aligning two protein sequences, which causes errors as even the best known alignment schemes is about 70% accurate (Zhang, 2008). There are also methods to predict protein structures which are known to have the same fold, by using a database to score proteins and associate each score with a structure, and inferring the highest scoring protein to be the one that really exists. This is known as *Protein threading.*

We need to validate the predicted structures for their accuracy and for that we use energy scoring functions or statistical potentials. Initial work by Anfinsen (Anfinsen, 1973) showed that proteins which were unfolded instantly refolded into a three dimensional structure. This led to the conclusion that the native state of a protein has the lowest free energy of all possible structures of a protein, also known as the Hypothesis of Thermodynamics of Protein Folding (Y. Zhou & Linhananta, 2002). So, we can construct a potential function which corresponds to the effective energy of a protein. Many computational methods have been developed, some of which use

quantum mechanics and thermodynamics to obtain the parameters for statistical potentials (Sippl, 1990). But the molecules considered were small, due to the limitations in quantum mechanical calculations. This was extrapolated to larger molecules, using the canonical ensemble hypothesis of statistical mechanics. These kinds of potentials were called “physics based potentials.” Being atomic scale models, these are extremely computationally intensive processes.

The other kinds of statistical potentials, which use previously known structures to come up with parameters for the statistical potential calculation are formulated. These type of potentials are known as “knowledge based potentials.”(Sippl, 1993).

1.2 Knowledge-Based Statistical Potentials

Knowledge based statistical potential is “an energy function derived from an analysis of known protein structures in the Protein Data Bank” (Berman et al., 2000).

$$E = - \log \left(\frac{Observed \;\; Probability}{Expected \;\; Probability} \right)$$

It employs the idea that if a particular interaction is occurring more frequently, it is more preferred while folding. Knowledge-based statistical potentials are based on the assumption that frequently observed structural features correspond to low-energy states. Tanaka and Scheraga were the first to employ the above assumption to estimate pairwise amino acid interaction potentials by converting the observed frequencies of amino acid pairs into effective free energies (Tanaka & Scheraga, 1976). Since then many variants of pairwise amino acid potentials have extended this idea (Miyazawa & Jernigan, 1996).

Although they use a pseudo-energy function as a measure of the real energies of a model, they take much lesser time, and computational power, therefore being cheaper and faster. Most of the knowledge based statistical potentials use the idea that low energy states are represented by frequently occurring interactions. Some of the existing potentials are GOAP (H. Zhou & Skolnick, 2011), DFIRE (H. Zhou & Zhou, 2002), DOPE(Shen & Sali, 2006), Rosetta(Kim T Simons, Bonneau, Ruczinski, &

Baker, 1999), R(Kim T Simons et al., 1999), Rosetta (K T Simons, Kooperberg, Huang, & Baker, 1997), ANOLEA (F Melo, Devos, Depiereux, & Feytmans, 1997), ModPipe (Eramian et al., 2006) etc.

Most of the knowledge based distance oriented pairwise statistical potentials are contact based i.e. they work on distance oriented pairwise interactions between two units of the protein. If they are within a certain cut-off distance then they are said to be interacting with each other. We assign a probability or energy score to each possible pairwise interaction (defined as two units within a certain cut-off distance of each other) between two units. The energy of the model is then the sum total of combined energy of all pairwise contacts in the structure. Now, these units can be residues or atoms.

1.3 Chemical Group and Orientation Based Potential

Each residue is essentially made up of atoms which are covalently connected. Here, we have taken these covalently connected atoms called chemical groups as units instead of amino acids and aim to construct a new chemical-group based statistical potential which could give better resolution and predict the structures more accurately. The distance between the groups is defined as the distance between their centroids. These chemical groups are covalently bonded atoms within the residues and linear combination of these groups can form all the 20 residues. As the number of chemical groups is more than the residues, we can have more possible pairwise interactions which will give us a more accurate statistical potential with better resolution.

We also take into account the orientation of the chemical groups with respect to each other by defining an initial direction vector for these groups for differentiating between the relative orientations of these groups when they occur as pairs. This potential should perform better than conventional distance based and pairwise potentials because it can differentiate between the angles of the groups with giving higher scores to preferred orientations and vice-versa. The groups in Fig 1.1 illustrates an example where both the groups are within the same distance but have different orientations w.r.t. each other. These two interactions will get the same score by an only distance-based potential however in our case we will be able to identify the

difference between the two interactions. We formulate two potentials: one distance based and another distance-angle based.

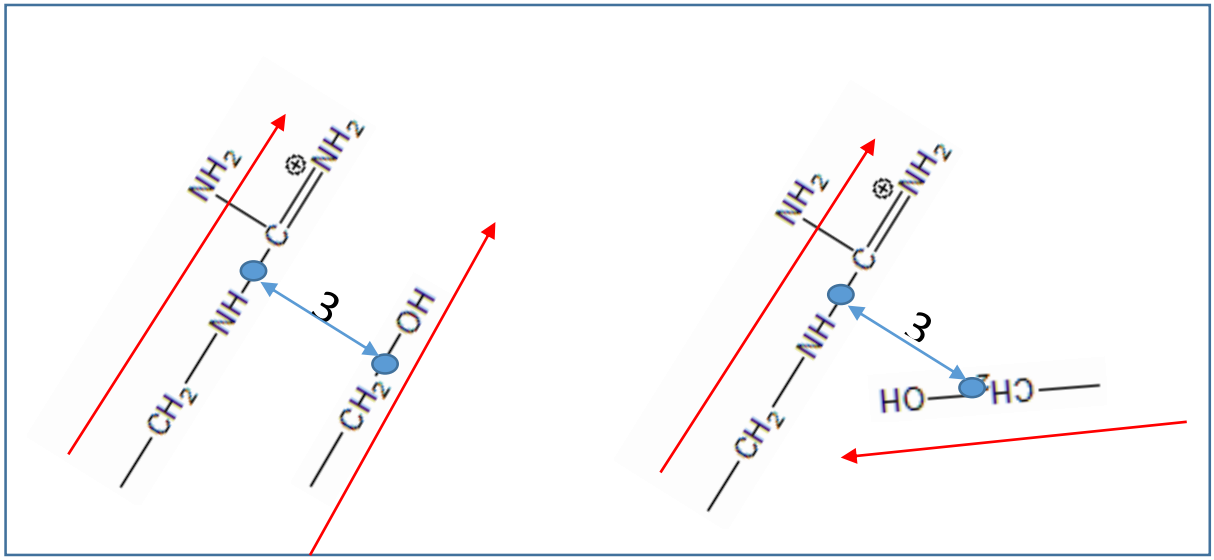

Fig 1.1: Both the pairs will be given the same score by a distance based potential as they both have equal distances between their centroids. However our distance-angle based potential will be able to differentiate between these two interactions.

This is a pairwise potential which can be extended and generalized for multibody interactions to formulate a multibody potential using n body cliques. A clique is a collection of neighboring groups where each constituent group is a neigbhor of the other i.e. all of them are within the given cut-off distance. For now, we try to construct a distance and angle based knowledge-based statistical potential.

Chapter 2

Methods

2.1 Defining Chemical Groups and their Orientation

We are going to define 16 functional groups, whose combination will form all the 20 naturally occurring amino acids. The Hydrogen atoms in the residues have not been included in these groups. In these 16 chemical groups, the atoms are differentiated based on the atom they are covalently bonded to i.e. a Carbon atom bonded to two Carbon atoms(and two implicit Hydrogens) is different than a Carbon atom bonded to one Nitrogen and one Carbon atom(and two implicit Hydrogens). Also primary, secondary and tertiary Carbon atoms are treated differently. Following figures, illustrate that how the groups have been named and how all the twenty amino acids can be divided into the sixteen groups.

The chemical groups (hereafter mentioned as groups) are named as r_1, r_2 …r_{16}. All the groups have been illustrated in Figure 2.2. The main chain is named r_1 which includes the peptide bond from the next residue. Glycine is like main chain and counted as r_1 and Proline is considered as r_{11}, which also includes the peptide bond with –N from next residue (If the next residue is also Proline, then the peptide bond with the N from the cyclic side chain of that Proline is considered.) The N atom within the same Proline residue is counted twice, once with the main chain of the previous residue to include the peptide bond and once with r_{11}. The Nitrogen atom involved in the peptide bond formation is always counted with the main chain of the previous residue. The grouping of r_1 and r_{11} groups is illustrated in figure 2.1. For the terminal residues, we are left with the $–NH_2$ group of the first residue (given first group is not Proline) and -C-COOH of the last residue. This $–NH_2$ group is taken as r_5, and the missing Carbon is ignored. The –C-COOH group is counted as r_6 which includes the terminal Oxygen (OXT). The last residue does not contain r_1 as its –C-COOH is counted with r_6 and N is counted with r_1 from previous residue.

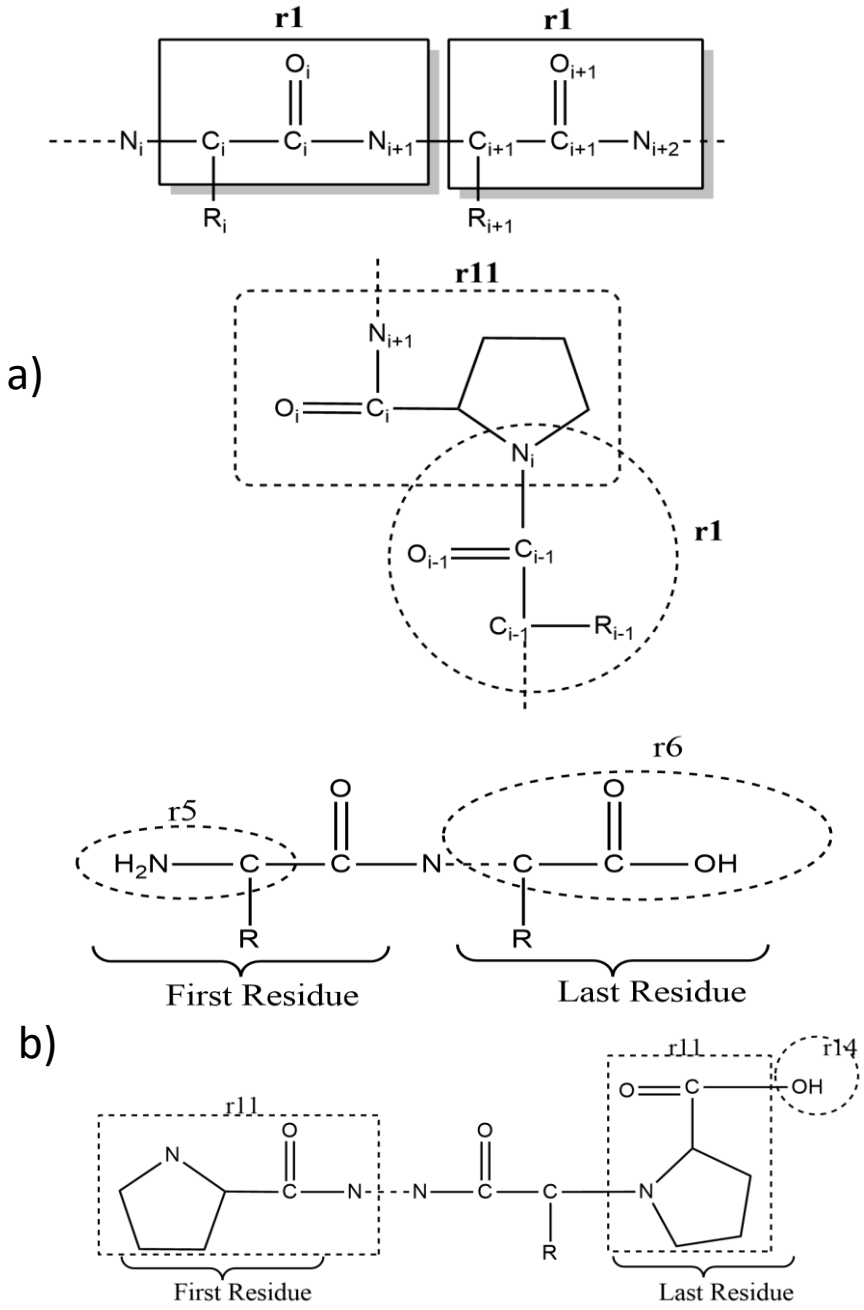

Fig 2.1: a) r_1 and r_{11} groups, where i is the current residue number. Notice, that N from cyclic side chain of Proline is being counted twice. **b)** Grouping of terminal residues. The C-alpha from first residue is counted twice with r5.

However, if the last residue is Proline then it is counted as $r_{11} + r_{14}$ i.e. –CO of –COOH is counted with r_{11} and remaining OH (OXT) is counted as r_{14}. The r_{11} from the terminal residue has only one N. The algorithm for converting from atoms to groups is as follows:

- If (residue number equals starting residue) and (residue is not Proline) then groups are r5, r1 and groups from the sidechain
- If (residue number not equals ending residue) and (residue is not Proline) then groups are r1 and groups from the sidechain
- If (residue number not equals ending residue or residue number equals starting residue) and (residue is Proline) then group is r11
- If (residue number equals ending residue) and (residue is not Proline) then groups are groups from the sidechain and r6
- If (residue number equals ending residue) and (residue is Proline) then groups are r11 and r14. This r11 won't include the peptide bond from the next residue as this is the terminal residue.

For the Cartesian co-ordinates of each group we took the centroid of the group by taking the average of the co-ordinates of the constituent atoms. After this we define a direction vector for each group to get an angle for the relative orientation of one group with respect to another when they occur as neighbors. We chose two atoms for each group and take the direction vector along the line joining those two atoms as the initial orientation of that group. If the group contains only one atom (like r2, r8) then we take covalently bonded atom attached to this group as the second atom. If we have two atoms A and B then direction vector is defined as:

$$\overrightarrow{D} = (x_A - x_B)\hat{i} + (y_A - y_B)\hat{j} + (z_A - z_B)\hat{k}$$

Now to find the angle between two groups, we find the angle between their direction vectors. This way, for each pair of groups we have a unique angle calculation because the initial direction vector of each group is fixed. The angle f for two vectors 1 and 2 is calculated as:

$$\theta = \cos^{-1}\left(\frac{x_1x_2 + y_1y_2 + z_1z_2}{\sqrt{x_1^2 + y_1^2 + z_1^2}\sqrt{x_2^2 + y_2^2 + z_2^2}}\right)$$

The angle ranges from 0 to 180 degrees. The combination of groups which forms all the 20 naturally occurring amino acids is listed in Table 2.1. The atoms which

are used to calculate the direction vectors for the groups have been listed in Table 2.2. The sixteen groups and their direction vectors are given in Figure 2.2.

1.	Arginine	r1 + r2 + r2 + r3	11.	Glycine	r1
2.	Histidine	r1 + r2 + r4	12.	Proline	r11
3.	Lysine	r1 + r2 + r2 + r2 + r5	13.	Alanine	r1 + r8
4.	Aspartic Acid	r1 + r2 + r6	14.	Valine	r1 + r12 + r8 + r8
5.	Glutamic Acid	r1 + r2 + r2 + r6	15.	Isoleucine	r1 + r12 + r8 + r2 + r8
6.	Serine	r1 + r7	16.	Leucine	r1 + r2 + r12 + r8 + r8
7.	Threonine	r1 + r7 + r8	17.	Methionine	r1 + r2 + r13
8.	Asparagine	r1 + r2 + r9	18.	Phenylalanine	r1 + r2 + r14
9.	Glutamine	r1 + r2 + r2 + r9	19.	Tyrosine	r1 + r2 + r14 + r15
10.	Cysteine	r1 + r10	20.	Tryptophan	r1 + r2 + r16

Table 2.1: The constituent groups for all the 20 types of residues

r1	N to C	r9	CG to ND2
r2	Previous C to this C	r10	CB to SG
r3	CD to NH1	r11	N to CG
r4	CG to NE2	r12	Previous C to this C
r5	CE to NZ	r13	Previous C to this C
r6	CG to OD1	r14	CG to CZ
r7	CB to OG	r15	Previous C to this O
r8	Previous C to this C	r16	CG to CH2

Table 2.2: The initial direction vectors for the 16 groups. The names of the atoms follow standard RCSB-PDB(Berman et al., 2000) naming convention.

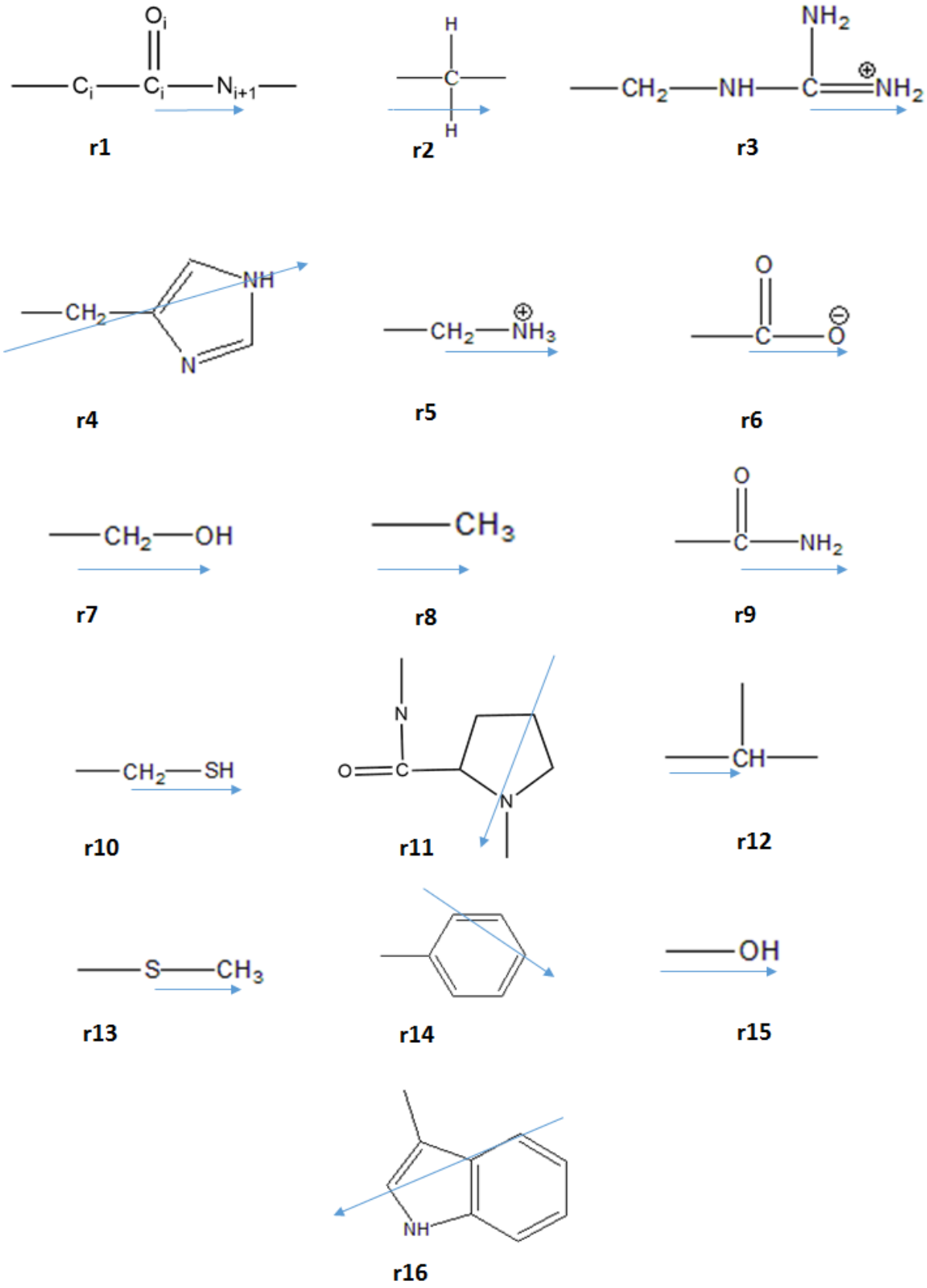

Fig 2.2: List of all 16 groups with their direction vectors marked with blue arrows

2.2 Construction of Sample dataset and DEPTH calculations

We take high resolution X-ray structures from RCSB Protein Data Bank (Berman et al., 2000). The PDB's were culled using the PISCES web server (Wang and Dunbrack, 2005) with the parameters given in Table 2.3. We obtained 2579 non-redundant structures and these were used to construct the sample dataset. The PDB codes for the sample dataset are listed in Appendix 1. The missing residues and atoms in the PDB's were built in by using MODELLER (Sali and Blundell, 1993). We only take single chains from the PDB's. The data from this training set was converted to group-form and along with the centroid and direction vectors of each group, the average DEPTH(Tan, Varadarajan, & Madhusudhan, 2011) of each group was also calculated.

DEPTH measures the extent of how much a particular atom/residue (or group) is buried within a protein. Depth of an atom in a protein is the distance from its nearest bulk solvent (water) (Chakravarty and Varadarajan, 1999). Depth of an atom is an important attribute of the protein and it co-relates with the protein stability. Chemical group depth is defined as the average of the depths of its constituent atoms.

We look at the depth profiles of various groups and it gives us important information about the hydrophilic/hydrophobic nature of the groups. Generally, hydrophobic groups will have more depth compared to hydrophilic groups because they are more deeply buried in the protein from the surface. We will be looking at the depth preferences of the groups and the average depth of the neighbors of a group at various depth. The standard deviation of the depth of the neighboring groups of a particular group conveys important information about the preferred interactions and hydrophilic/hydrophobic interactions of that group.

Sequence Percentage Identity	<= 30 %
Resolution	0.0 to 1.81 Å
R-Factor	0.25
Sequence Length	100 to 220
Non X-ray entries	Excluded
CA-only entries	Excluded
Cull PDB by	Chain
Cull chains within entries	Yes

Table 2.3: Culling Parameters for the PISCES Web Server for selecting the Training Set

2.3 Construction of Statistical Potential

We use the Cartesian coordinates of the centroid and direction vectors of all the groups from 2579 PDB's to construct pairwise distance based statistical potentials. We construct pairwise potentials for all 136 different possible pairs possible at various distance and angle distribution. For this, we need to define a neighbourhood distance cut-off. Two groups are said to be neighboring groups if the distance between their centroids is less than equal to a given cut-off distance. If two groups are neighboring groups, it is assumed that those two are interacting with each other and we count them as a single pair. We took the distance cut-off as 6 Å and to find all the pairs within this distance we used cell-list (taken from Neelesh Soni).We also calculate the relative orientation of one group with respect to the other (angle) in a pair by aforementioned inverse cosine formulae.

For getting a distribution of the potentials at various separations, we divide the distance domain into distance bins and for angles, into angle bins. The distance bins are as such: <1.3 Å, 1.3-1.4 Å, 1.4-1.5 Å….up to 6 Å. Each distance bin is further sub-divided into 18 angle bins from 0 to 180 degrees. We will now formulate the distance (mp_dist) and angle (mp_ang) based statistical potentials. For each type of pair we have 48 distance bins and 18 angle bins for each distance bin.

Total Number of Distinct pairs for 16 groups = 17*16/2 = 136

Total Number of Statistical Potential bins for distance = 136*48 = 6528

Total Number of Statistical Potential bins for angle = 6528*18 = 117504

We have two type of potentials-

1) mp_all: Groups coming from same residue are also considered as pairs i.e. even if they are covalently bonded they are counted as pairs if the inter-centroid distance is less than the distance cut-off, which is most likely true for all residues as all of the groups from the same residues will not have a separation of more than 6 Å. So, for any group we will always count the same residue groups as neighbors.

2) mp_diff: Groups from the same residues and covalently bonded groups are excluded in the pair count. In this case, for r1-r1, r1-r11 and r11-r11 pairs we do not consider consecutive groups as pair because they are always covalently bonded via the peptide bond. The difference between the residue numbers of the parent residues should be greater than equal to two in such cases to ensure that no two covalently bonded groups are counted as pairs.

2.3.1 Distance based Statistical Potential

$$S_{i.j}(d) = -\ln\left(\frac{\frac{f_{i,j}(d)}{\sum_{i,j=1}^{16} f_{i,j}(d)}}{\frac{f_i(d)\,f_j(d)}{\left(\sum_{total} N(d)\right)^2}}\right)$$

where,

$$f_{i,j}(d) = \text{Frequency of observed } r_i - r_j \text{ pairs at } d \text{ distance}$$

$$\sum_{i,j=1}^{16} f_{i,j}(d) = \text{Frequency of all observed pairs at } d \text{ distance}$$

$$f_i(d) = \text{Frequency of } r_i^{th} \text{ group at distance } d$$

$$f_j(d) = \text{Frequency of } r_j^{th} \text{ group at distance } d$$

$$f_i(d)f_j(d) = \text{Number of maximum possible } r_i - r_j \text{ pairs at } d \text{ distance}$$

$$\left(\sum_{total} N(d)\right)^2 = \text{Number of maximum possible pairs of all type at } d \text{ distance}$$

If the frequency of observed pairs for a particular r_i-r_j pair is 0, then we give it a background count of 1. While scoring, this distance bin will get a bad score because we are taking negative logarithm.

This formulation is similar to free energy approximation with the numerator as observed probability and denominator as expected probability i.e. r_i-r_j came together by chance. This is similar to a potential energy measure of observed interacting pairs, with favourable interactions getting a lower energy score. We will have two type of potentials from this:

1) mp_dist_all : all the pairs are included
2) mp_dist_diff : pairs from the same residues and covalently bonded pairs are excluded

The distance distribution of the mp_dist_diff for one pair (r_1-$r1_2$) has been shown in Fig 2.3. We have 136 such pairs.

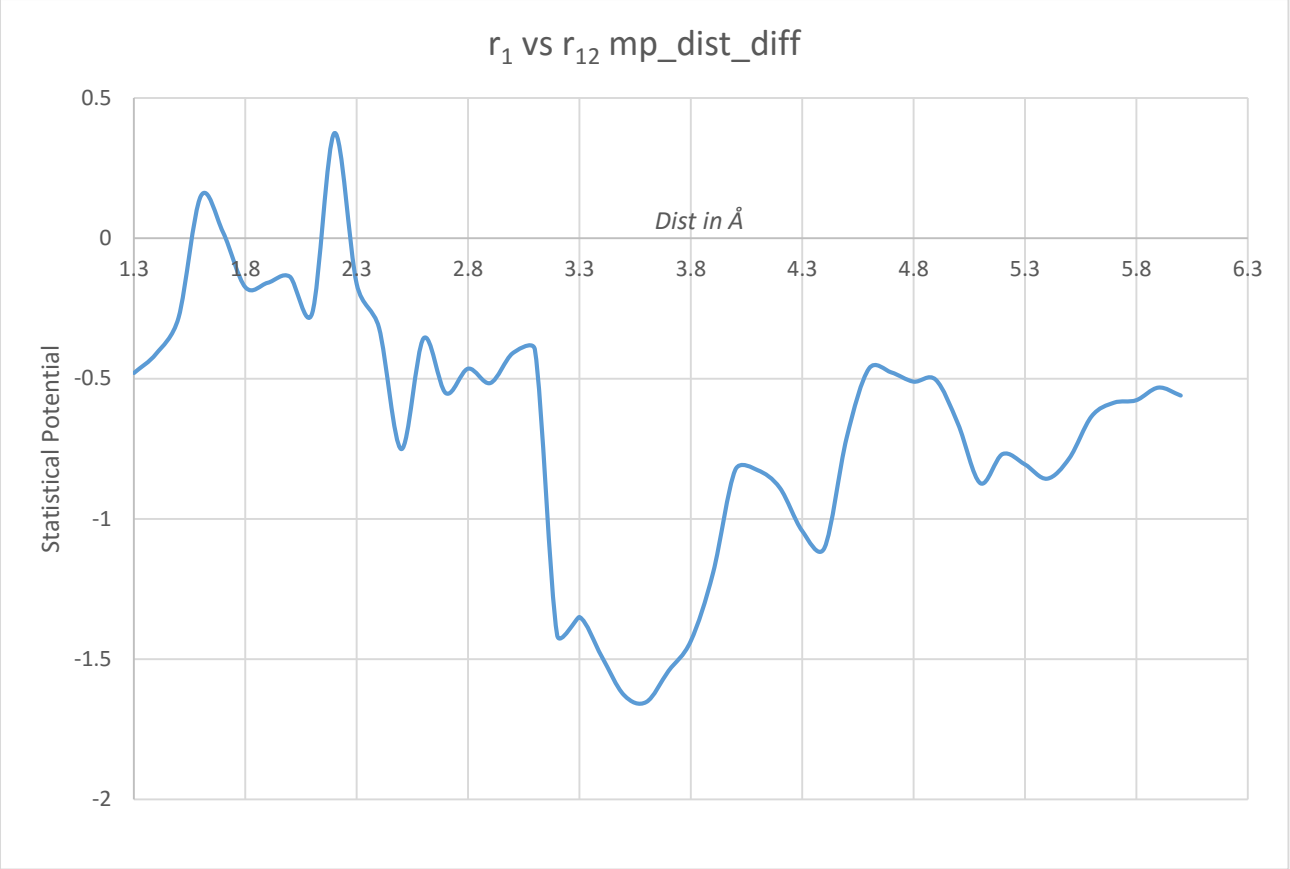

Fig 2.3: Distribution of Statistical potential (mp_dist_diff) for r_1-r_{12} pair.

2.3.2 Angle based Statistical Potential

At each distance bin we have 18 angle bins and essentially the Statistical potential at each distance bin is the sum of the Statistical potential from the 18 angle bins. So, we can divide the mp_dist potential according to the ratio of the observed counts from each angle bin to total count at that distance bin to obtain the mp_ang_1 potential. The potential is:

$$S_{i,j}(d,\theta) = \frac{f_{i,j}(d,\theta)}{f_{i,j}(d)} * \left(-\ln\left(\frac{\frac{f_{i,j}(d)}{\sum_{i,j=1}^{16} f_{i,j}(d)}}{\frac{N_i(d)N_j(d)}{\sum}} \right) \right) = \frac{f_{i,j}(d,\theta)}{f_{i,j}(d)} * S_{i,j}(d)$$

$f_{i,j}(d,\theta)$ = Frequency of observed $r_i - r_j$ pairs at d distance and θ orientation angle b/w their direction vectors.

$f_{i,j}(d,\theta)/f_{i,j}(d)$ = Fractional contribution of θ^{th} bin to d^{th} distance bin

We can also multiply the distance score by fractional contribution first and then take negative logarithm for individual distance bins. We will call this potential mp_ang_2.

$$S_{i,j}(d,\theta) = -\ln\left(\frac{f_{i,j}(d,\theta)}{f_{i,j}(d)} * \frac{\frac{f_{i,j}(d)}{\sum_{i,j=1}^{16} f_{i,j}(d)}}{\frac{N_i(d)N_j(d)}{\sum}}\right)$$

$$S_{i,j}(d,\theta) = -\ln\left(\frac{\frac{f_{i,j}(d,\theta)}{\sum_{i,j=1}^{16} f_{i,j}(d)}}{\frac{N_i(d)N_j(d)}{\sum}}\right)$$

This construction is similar to the mp_dist potential except in place of the frequency of observed r_i-r_j pairs at a particular distance bin, we take the frequency for the said angle bin at that distance. The background count for angle bins with observed frequency =0 is 1/18. Earlier, we took background count for distance bin as 1 and this is equipartitioned among the 18 angle bins. The sum of the observed frequencies of all 18 angle bins of a particular distance bin will be equal to observed frequency for that distance bin. Here also we will have two kinds of potential: one with all residues (mp_ang_1_all, mp_ang_2_all) and one which excludes same-residue and covalently bonded pairs (mp_ang_1_diff, mp_ang_2_diff). All the potentials, which we have formulized are listed in Table 2.4.

Distance Potential	mp_dist_all
	mp_dist_diff
Angle Potential	mp_ang_1_all
	mp_ang_1_diff
	mp_ang_2_all
	mp_ang_1_diff

Table 2.4: Listed here are all the constructed potentials. _all potentials are the one which include pairs from all residues and _diff are the ones which exclude same-residue and covalently bonded pairs.

2.4 Scoring PDB Structures

To test a PDB using our statistical potentials, we find out all the pairs of groups within 6 Å cut-off distance using cell list. We also find the orientation of these groups with respect to each other i.e. angle between their direction vectors. Now, each of these pairs will fall into a particular distance and angle bin. We assign the statistical potential score of that bin to that pair and we do this for all the pairs in the PDB. The sum of statistical potential scores for all the pairs in the PDB will give us the total score of that PDB. When we have to calculate the angle potential, we look at the angle bins and distance bins for distance potential. While scoring also, we'll leave out within same residue pairs and covalently bonded pairs when calculating mp_diff potentials. All pairs will be included when we are calculating mp_all statistical potential sores.

We also add the scores for the pairs of each residue and get a residue wise statistical potential score. We build a residue wise profile and can find out the residues where the score is higher corresponding to nearby residues (i.e. unusually high peak in residue number vs statistical potential score graph). This helps us in identifying the sub-optimally folded regions in the protein and we can try to correct that. We can also normalize the total PDB score using the number of groups. This helps us in comparing proteins which have different residue numbers.

2.5 Benchmarking of statistical potentials

The statistical potentials obtained are benchmarked by testing and scoring six set of decoy structures and rank ordering the native structure. The quality of a statistical potential is benchmarked on its capability to recognize the best model in a set of decoys. Ideally, the native structure should always get rank 1 because it is the most stable structure with least potential energy. We compare the results of our potentials with six other potentials (DOPE, DFIRE, Rosetta, ModPipe-Pair, and ModPipe-Surf). We also calculate score-error Pearson correlation coefficients between the C alpha RMS errors and statistical potential scores. This helps us in correlating the structural similarity of a decoy model to its native state.

Accuracy of the statistical potentials is also benchmarked based on the basis of its ability to select the most accurate non-native model. We measure this by delta RMSD i.e. the difference between the RMSD of the model selected by our statistical potential as most accurate and the actual best non-native model (based on Cα RMSD). If delta RMSD is 0.0 Å, then our statistical potential is successful in predicting the most accurate non-native model. The identification of the non-native structure which is closest to the native structure from a set of decoys is generally considered significantly more difficult than identification of the native structure because in realistic model assessment the native structure is not known.

We also try to find a relation between per-residue score and Cα RMSD. For this we will take a large set of decoy structures with varying RMSD's and calculate their per-residue score. Per-residue score allows us to compare structures with different residue lengths. For identifying the sub-optimally packed regions in a structure we will do residue-wise and group-wise profiling of their statistical potential score. We can either compare these residue-wise/group-wise scores to the corresponding residue/group in the native structure or find out abnormally high peaks (least negative scores/positive scores) in the profile. Thus, we can identify those regions in the structure which are incorrectly packed and try to rectify those regions.

Chapter 3

Results

3.1 Depth Analysis of the Chemical Groups

We find the average depth of all the 16 groups using DEPTH (Tan et al., 2011). The preferences are plotted in Fig 3.1. As expected, the hydrophilic groups (r3, r5, r9 and r15) have lower depths than the hydrophobic groups (r8, r10, r12, r13, r14, r16).

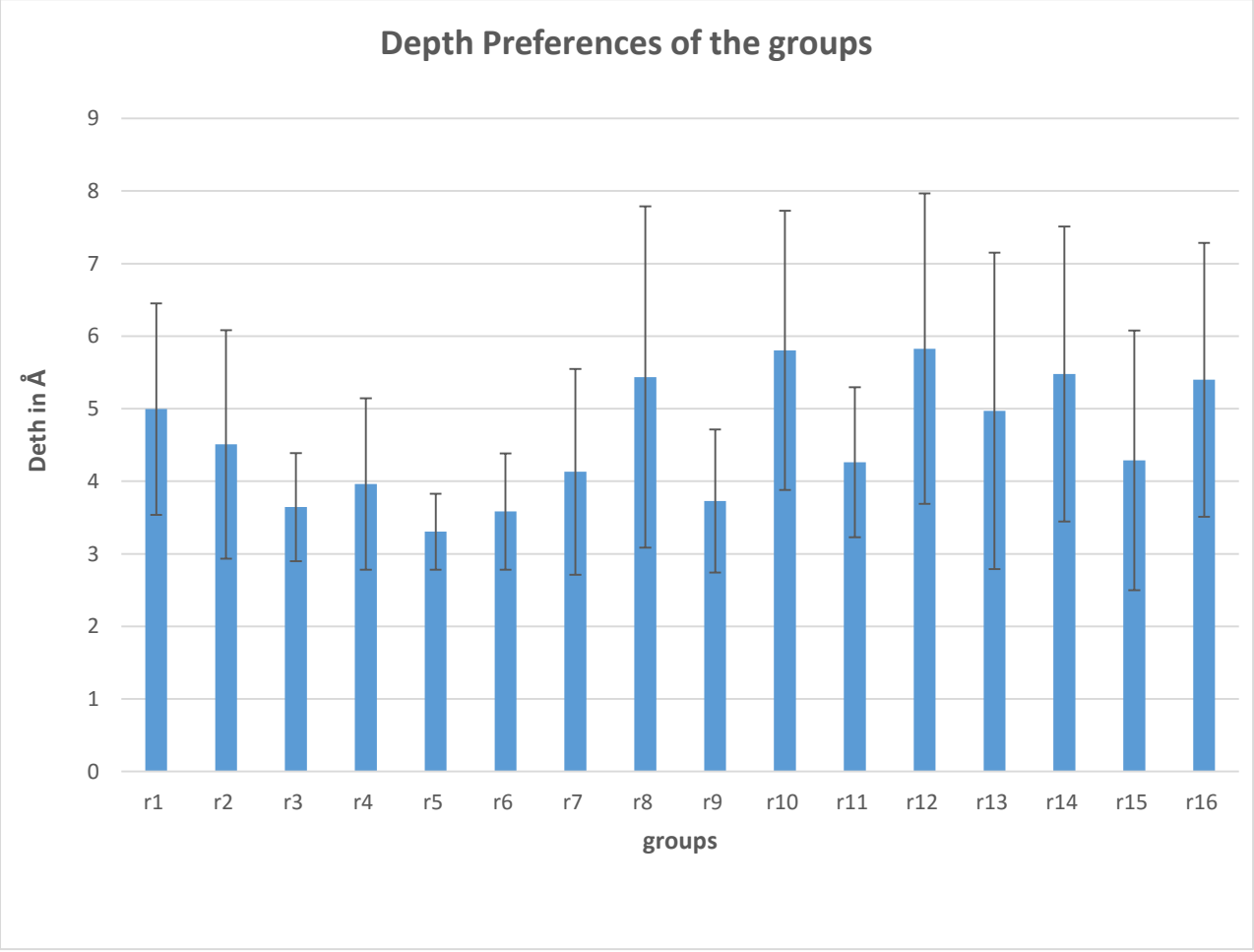

Fig 3.1: Average Depth Preferences of the groups in A^0

We calculate the average depth of the neighboring groups at each depth bin and find it to be a straight line. This tells us that the more we increase the depth, the depth of the neighboring groups will also increase i.e. the more buried groups will interact with other groups at the same buried depth. Hydrophobic groups will have higher depth and so their neighbors also will have higher depth and these groups will have hydrophobic interactions. Same is true for hydrophilic groups. They will reside at the surface forming hydrogen bonds with the solvent and their neighboring groups will also be hydrophilic, encouraging hydrophilic interactions.

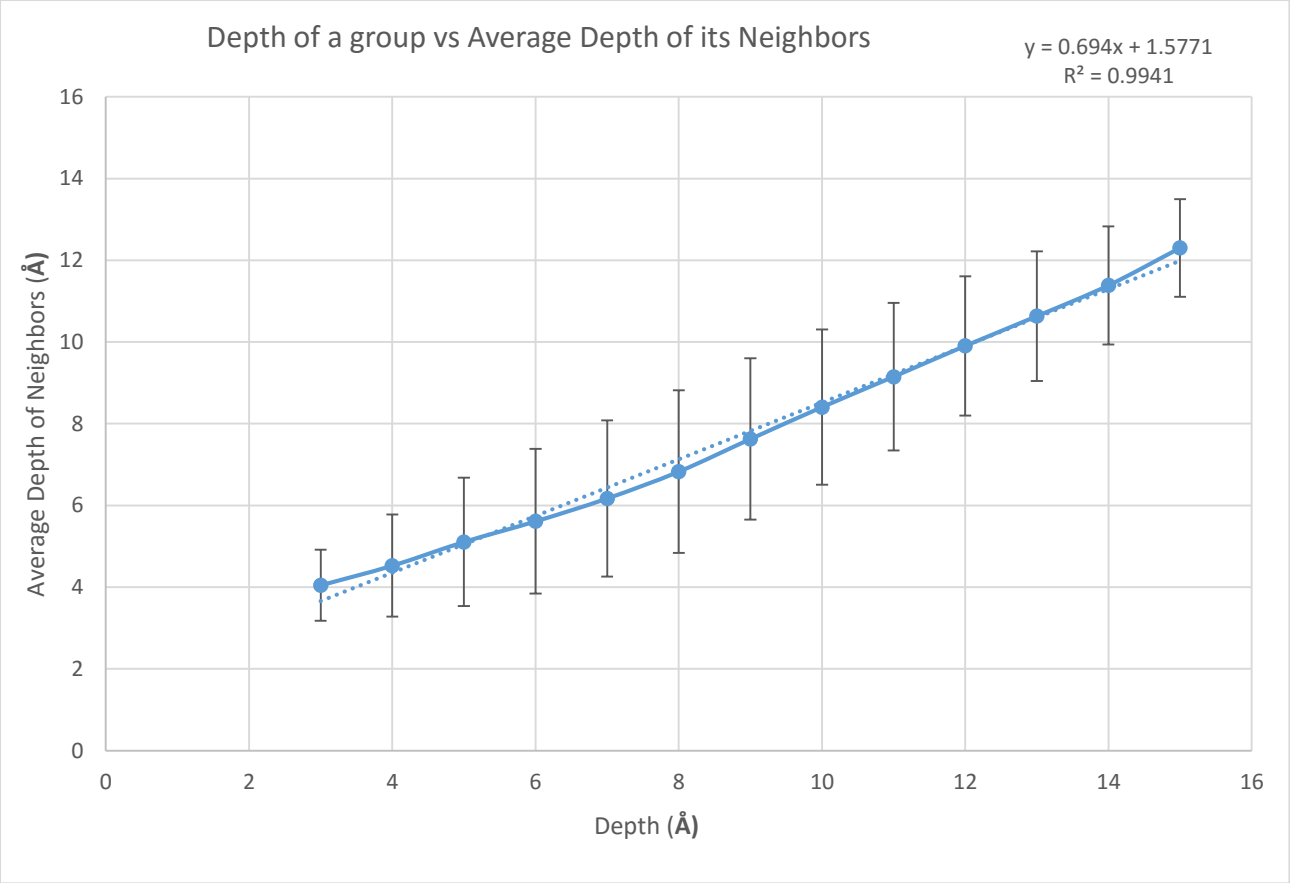

Fig 3.2: The plot between Depth of a group and Mean Depth of the neighboring groups. The dotted blue line represents the linear regression line. We see that this is a straight line, so with increasing depth of the groups, the depth of its neighbors also increases linearly.

3.2 Benchmarking of statistical potentials

3.2.1 Native State Detection

MOULDER SET

We first assess our statistical potentials on their ability to correctly identify the native states in the moulder decoy set (Eramian et al., 2006). This set is constructed using iterative target-template alignment and comparative model building of 20 target sequences using MODELLER-6 (John & Sali, 2003). This set consists of 300 models for each target and we rank order the native structure. If the native structure is assigned rank 1, then it is correctly identified. We also compare our results with six other scoring functions. The native ranks for our potentials are listed in Table 3.1.

	MP_DIST_ALL	***MP_ANG_1_ALL***	***MP_ANG_2_ALL***	***MP_DIST_DIFF***	***MP_ANG_1_DIFF***	***MP_ANG_2_DIFF***
1BBH	1	225	182	1	1	246
1C2R	3	292	178	1	2	233
1CAU	3	220	184	1	1	248
1CEW	4	283	213	1	1	254
1CID	23	301	142	1	1	174
1DXT	1	100	78	1	1	145
1EAF	1	239	221	1	1	374
1GKY	1	293	243	1	1	282
1LGA	1	1	101	1	1	184
1MDC	1	272	120	1	1	179
1MUP	263	301	9	99	30	23
1ONC	1	99	192	1	1	229
2AFN	1	199	244	1	1	265
2CMD	1	32	106	1	1	254
2FBJ	7	2	66	7	1	135
2MTA	6	78	282	1	2	235
2PNA	225	301	62	158	114	82
2SIM	1	236	87	1	1	166
4SBV	2	1	168	1	1	194
8I1B	22	227	24	51	7	45
CORRECT	***10***	***2***	***0***	***16***	***15***	***0***
ACCURACY	***50.00%***	***10.00%***	***0.00%***	***80.00%***	***75.00%***	***0.00%***

Table 3.1: Ranks of the native structures for the 20 target proteins for MOULDER set.

We see that mp_all potentials perform very poorly. The mp_ang_2 potentials (mp_ang_2_all and mp_ang_2_diff) are not accurate compared to mp_diff potentials. So, from now on we only consider two potentials- mp_dist_diff as mp_dist (universal distance based potential) and mp_ang_2_diff as mp_ang (universal orientation based potential). The mp_dist potential correctly identifies 16 and mp_ang correctly identifies 15 native structures. DOPE (Shen & Sali, 2006), DFIRE (H. Zhou & Zhou, 2002), Rosetta(Kim T Simons et al., 1999), ModPipe-Pair, and Modpipe-Comb (Francisco Melo, Sánchez, & Sali, 2002) are able to correctly identify 19 out of the 20 native structures from the total of 301 conformations (Shen & Sali, 2006). Modpipe-Surf (Francisco Melo et al., 2002) failed to identify two native structures (2pna and 1mup) and ranked the other eighteen native structures correctly (Shen & Sali, 2006). All the potentials miss the native state of 2pna which is the only NMR structure in this set.

Decoys 'R' US Set

We take four multiple target decoy sets- *4state_reduced, fisa, fisa_casp3* and *lmds.* These are available at Decoys 'R' Us (Samudrala & Levitt, 2000) (http://dd.stanford.edu). The *4state_reduced* decoy set, has around 632 to 689 models for each of the seven targets and it is generated using a "four-state off-lattice model with a conformational relaxation method" (Park & Levitt, 1996). The *fisa* and *fisa_casp3* decoy sets each have four and three targets (500-1400 models per target), respectively which were generated using "a combination of a Bayesian scoring function and a simulated annealing protocol" (K T Simons et al., 1997). The *lmds* decoys set has 10 primarily short targets, with 215-500 models for each target. This set is obtained by a "local optimization method and a reduced ENCAD energy function" (Keasar & Levitt, 2003).

Our mp_dist correctly identifies 19 and mp_ang correctly identifies 12 native structures for 32 targets in these four multiple target decoy sets, while DFIRE, Rosetta, ModPipe-Pair, ModPipe-Surf, ModPipe-Comb and DOPE are successful for 19, 7, 11, 4, 10, and 20 targets, respectively (Table 3.3). Our mp_dist potential and mp_ang are the only potentials which correctly identify all 4 targets in the *fisa* set. The mp_dist

potential is the most accurate on lmds set as well with correctly identifying 8 native structures out of 10 targets, Dope being second with correctly identifying 7 of them. All scores fail to identify the native state of 1fc2 in the *fisa* set (except for mp_dist and mp_ang potentials) and 1b0n-B (except for mp_dist), 1bba, and 1fc2 in the *lmds* set. DOPE is the best performer on this set with 83.33% accuracy. Our mp_dist potential is tied with DFIRE at second place with 79.17% accuracy and mp_ang potential is at the third place with 50% accuracy.

Overall for all the five decoy sets (4state-reduced, fisa, fisa-casp3, lmds and MOULDER), mp_dist correctly identifies 35(80%), mp_ang identifies 27 (61%), DFIRE identifies 38 (86%), Rosetta identifies 26 (59%), ModPipe-Pair identifies 30 (68%), ModPipe-Surf identifies 22 (50%), ModPipe-Comb identifies 29 (66%), DOPE identifies 39 (87%) native states for 44 targets. All the target native structures in the five decoy sets are X-ray crystallography structures, except for 1bba in *lmds* and 2pna in *MOULDER* which are NMR structures. The DFIRE results are taken from original publication (H. Zhou & Zhou, 2002) The assessment results for DOPE, Rosetta, ModPipe-Pair, and ModPipe-Surf potentials are taken from original DOPE paper (Shen & Sali, 2006). The rankings of the potential based on their percentage accuracy is listed in Table 3.2.

	4 state-reduced	***fisa***	***fisa_casp3***	***lmds***	***MOULDER***	**Total**
1. DOPE	7/7	3/4	3/3	7/10	19/20	39/44= 87%
2. DFIRE	6/7	3/4	3/3	7/10	19/20	38/44= 86%
3. mp_dist	5/7	4/4	2/3	8/10	16/20	35/44= 80%
4. ModPipe-Pair	6/7	0/4	2/3	3/10	19/20	30/44= 68%
5. ModPipe-Comb	6/7	0/4	0/3	4/10	19/20	29/44= 66%
6. mp_ang	3/7	4/4	0/3	4/10	15/20	27/44= 61%
7. Rosetta	4/7	1/4	0/3	2/10	19/20	26/44= 59%
8. ModPipe-Surf	1/7	1/4	0/3	2/10	18/20	22/44= 50%

Table 3.2: Ranking of the potentials based on their percentage accuracy. The column indicates the decoy set and the rows indicate the scoring function.

	DFIRE	Rosetta	ModPipe-Pair	ModPipe-Surf	ModPipe-Comb	Dope	mp_dist	mp_ang
4state-reduced								
1ctf	1	1	1	1	1	1	1	2
1r69	1	2	1	17	1	1	1	1
1sn3	1	1	1	7	1	1	1	1
2cro	1	5	1	103	1	1	1	1
3icb	4	6	15	33	8	1	8	17
4pti	1	1	1	71	1	1	145	308
4rxn	1	1	1	18	1	1	1	3
Correct	6	4	6	1	6	7	5	3
fisa								
1fc2	254	158	491	1	453	375	1	1
1hdd-C	1	90	293	18	135	1	1	1
2cro	1	26	11	146	19	1	1	1
4icb	1	1	196	2	167	1	1	1
Correct	3	1	0	1	0	3	4	4
fisa_casp3								
1bg8-A	1	1068	1	1180	282	1	1	5
1bl0	1	960	4	912	86	1	31	13
1jwe	1	1177	1	1119	6	1	1	1
Correct	3	0	2	0	0	3	2	1
lmds								
1b0n-B	430	300	56	186	18	34	1	13
1bba	501	174	501	117	444	501	7	20
1fc2	501	291	325	54	222	476	346	83
1ctf	1	1	1	1	1	1	1	1
1dtk	1	9	4	1	1	1	1	1
1igd	1	1	1	3	1	1	1	2
1shf-A	1	5	24	18	7	1	1	1
2cro	1	2	4	28	12	1	1	1
2ovo	1	29	5	8	2	1	1	2
4pti	1	4	1	44	1	1	1	7
Correct	7	2	3	2	4	7	8	4
Total Correct	**19**	**7**	**11**	**4**	**10**	**20**	**19**	**12**
Accuracy	**79.17%**	**29.17%**	**45.83%**	**16.67%**	**41.67%**	**83.33%**	**79.17%**	**50.00%**

Table 3.3: Ranks of the native structures for the 24 target proteins in four multiple target decoy sets. The columns indicate tested scoring functions and the rows indicate pdb codes of the targets.

3.2.2 Correlation between the score and model error

If we want to describe the correlation between the score of a model and how similar it is to the native state, we find the Pearson correlation coefficient r between the Cα RMSD error of a decoy and its score. The value of r closest to 1 is the best and -1 the worst. We calculate the individual and average score-error correlation coefficients for the eight scoring functions for the *MOULDER* decoy set (Table 3.4). We get the lowest values for the coefficients and average value is 0.61 for mp_dist and 0.58 for mp_ang. However, this correlation is based on the assumption that Cα RMSD is the best measure of a structure to its native state which is not entirely true as it leaves out the side chains which plays a crucial role in determining the three dimensional structure and our potential gives us the overall packing of the protein structure which is more robust.

	DFIRE	Rosetta	ModPipe-Pair	ModPipe-Surf	ModPipe-Comb	Dope	mp_dist	mp_ang
1bbh	0.93	0.80	0.84	0.85	0.89	0.92	0.89	0.86
1c2r	0.92	0.81	0.66	0.84	0.82	0.93	0.80	0.79
1cau	0.76	0.86	0.57	0.67	0.66	0.80	0.63	0.61
1cew	0.68	0.75	0.59	0.61	0.63	0.68	0.44	0.31
1cid	0.88	0.86	0.54	0.83	0.77	0.89	0.64	0.57
1dxt	0.90	0.91	0.89	0.84	0.92	0.94	0.78	0.74
1eaf	0.82	0.80	0.74	0.80	0.81	0.84	0.56	0.57
1gky	0.57	0.79	0.85	0.84	0.87	0.73	0.45	0.51
1lga	0.89	0.89	0.83	0.82	0.87	0.91	0.66	0.67
1mdc	0.84	0.84	0.77	0.79	0.82	0.88	0.60	0.61
1mup	0.87	0.88	0.68	0.87	0.78	0.85	0.55	0.58
1onc	0.94	0.91	0.73	0.90	0.89	0.93	0.90	0.86
2afn	0.83	0.90	0.73	0.90	0.87	0.87	0.43	0.37
2cmd	0.88	0.87	0.86	0.81	0.86	0.90	0.57	0.57
2fbj	0.84	0.88	0.75	0.79	0.82	0.84	0.69	0.70
2mta	0.92	0.78	0.60	0.63	0.72	0.92	0.87	0.85
2pna	0.89	0.79	0.68	0.87	0.83	0.90	0.88	0.85
2sim	0.88	0.88	0.88	0.49	0.90	0.90	0.67	0.38
4sbv	0.77	0.83	0.68	0.66	0.75	0.83	0.00	0.02
8i1b	0.91	0.90	0.74	0.86	0.85	0.90	0.11	0.14
Average	**0.85**	**0.85**	**0.73**	**0.78**	**0.82**	**0.87**	**0.61**	**0.58**
Median	**0.88**	**0.86**	**0.73**	**0.82**	**0.83**	**0.89**	**0.60**	**0.59**

Table 3.4: Pearson correlation coefficient between the Cα RMSD error of a decoy and its score for the *MOULDER* decoy set for the eight scoring functions.

3.2.3 Selecting the most accurate non-native model

Generally, we don't have the native structure available and the task of the identification of most accurate non-native model is done in realistic model assessment. So, we assess the 300 decoy structures for the 20 target proteins for the *MOULDER* set and identify the model with the best score using our eight scoring functions. The difference between the structure with the lowest RMSD in the decoys (considered here to be the actual best non-native structure) and the best model selected using the scoring function gives us an measure of the accuracy of that scoring function. The lower the delta RMSD better is the accuracy of the scoring function. Here also we get the lowest accuracy for our mp_dist and mp_ang potential with average Δ RMSD = 4.14 Å and 4.49 Å.

	DFIRE	**Rosetta**	**ModPipe -Pair**	**ModPipe -Surf**	**ModPipe -Comb**	**Dope**	**mp_dist**	**mp_ang**
1bbh	0.00	0.07	0.00	0.07	0.07	0.00	0.23	0.95
1c2r	0.02	0.86	1.79	3.25	2.00	0.00	2.12	2.66
1cau	2.92	0.95	8.70	0.42	0.42	2.92	5.65	5.65
1cew	3.47	2.73	2.06	2.16	2.06	2.16	6.59	8.87
1cid	0.08	0.37	1.15	1.15	1.15	1.15	0.63	9.60
1dxt	1.11	0.55	1.03	4.18	0.00	0.55	1.99	1.42
1eaf	0.47	0.34	0.99	1.68	1.68	0.47	2.60	2.60
1gky	1.14	0.00	0.48	0.01	0.01	0.01	0.94	6.82
1lga	0.80	0.00	2.91	6.33	2.70	0.99	3.80	1.99
1mdc	0.22	0.05	0.74	3.75	0.74	0.02	0.13	0.30
1mup	0.67	0.08	0.40	0.69	0.40	0.26	2.72	2.72
1onc	0.40	0.48	0.72	0.40	0.72	0.35	3.41	0.66
2afn	0.12	0.00	0.88	0.50	0.71	0.12	3.07	4.55
2cmd	0.23	0.68	1.07	2.75	0.84	0.84	5.12	5.12
2fbj	0.91	0.91	2.80	0.26	2.80	0.91	8.90	4.59
2mta	0.63	2.85	2.34	0.65	0.57	0.21	1.14	0.63
2pna	0.00	0.07	0.60	0.00	0.07	0.00	1.02	1.02
2sim	0.16	0.27	0.13	1.26	1.12	0.16	8.64	8.19
4sbv	0.00	5.58	5.29	5.93	5.78	0.00	13.62	11.04
8i1b	0.50	0.50	1.35	0.78	0.78	0.50	10.46	10.46
Average	**0.69**	**0.87**	**1.77**	**1.81**	**1.23**	**0.58**	**4.14**	**4.49**
Median	**0.44**	**0.43**	**1.05**	**0.97**	**0.76**	**0.30**	**2.90**	**3.64**

Table 3.5: Δ RMSD (best value, 0.0 Å**)** for the eight scoring functions.

3.3 Relation between per-residue score and RMSD

We take 10000 structures from the SVMod-MODPIPE set (Eramian et al., 2006) with RMSD values ranging from 0 to 20 Å. We then calculate the per-residue score by dividing the total score for the protein structure (PDB file) by number of residues in the structure. We draw a scatter plot for the per-residue score and RMSD for all the 10000 PDB's for mp_dist (Figure 3.3) and mp_ang potentials (Figure 3.4). We see that with increasing RMSD the per-residue score also increases, as expected however, the rate of increase slows down. So, we fit a logarithmic regression line on the plot of RMSD vs average per-residue score at that RMSD (Figure 3.5 and Figure 3.6).

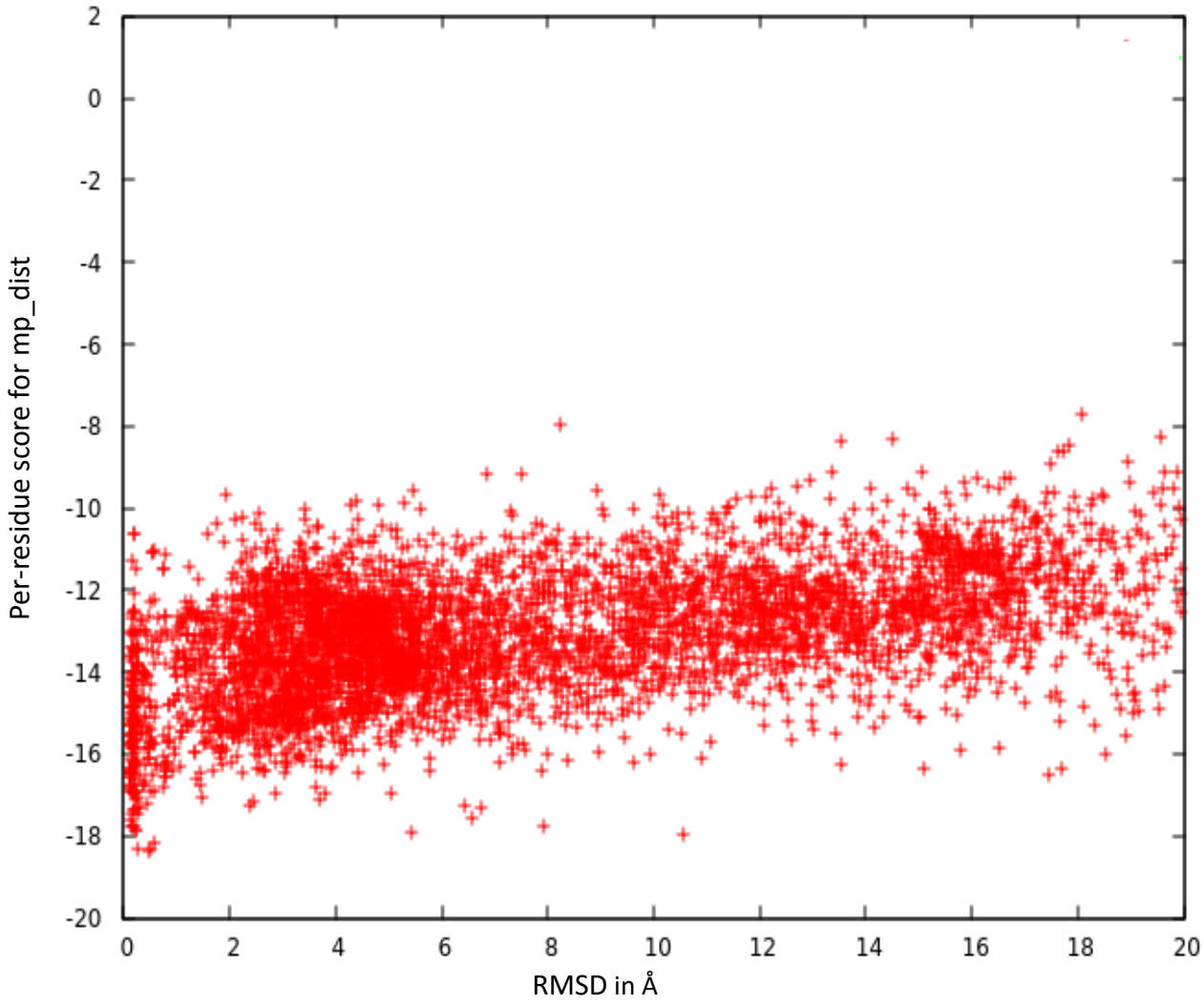

Fig 3.3: Scatter Plot of RMSD vs per-residue score for mp_dist potential for 10000 structures from MODPIPE set.

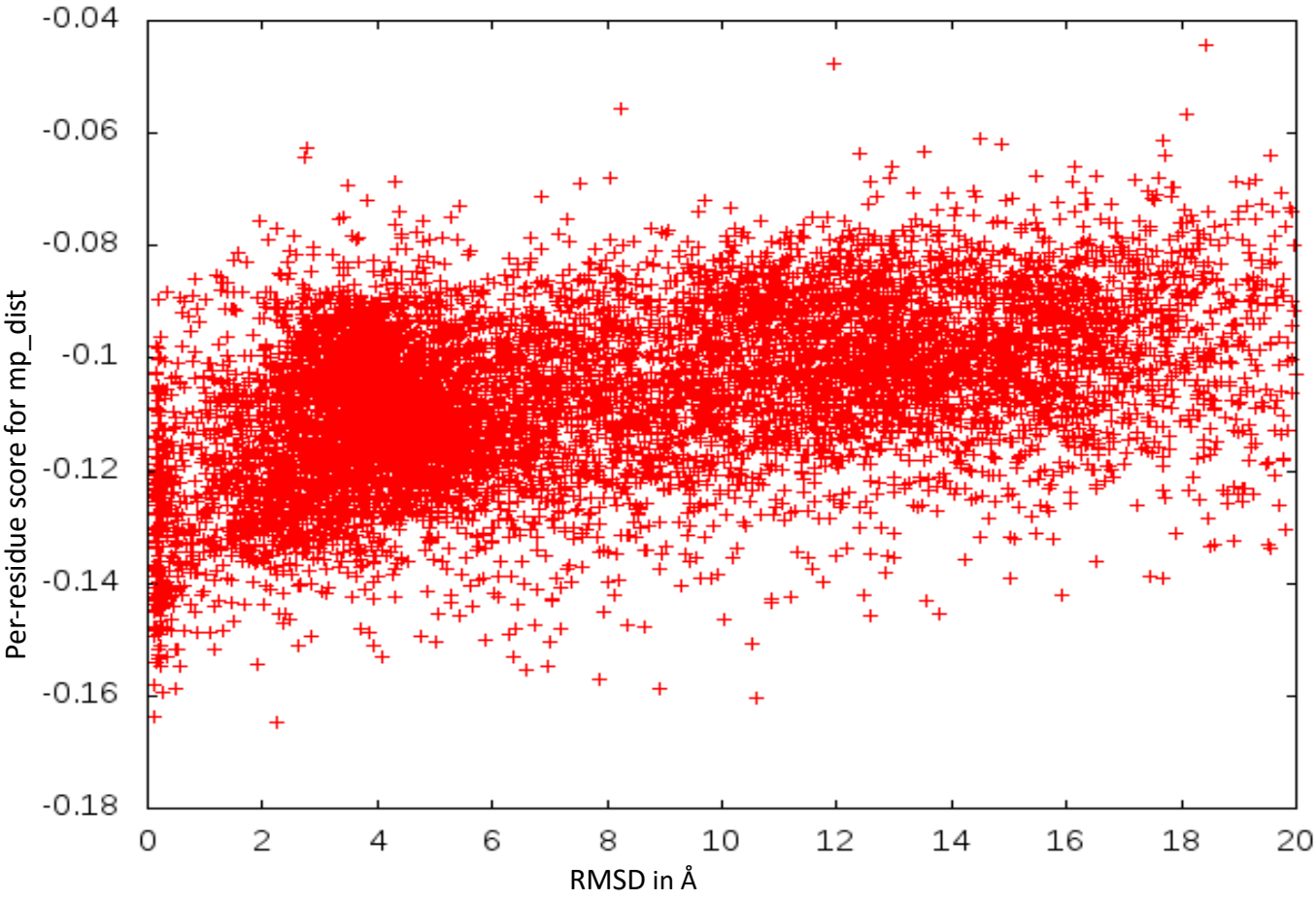

Fig 3.4: Scatter Plot of RMSD vs per-residue score for mp_ang potential for 10000 structures from MODPIPE set.

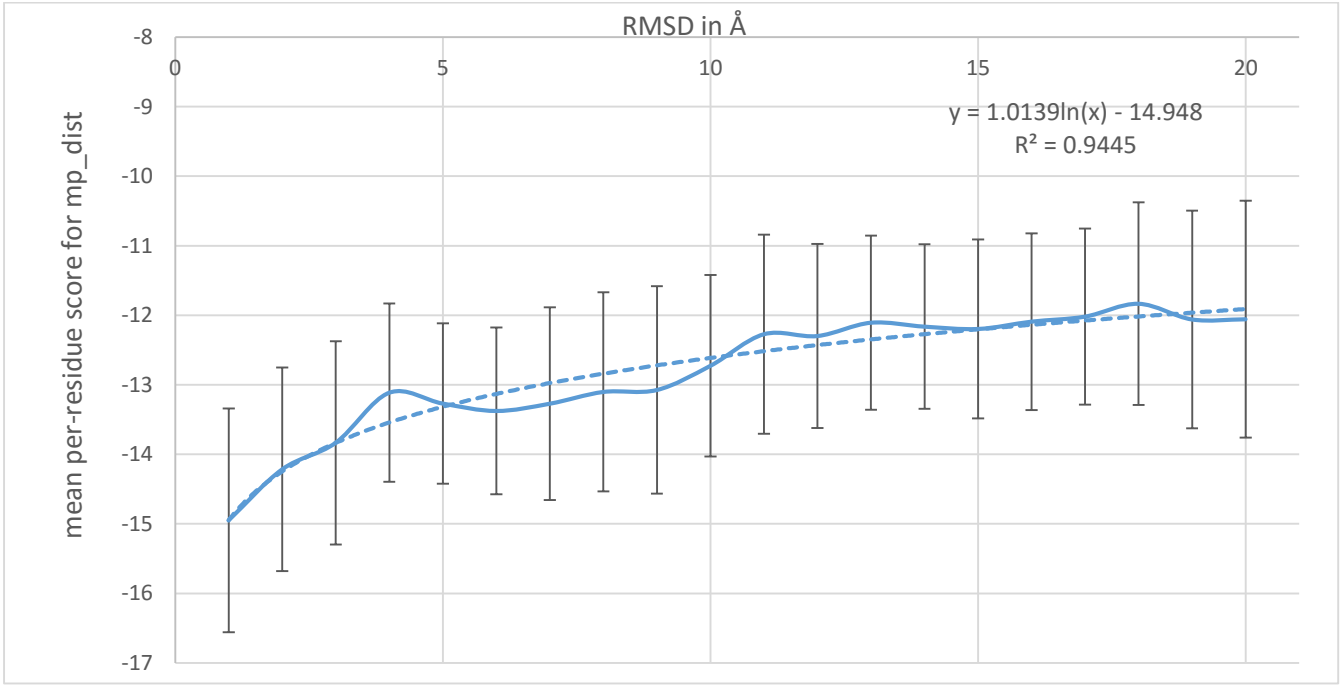

Fig 3.5: RMSD vs mean per-residue score for mp_dist potential. The dotted blue line is the fitted logarithmic regression line.

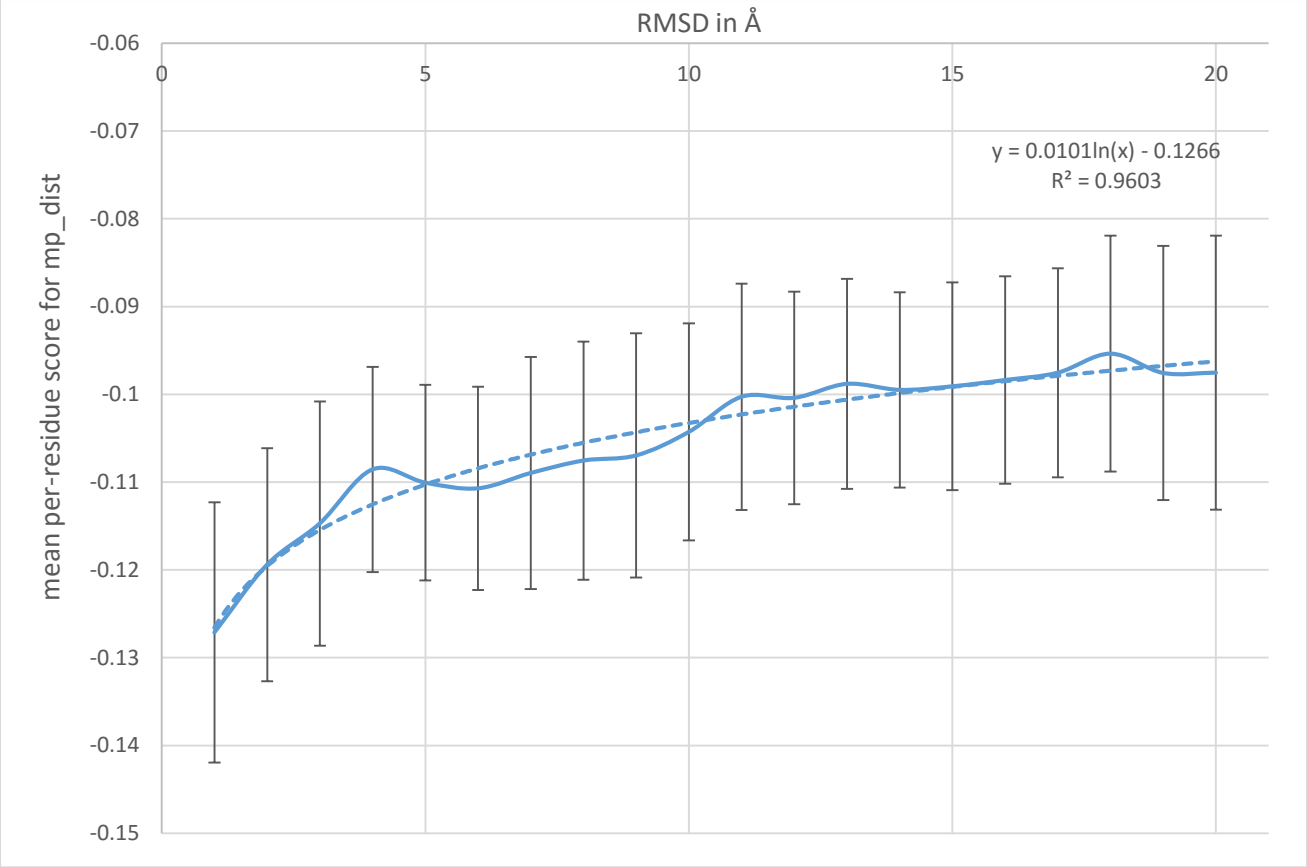

Fig 3.6: RMSD vs mean per-residue score for mp_ang potential. The dotted blue line is the fitted logarithmic regression line.

We can use the equation of the logarithmic regression line to back calculate the RMSD of any given PDB structure by calculating the per-residue mp_ang or mp_dist scores. We can then find the actual Cα RMSD of that PDB file using Chimera (Pettersen et al., 2004) or various other available tools. The Δ RMSD between the two RMSD values gives us an estimate of the accuracy of our statistical potentials and the dependence of the scores of a PDB structure on the RMSD value of that structure.

3.4 Residue wise profiling of the Structures

We can plot the potential scores for each residue for a protein structure w.r.t. residue number (residue wise) and see if there are any abrupt rise in scores i.e. are we getting any peaks which shouldn't be there. We can identify these peaks by looking at the scores of the neighboring environment, here, neighboring residues. The residues with unusually high scores (in magnitude) correspond to sub-optimally packed regions for the structure which we can look at and try to correct. Another way to identify such regions is superimposing the residue-wise plot of the decoy structure to the residue wise plot of the native structure (if it is known). We can look at the residue number where the decoy structure doesn't follow the same pattern as the native or it has unusually high score differences. The native structure should have the lowest over-all as well as lowest residue-wise scores. We did this for 1BBH protein structure and we've plotted the graphs for mp_dist (Fig 3.7) and mp_ang (Fig 3.8). The decoy structure has RMSD of 3.22 Å.

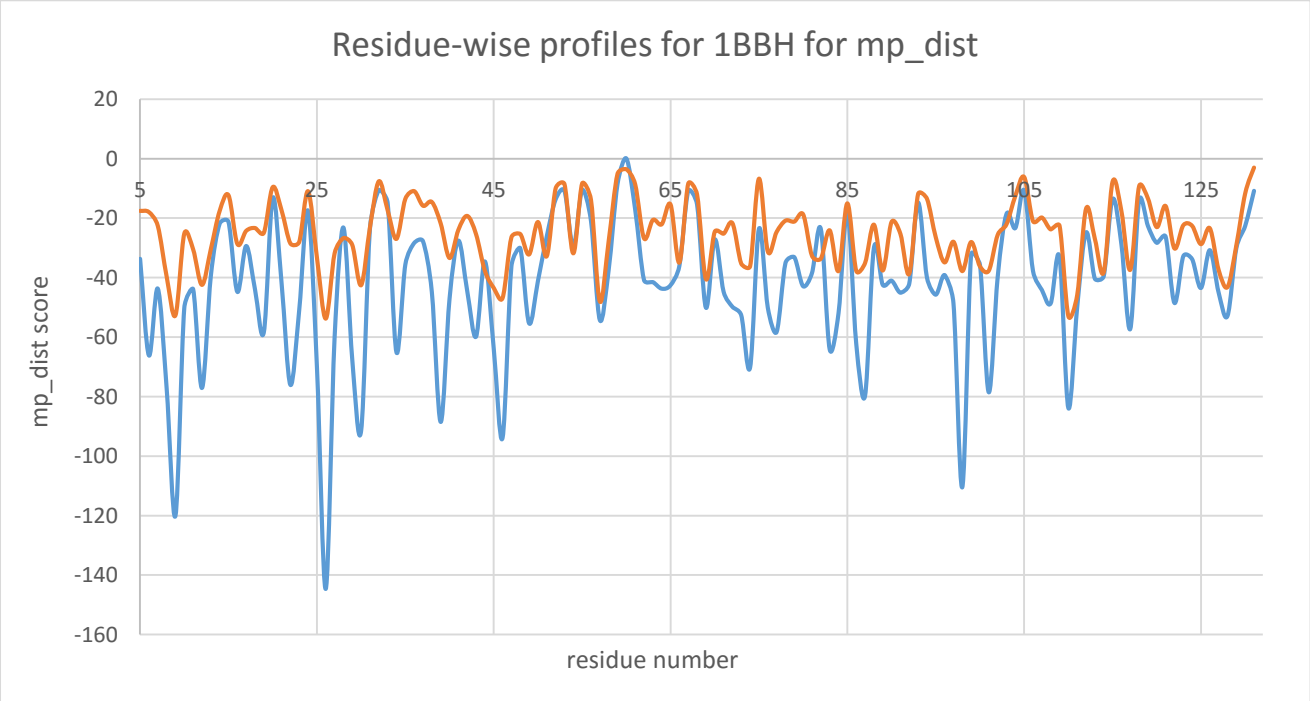

Fig 3.7: Residue wise profiling of 1BBH for mp_dist. The blue colour is for native structure and orange for the decoy structure

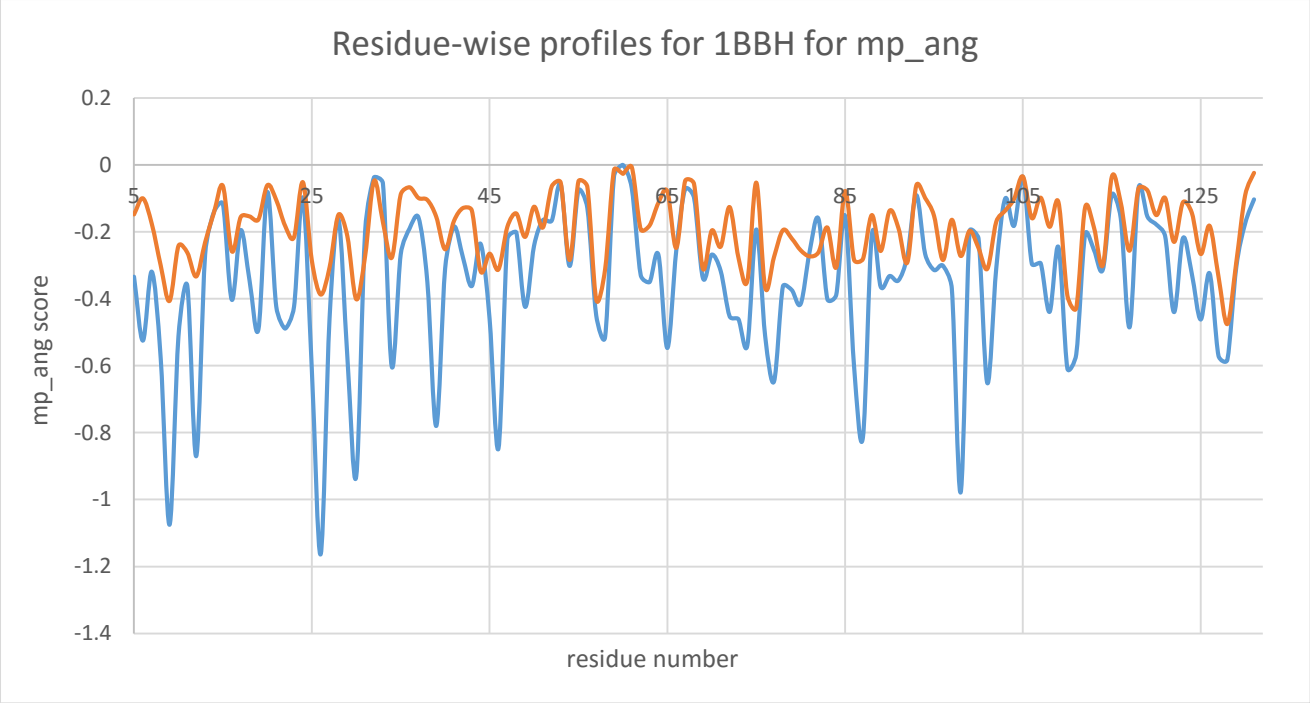

Fig 3.8: Residue wise profiling of 1BBH for mp_ang. The blue colour is for native structure and orange for the decoy structure

One region where we can clearly see that the decoy structure does not follow the pattern followed by the native structure (Fig 3.9) is at reside number 60. So we saw those regions(residue 59 to 64) using UCSF Chimera (Pettersen et al., 2004) and found that in the decoy structure it forms a loop while in the native it is part of an alpha-helix.(Fig 3.10). The 60th and 61st residue have been shown in Fig 3.11 to illustrate the difference in the packing.

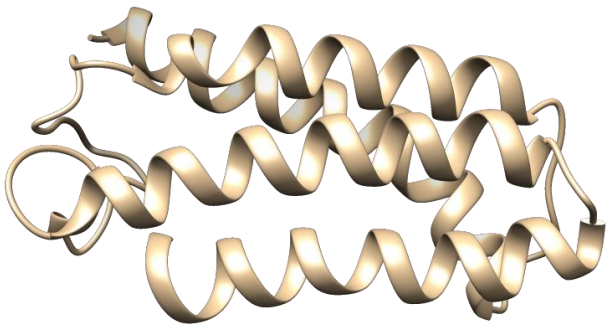

Fig 3.9: Native structure of 1BBH rendered using Chimera

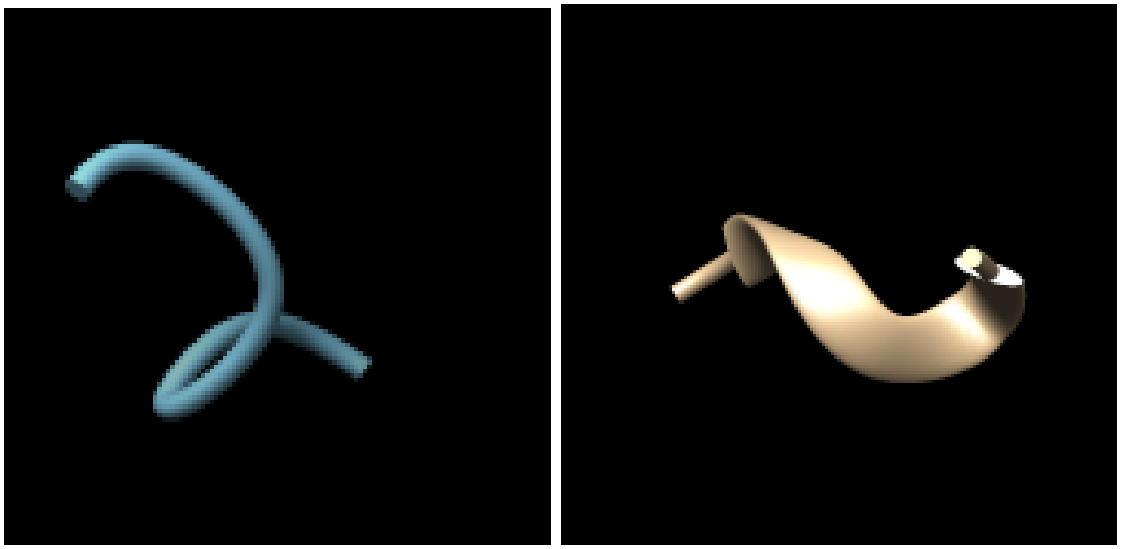

Fig 3.10: Native structure of 1BBH rendered using Chimera between residues 59 to 64. The blue one (part of a loop) is decoy structure and grey (starting of alpha-helix) is the native structure.

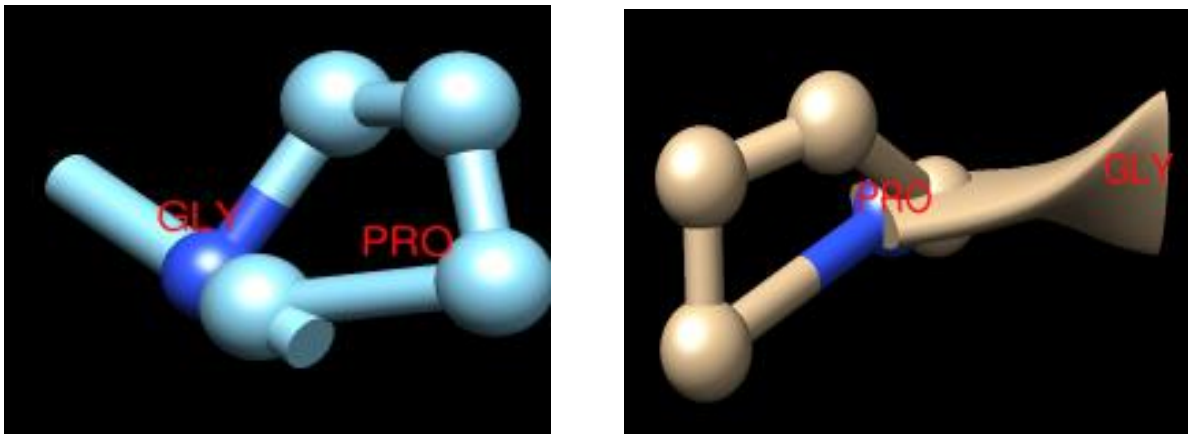

Fig 3.11: Native structure of 1BBH rendered using Chimera showing residue 60 and 61 with side chain of proline (61st residue). The blue one (part of a loop) is for the decoy structure and grey (starting of alpha-helix) is for the native structure.

If we want more resolution than we can do group-wise profiling of the scores instead of residue wise and identify the incorrectly packed regions in the protein.

Chapter 4

Discussion

4.1 Statistical Potential

The experimental methods for finding out the native state of the protein structure (X-ray and NMR spectroscopy etc.) are expensive and time-consuming. So, computational methods to determine the 3d-structure of the proteins are becoming very popular. The physics based computational method (ab initio modelling) for protein structure prediction try to build the three-dimensional structure from scratch without studying previously solved structures and require vast computational resources. The Comparative protein modeling methods use previously solved structures as templates and assume that the protein folds into a structure that is similar to already experimentally determined structures of another homologous model. Comparative modeling is comparatively faster and requires less resources and is extensively used for protein structure prediction. The structures obtained using comparative modeling needs to be validated for their accuracy. For that purpose we use potential functions.

Based on observations made on an experimental dataset, statistical potentials allow us to derive approximate functions which can be used to predict the energy of an unknown system. Statistical potentials are based on the Boltzmann principle states that frequently observed structural features correspond to the lowest energy states and are more stable. Statistical potentials can be based on physical principles or previously solved structures (knowledge based statistical potentials) or both (e.g. Rosetta). We construct a knowledge-based statistical potential because it is algorithmically and computationally much more feasible than the physics-based potentials. The statistical potential based on physical laws are heavily dependent on the accuracy of the structure of the proteins. Even slight errors in the Cartesian coordinates of the atoms leads to large deviations in the calculation of the approximate

energy of the model. Knowledge-based Statistical potentials are robust enough that slight errors do not affect the estimates of potentials by a very significant amount. We chose pairwise interactions as a reference state and most frequently observed states are energetically favoured than the less frequently ones and the protein will try to fold in such a way so as to maximize those states and minimize the energy.

Generally pairwise(contact based) statistical potentials are distance based and they imply the idea that if two units(residues/atoms/chemical group) within an protein are within a certain cut-off distance then they are said to form a neighboring pair or interacting with each other These interactions can be hydrophilic/hydrophobic, Van Der Waals interactions, electrostatic forces. We give a score to such interactions and energetically favourable interactions (by Boltzmann principle more frequently observed interactions) get high scores. Generally this interacting unit is a residue or atom. We constructed a knowledge-based, distance based pairwise potential with chemical groups as units. We observe the pairs of groups within a certain cut-off distance. The distance between the groups is defined as the distance between their centroids. These chemical groups are covalently bonded atoms within the residues and linear combination of these groups can form all the 20 residues. We defined an initial direction vector for these groups for differentiating between the relative orientations of these groups when they occur as pairs. That way we could differentiate between two pairs of interacting chemical groups at equal distances between their centroids but in different orientations w.r.t each other. The orientations (angle bins from 0 to 180 degrees) which are frequently observed were given higher scores. So, we constructed two potentials- distance based and distance-angle based potential. The distance-angle based potential is different than conventional contact potentials because it can also give us information of the interacting pairs of chemical groups.

4.2 The formulation of the potentials and choosing the best potentials

We formulated two types of potentials- mp_all and mp_diff. Each of these two types have 3 potentials: one distance based mp_dist and two angle based (mp_ang_1 and mp_ang_2). In mp_ang_1 we first took the negative log of the score at that distance

bin (to convert it into a free energy like formulation) and then further divided it into the 18 angle bins by taking the fractional contribution of each angle bin to construct mp_ang_1 potential. In mp_ang_2, we first fractionally divide the distance bin score and then take the negative log. This distance bin score is already very less (of the order of 0 or 1), dividing it into the angle bins further decreases it to the order of (-1 or -2). Taking the negative log of already very small numbers gives large values which does not correlate with the relative differences between the preferences of the various groups. This potential function makes zero correct predations, giving arbitrarily high positive values and hence we get a very random score for this function. So, we reject this statistical potential function.

The mp_diff potential excludes the pairs which are coming from the same residue and pairs which are covalently bonded while mp_all potentials consider all the pairs irrespective of whether they come from the same residue or they are covalently bonded or not. We see that all mp _all potentials work very poorly. The mp_dist_all predicts 10, mp_ang_1_all predicts 1 (mp_ang_2 has already been rejected so mp_ang_2_all is rejected by default) out of 20 for MOULDER set which is very poor success rate. Even, the lowest accurate potential after these (mp_ang_1_diff) predicts native structures correctly for 15 targets.

We reject all the mp_all potentials because they do not paint a correct picture of the group preferences. As it considers neighboring groups from the same residues and covalent bonded groups it does not weigh preferred scores based on the strength of the interactions(hydrophilic, hydrophobic or van der walls' etc.) In this potential groups within same residues always get high preferences for each other which highly skews up the statistical learning of the potential. We know that no covalent bond has length more than 6 Å (our cut-off distance) and two covalently bonded groups are always counted as pairs, getting very high preference scores. So, these potentials perform very poorly in rank ordering the native structure in the decoy sets. Hence, we reject all the mp_all potentials.

At last we are left with two potentials: the mp_dist_diff and mp_ang_1_difff potentials. We call mp_dist_diff as mp_dist universally which is our final distance

based potential and mp_ang_1_diff as mp_ang universally which is our final distance-angle based potential.

4.3 Limitations and Improvements

4.3.1 Observations

The mp_dist potential had an accuracy of 80 %(3rd best) and mp_ang potential (6^{th}) of 62% on the set of five multiple target decoys sets overall. The accuracy of mp_dist is only bettered by DOPE and DFIRE. In the MOULDER set mp_dist predicted 16 and mp_ang predicted 15 structures correctly. The failures were 1mup (ranked 99), 2pna (ranked 158), 2fbj (ranked 7) and 8i1b (ranked 51) for mp_dist and 1mup (ranked 30), 2pna (ranked 114), 2mta (ranked 2), 8i1b (ranked 7), 1c2r (ranked 2) for mp_ang. We see that 1mup, 2pna and 8i1b are common to both and rest of the individual failures are very close to rank 1(two are ranked 2 i.e. 2^{nd} best). Here, 2pna is a NMR structure and all the eight scoring fucntions fail for this PDB structure.

In the Decoys 'R' Us set, for three structures (1b0n-B, 1bba and 1fc2) all the remaining six scoring functions failed while our potentials were able to identify 1bon-B correctly and gave a good rank to 1bba as well. Overall all the failures ranked the native structures very close to rank 1 except for 4pti. We never see a failed case where mp_dist potential ranks the native structure very close to 1 and mp_ang giving it a very bad rank. Whenever a structure is ranked badly by one of the potential, it gets bad rank by the other one as well and vice-versa. The cases where mp_dist correctly predicts the native structure but mp_ang fails to predict the native structure have the native structure ranked very close to 1(the worst value is 13). We see only one case where mp_ang successfully predicts the native state but mp_dist fails. We also observe that if the native structure prediction fails for a target and they get a bad rank then given it's an x-ray structure the mp_ang rank is always better than mp_dist rank except for 4pti. Some of the observations are:

1) The structures with alpha helices are correctly identified.

2) If both potentials fail and mp_dist gives bad rank then mp_ang also give bad rank and rank for mp_dist is higher than that of mp_ang except for 4pti. These are the structures with many beta sheets.

3) If both potentials fail and mp_dist gives good rank then mp_ang also gives good rank and rank for mp_ang is higher than that of mp_dist. This is because of sparse data of training set and smaller size of test protein structures (sparse data of testing set).

4) If mp_dist gives rank 1 then mp_dist gives rank 2 or very close to 1. This is again because of sparse data of training set. If we increase the training set to include more PDB structures than mp_dist rank should eventually become 1.

5) Except for 2fbj there is no case where mp_ang correctly identifies a native structure but mp_dist doesn't. This strengthens the idea that the major reason of failure for these potentials is sparse data and if we increase our training set the accuracy will improve. The 2fbj protein also has large number of beta sheets. (Table 4.1)

PDB Code	***mp_dist***	***mp_ang***	***Probable Reasons for failure***
1c2r	1	2	sparse training set data
1mup	99	30	large number of beta sheets
2fbj	7	1	large number of beta sheets
2mta	1	2	sparse training set data
2pna	158	114	Fails all scores. NMR Structure
8i1b	51	7	large number of beta sheets
1ctf	1	2	sparse training set data
3icb	8	17	sparse training set data
4pti	145	308	Unknown, probably because small protein
4rxn	1	3	sparse training set data
1bg8-A	1	5	sparse training set data Set
1bl0	31	13	large number of beta sheets
1b0n-B	1	13	Small residue- sparse testing set data
1bba	7	20	Fails all scores. Small protein
1fc2	346	83	Fails all scores. Beta sheets and small protein
1igd	1	2	sparse training set data
2ovo	1	2	sparse training set data
4pti	1	7	sparse training set data

Table 4.1: List of failures for mp_dist and mp_ang with the ranking of the native structure for each target and probable reasons for the failure. The red-coloured are those cases which fail for both the potentials.

Excluding the NMR structures, the failures are because of three reasons: less data for the training set (sparse data for training set), small size of test proteins (sparse data for testing set) and large number of beta sheets.

4.3.2 Sparse Training Data set

As, both the statistical potentials (mp_ang and mp_dist) have same reference states (except the bin size), they should be able to predict structures correctly at the same time (mp_ang should perform better). Which is not the case. The mp_dist contains 6528 bins with same no. of potentials and mp_dist has 6528*18= 117504 potentials. We have taken 2579 PDB structures and sometimes we see that a particular distance bin/angle bin never occurs for a particular r_i-r_j pair or has a very low count. Ideally speaking, this means that if a particular interaction at a particular distance (for mp_dist) or orientation (for mp_ang) between those two groups is getting zero count or getting a very low count in the entire domain then those two group do not like to interact with each other at that distance or orientation, period. However, we have not covered the entire domain of the protein interaction angles and distances for the 136 combinations of the groups. They can be attractive or repulsive groups and to confirm that we need to increase our training dataset to increase the domain so that we cover more range of interactions as we might have missed all those pdb structures with structural features where that particular interaction is highly favoured.

There are millions of such interactions possible and an ideal knowledge based statistical potential should learn all the favoured interactions Our assumption is that most likely we capture the favourable protein interactions with 2579 already solved and formulate our potentials based on these observed interactions which is not 100% true (10% of distance bins have observed pair count =1).The problem of statistical potentials not correctly able to capture all the favourable states in the structures and giving incorrect scores (sub-optimal or bad scores) is the common occurring problem of sparse statistics. However, increasing our training set to include more pdb structures with various other commonly occurring structural features can improve the results especially for mp_ang as it has around 0.1 million different statistical potential bins. I believe that if we increase our training set then all the cases where mp_dist

ranks the structure as 1 and mp_ang ranks close to 1, will then get rank 1 from both the potentials. Also cases with ranks for both the potentials near to 1 will should also be solved.

4.3.3 Pseudo-Count

Sparse statistics is the most trivial and common source of errors in statistical potentials. Many times the bins have zero count-Two workarounds for this problem are: pseudo-counts (Samudrala & Moult, 1998) and a weighted background score scheme suggested by Sippl (Sippl, 1990). Pseudo-counts involve simply adding a count of one (unity) to every empty count to avoid a division by zero when calculating fractions or taking logarithms which could result in arbitrarily very high positive values (like the case of mp_ang_2) in some cases. The weighting scheme method assigns some fraction of the average of all interaction types to the bins in the case of an empty count. The pseudo-count is known as background score in our case. We assigned a pseudo-count of 1 to the distance bins and 1/18= 0.06 to angle bins with empty count. The limitation here is that roughly 10% of the distance bins have actual count equal to 1(before giving pseudo-count). The less frequently occurring groups (like r16 count-5402 and r10 count -5292 out of total 1.3 million groups) will have less number of pairs and we might fail to capture a highly preferred interaction for a particular distance/angle bin as we get a very low count for that bin (maybe even 1 or 0=pseudo-count of 1), giving a bad score to that bin. Now, when that interaction occurs in protein to be tested, it gets a bad score which should not be happening. Also, the interactions with zero count which are highly unfavourable still gets a score of 1 and actually favourable interactions for less frequent groups gets a score of 1(or very less count) and this creates problem while testing protein structures because favourable interactions get bad scores and unfavourable interactions get scores as good as the favoured interactions (at least 1). This problem is sparse statistics with the training set and can be corrected by increasing the PDB'S in the training set. The pseudo-count problem can be taken care of by weighting the pseudo-count with the interaction frequency of each group

4.3.4 Size of the tested protein (sparse testing data set)

If the protein structure being tested is small as in the case of 1b0n-B, 1bba and 1fc2 with 31, 36 and 43 residues respectively, our potential performs better than the other six potentials which gives very high ranks to the native structures and are not able to predict even one correct native structure. Our mp_dist ranks selects the correct native structure for 1b0n-B, gives 7^{th} rank for 1bba while mp_ang gives 13^{th} rank to 1b0n-B and 20^{th} rank to 1bba. Our potentials (meaning all eight potentials fail) to recognize 1fc2. Statistical potentials are generally less accurate for smaller proteins. This is because small proteins have smaller number of pairwise interactions, resulting in a large statistical error. This is an example of sparse statistics on the testing set.

As we cannot do anything about the protein to be tested, we can try to rectify this problem to some extent by quantifying our statistical potential scores by increasing the training set to include more PDB structures so that even these lesser numbers of interactions give us enough information to distinguish the PDB from decoys. The point to note here is that our potentials perform best on small proteins, even better than DOPE and DFIRE and if we sufficiently increase our training data set then we should be able to predict the native structures more accurately. The cases where both the potentials fail but have native rank very close to one should be solved by this as proteins with small residue length is one of the reasons for such cases.

4.3.5 The Beta Sheet Problem

We see that almost all the cases in which both the potentials give bad ranks to the native structure are the ones which have large number of beta sheets and no or only one alpha helix. The rank for the mp_ang potential is better than the rank for mp_dist. The failure of our potential to correctly identify the native structures containing large number of beta sheets tells us that our potentials failed to capture information about some interactions which only occur in the beta-sheets and hence when scoring those interactions, the potentials give them a bad score. The interesting thing to note here is that the mp_ang potential performed better than mp_dist potential (only case where

our distance-angle based potential outperformed just the distance-based potential). The mp_dist potential gives these interactions a bad score but mp_ang gives a lesser bad score i.e. if the interactions are considered unfavourable by our potentials, there are certain orientations (angles) which are less unfavourable.

This could be because of certain r_i-r_j pair interactions or because either the training data set which we took had no such interactions or they had very few beta sheets. To rectify this problem we can once again as always increase our training set to include more pdb structures, moreover adding those pdb structures which have these particular interactions make sure that our potentials learn these interactions. One other way can be to sub-divide the potentials based on the secondary structures (alpha helices, beta sheets, coils and loops) and motifs. That will capture all the information required including the beta-sheets interactions to improve our statistical potential.

4.3.6 Accuracy of the potentials considering the sparse statistics problem

In 'rank ordering the native structure method' to compute the accuracy of the statistical potential, rank of 1 is success and any rank higher than that is considered a failure irrespective of the rank. Even a rank of 2 for the native structure is considered a failure. However, in most of our failures the ranks are very close to 1(lesser than 20). For the failures of other potentials the ranks are very high. And now, we've established that failures with ranks close to 1 are most likely because of sparsed data for training set and these rankings can be improved if we increase our training pdb database. To test the accuracy of all the eight potentials on the Decoys 'R' Us set, we add the ranks of the native structure for 24 targets. This number gives us a clearer picture of the accuracy as it also takes into account the ranks for the failed cases and how close that rank is to 1.The lower the sum, more accurate is the potential. 100% performing potential will have a score of 24.

The sum is 1728, 4320, 1951, 4092, 1881, 1426, 575 and 499 for DFIRE, Rosetta, ModPipe-Pair, Modpipe-Surf, ModPipe-Comb, DOPE, mp_dist and mp_ang

respectively. Here, we see that mp_ang has the lowest score with mp_dist close second. So, if we are able to create a very accurate database of PDB's with all the possible interactions and train our potentials on that database which can quantify even the most uncommon protein interactions we should have a very accurate statistical potential function. Even now, based on the sum of native ranks our mp_dist and mp_ang potentials seem most accurate. The mp_ang outperforms mp_dist potential which theoretically it should as it encapsulates more information, given the limitations because of sparsed data are solved. (mp_ang potential is much more affected by sparse data than mp_dist).

4.3.7 Improvement

The statistical potentials can be improved by:

1) Increasing the training set to include PDB structures with varying structural features with all types of interactions (including beta sheets) and reduce pseudo-counts. This also takes care of sparse training dataset and sparse testing data set (somewhat).

2) Avoiding incomplete PDB's from the RCSB database i.e. excluding the PDB's with missing residue and atoms. At present, we built in the missing atoms and residues using MODELLER but it leads to clashes in the sidechains sometimes and also cause redundancy if many residues/atoms are missing.

3) Sub-dividing the statistical potentials to encapsulate the secondary structure information of the proteins

4) Reducing the number of angle bins/distance bins to reduce sparse data. However, this will decrease the resolution of the statistical potentials.

5) Avoiding double counts (N in r11 and C in first residue for r5) while counting these groups as it sometimes lead to clashes by reconstructing these groups or adding new groups to include such cases.

6) Instead of a universal pseudo-count, using group-pairwise based pseudo-count weighted by the frequency of occurrence of that particular interaction and numbers of those groups.

7) Instead of having a universal cut-off of 6 Å for each group, using group wise cut-off distances which are proportional to the size of that group. That way, large groups like r16 will have larger cut-offs and smaller groups like r8, r12 have smaller cut-off distances for the neighboring pair.
8) We can also incorporate depth preferences of the individual groups into the potential (similar to what we do for including secondary structures) which will give us more information about the interactions
9) We can go from pairwise to multibody potential based on cliques. This is a future work we plan to do.

4.4 Advantages of Group based potentials

We saw the correlation between Cα RMSD error of the decoys and our scoring functions is not that good compared to the other six scoring functions.
However, the assumption that the Cα RMSD is the best measure of the similarity between the decoy structure and the native structure is not necessarily true. The RMSD only takes Cα atoms into account without considering the sidechains, but our potential takes both the main chain and the side chain into account giving us an idea about the overall packing of the protein structure which is a more robust assessment of the decoy structures. Finding out the accurate packing of the protein sidechains is an important problem in protein structure prediction and our potentials give us important information about the same. Moreover, our potentials have very high resolution and they are able to differentiate between two interactions which differ by 0.1 Å distance and 10 degrees angle.

The important aspect of our statistical potential is that it can be used to improve and refine the incorrectly packed regions in the by plotting a residue/group profile of the structure. The residues or groups which have high scores (i.e. more energy) and are energetically unflavoured can be identified and we can improve these sub-optimally packed regions locally. This is perhaps the most important feature of our potential as it can be used to refine structures even locally. Our potential is also very useful in

scoring very small proteins compared to all the other six potentials. It is the most accurate scoring function and is successful for proteins as small as 31 residue length.

4.5 Future Work

The statistical potential will be further improved by adding the information of the preferred depth of the groups into the statistical potentials. We have already done analysis for the depth preferences of each of the groups and we can incorporate this information in the distance: angle based potential to include a third dimension of depth. We can divide the observed depth range for the groups into small depth bins, like we did for the distance and angles. This will give us a three-dimensional statistical potential which will be able to differentiate two interactions for the same groups at same distance and angle bins but different depth. When the groups are interacting in the depth bins which they prefer more, they'll get a higher score and vice-versa.

We will also try to develop a multibody potential. This is based on the pairwise interactions within a clique. A clique in our case is defined as a collection of neighboring groups such that the distance between each pair of the groups within the clique is less than the cut-off potential. We will be using n-body cliques to formulate the potential. Given n body clique for a neighborhood we should be able to predict the most preferred n+1th group and so on and so forth. We plan on going up to 5-body cliques. This potential should work very well for refining proteins locally because we have information about which of the cliques are statistically more observed in the dataset and which are unfavourable. If any collection of neighboring groups occur where non favourable cliques are being formed we can refine that region by changing the problematic group. This potential will also include orientation of the groups w.r.t each other and the depth of the groups. We can further add torsional angles for 4 and 5 body cliques.

A web-server for the public domain will also be developed where people can submit their protein structures to get a statistical potential score. We also plan to submit our potential for CASP (Critical Assessment of protein Structure Prediction) (Moult, Pedersen, Judson, & Fidelis, 1995).

References

Alberts, B., Bray, D., Hopkins, K., Johnson, A., Lewis, J., Raff, M., … And Walter, P. (2014). *Essential Cell Biology*. *Archives of Biochemistry and Biophysics* (Vol. 231). Retrieved from http://discovery.ucl.ac.uk/109974/

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2008). *Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter*. *Biochemistry and Molecular Biology Education* (Vol. 36). http://doi.org/10.1002/bmb.20192

Anfinsen, C. B. (1972). The formation and stabilization of protein structure. *The Biochemical Journal*, *128*(4), 737–49. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1173893&tool=pmcentrez&rendertype=abstract

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, *181*(96), 223–230. http://doi.org/10.1126/science.181.4096.223

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, *28*(1), 235–242. http://doi.org/10.1093/nar/28.1.235

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1166865&tool=pmcentrez&rendertype=abstract

Eramian, D., Shen, M., Devos, D., Melo, F., Sali, A., & Marti-Renom, M. a. (2006). A composite score for predicting errors in protein structure models. *Protein Science : A Publication of the Protein Society*, *15*, 1653–1666. http://doi.org/10.1110/ps.062095806

Eswar, N., Webb, B., Marti-Renom, M. a, Madhusudhan, M. S., Eramian, D., Shen, M.-Y., … Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.]*, *Chapter 2*, Unit 2.9. http://doi.org/10.1002/0471140864.ps0209s50

John, B., & Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research*, *31*(14), 3982–92. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165975&tool=pmcentrez&rendertype=abstract

Keasar, C., & Levitt, M. (2003). A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics. *Journal of Molecular Biology*, *329*(1), 159–174. http://doi.org/10.1016/S0022-2836(03)00323-1

Melo, F., Devos, D., Depiereux, E., & Feytmans, E. (1997). ANOLEA: a www server to assess protein structures. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *5*, 187–190.

Melo, F., Sánchez, R., & Sali, A. (2002). Statistical potentials for fold assessment. *Protein Science : A Publication of the Protein Society*, *11*(2), 430–48. http://doi.org/10.1002/pro.110430

Miyazawa, S., & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, *256*(3), 623–644. http://doi.org/10.1006/jmbi.1996.0114

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A Large-Scale Experiment to Assess Protein Structure Prediction Methods. *Proteins: Structure, Function, and Bioinformatics*, *23*(3), ii–iv. http://doi.org/10.1002/prot.340230303

Nitta, T. (2008). *Advances in IMAGING AND ELECTRON PHYSICS*. *Advances in Imaging and Electron Physics* (Vol. 152). http://doi.org/10.1016/S1076-5670(08)00604-6

Park, B., & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *Journal of Molecular Biology*, *258*(2), 367–392. http://doi.org/10.1006/jmbi.1996.0256

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. http://doi.org/10.1002/jcc.20084

Samudrala, R., & Levitt, M. (2000). Decoys “R” Us: a database of incorrect conformations to improve protein structure prediction. *Protein Science : A Publication of the Protein Society*, *9*(7), 1399–1401. http://doi.org/10.1110/ps.9.7.1399

Samudrala, R., & Moult, J. (1998). An all-atom distance-dependent conditional

probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, *275*(5), 895–916. http://doi.org/10.1006/jmbi.1997.1479

Shah, N., Sattar, A., Benanti, M., Hollander, S., & Cheuck, L. (2006). Magnetic resonance spectroscopy as an imaging tool for cancer: a review of the literature. *J Am Osteopath Assoc*, *106*(1), 23–27. http://doi.org/106/1/23 [pii]

Shen, M.-Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science : A Publication of the Protein Society*, *15*(11), 2507–24. http://doi.org/10.1110/ps.062416606

Simons, K. T., Bonneau, R., Ruczinski, I., & Baker, D. (1999). AB INITIO: PREDICTION REPORTS Ab Initio Protein Structure Prediction of CASP III Targets Using ROSETTA. *Proteins: Structure, Function, and Bioinformatics*, *37*(May), 171–176. http://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.0.CO;2-Z

Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, *268*(1), 209–225. http://doi.org/10.1006/jmbi.1997.0959

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*. http://doi.org/10.1016/S0022-2836(05)80269-4

Tan, K. P., Varadarajan, R., & Madhusudhan, M. S. (2011). DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Research*, *39*(SUPPL. 2), 1–7. http://doi.org/10.1093/nar/gkr356

Tanaka, S., & Scheraga, H. a. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, *9*(6), 945–50. http://doi.org/10.1021/ma60054a013

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, *18*(3), 342–348. http://doi.org/10.1016/j.sbi.2008.02.004

Zhou, H., & Skolnick, J. (2011). GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical Journal*, *101*(8), 2043–2052. http://doi.org/10.1016/j.bpj.2011.09.012

Zhou, H., & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science : A Publication of the Protein Society*, *11*(11), 2714–26. http://doi.org/10.1110/ps.0217002

Zhou, Y., & Linhananta, A. (2002). Thermodynamics of an All-Atom Off-Lattice Model of the Fragment B of Staphylococcal Protein A: Implication for the Origin of the Cooperativity of Protein Folding. *The Journal of Physical Chemistry B*, *106*(6), 1481–1485. http://doi.org/10.1021/jp013824r

Appendix A

Training Set PDB Codes

1A2PA	1A62A	1A73A	1ATZA	1AUOA	1BBHA	1BGCA	1BGFA	1BKRA	1BTKA	1C1KA	1C52A	1C7KA	1CCWA
1CCZA	1CHDA	1CMCA	1CQMA	1CV8A	1CXQA	1D02A	1D0QA	1D1QA	1D2SA	1D2VA	1D4OA	1D4TA	1DBWA
1DDWA	1DG6A	1DGWA	1DHNA	1DK8A	1DLFL	1DNLA	1DOWA	1DQGA	1DUSA	1DWKA	1DY5A	1DZ3A	1DZKA
1E1HB	1E29A	1E5KA	1E7LA	1EAQA	1EB6A	1ECAA	1EEXG	1EJ8A	1ELKA	1ELWA	1EUWA	1EW0A	1EW4A
1F1EA	1F1MA	1F2TA	1F2TB	1F32A	1F3UA	1F3UB	1F46A	1F7DA	1F86A	1FM0E	1FT5A	1G12A	1G1TA
1G2RA	1G3PA	1G57A	1G5TA	1G66A	1G8EA	1G8KB	1GGZA	1GM7A	1GMIA	1GMUA	1GMXA	1GNYA	1GO3E
1GO3F	1GP0A	1GS9A	1GTVA	1GU2A	1GV2A	1GWMA	1GY7A	1H03P	1H0HB	1H2CA	1H32B	1H4AX	1H4XA
1H6FA	1H6HA	1H8UA	1H97A	1H9MA	1HDKA	1HDOA	1HFES	1HH8A	1HQ1A	1HXIA	1HXNA	1HXRA	1HZTA
1I0RA	1I12A	1I2MA	1I4JA	1I4UA	1I58A	1I5GA	1IBYA	1ID0A	1IDPA	1IFRA	1IIBA	1IJBA	1IJYA
1IKTA	1IO0A	1IOOA	1IQ4A	1IQQA	1ISPA	1IT2A	1IUJA	1IX9A	1J0PA	1J24A	1J27A	1J2RA	1J34B
1J3AA	1J3WA	1J77A	1J83A	1J8BA	1J8RA	1J98A	1JATB	1JAYA	1JB3A	1JBEA	1JF3A	1JF8A	1JFXA
1JH6A	1JHGA	1JHJA	1JIDA	1JIWI	1JKEA	1JL1A	1JNIA	1JOSA	1JUVA	1JYAA	1JYHA	1K2XA	1K2XB
1K4NA	1K5NB	1K6KA	1K7JA	1KAFA	1KCQA	1KGDA	1KHIA	1KMTA	1KMVA	1KNGA	1KOEA	1KPTA	1KQ6A
1KQFC	1KQRA	1KR4A	1KT6A	1KTGA	1KUXA	1KXOA	1L3KA	1L6PA	1L6XA	1L8RA	1LB6A	1LF7A	1LKKA
1LM5A	1LMIA	1LN4A	1LO7A	1LQ9A	1LQVA	1LU4A	1LWBA	1LXJA	1LYQA	1M1FA	1M2DA	1M4IA	1M4JA
1M55A	1M70A	1M9ZA	1MB3A	1MBAA	1MC2A	1MF7A	1MG4A	1MIXA	1MJHA	1MK4A	1MKZA	1MN8A	1MQVA
1MR7A	1MTYG	1MWQA	1MY7A	1N08A	1N13B	1N62A	1N71A	1N8VA	1N9PA	1NEPA	1NF9A	1NG2A	1NG6A
1NJHA	1NKIA	1NLQA	1NNXA	1NQJA	1NQZA	1NRJA	1NRJB	1NRVA	1NS5A	1NTVA	1NU0A	1NULB	1NWWA
1NWZA	1NXMA	1NYCA	1NYKB	1NZ0A	1O6DA	1O7IA	1O8BA	1OA2A	1OA8A	1OB8A	1OBOA	1OCYA	1OD3A
1OD6A	1OH0A	1OH4A	1OI0A	1OI6A	1OOHA	1OQVA	1ORUA	1OSYA	1OU8A	1OUWA	1OW1A	1OW4A	1OXJA
1P0ZA	1P3CA	1P57A	1P5VB	1P6OA	1P90A	1PBJA	1PDOA	1PKHA	1PKOA	1PM4A	1PMHX	1PP0A	1PQHA
1PSRA	1PU6A	1PVMA	1PWBA	1PYOB	1PZ4A	1PZ7A	1Q1FA	1Q3OA	1Q4UA	1Q5ZA	1Q6OA	1Q7LA	1Q8DA
1Q9UA	1QAUA	1QDDA	1QF8A	1QFTA	1QHQA	1QJPA	1QKRA	1QNAA	1QV1A	1QVEA	1QW2A	1R0UA	1R29A
1R45A	1R62A	1R77A	1R8SE	1R9WA	1RFYA	1RG8A	1RGXA	1RKIA	1RKUA	1RLIA	1RO0A	1ROCA	1RTTA
1RUTX	1RW1A	1RXQA	1RYLA	1S3CA	1S5AA	1S5UA	1S7IA	1S99A	1S9UA	1SAUA	1SBXA	1SC3A	1SDIA
1SENA	1SH8A	1SHUX	1SJWA	1SJYA	1SLUA	1SZHA	1T1JA	1T2WA	1T3YA	1T6UA	1T82A	1T92A	1T9IA
1TC1A	1TD4A	1TFEA	1TJOA	1TP6A	1TQGA	1TR0A	1TS9A	1TT8A	1TU9A	1TUAA	1TVGA	1TW9A	1TX4A
1TXLA	1TY0A	1U14A	1U53A	1U5DA	1U5XA	1U7IA	1U9KA	1UB3A	1UCDA	1UEBA	1UFYA	1UI0A	1UKFA
1UKKA	1UNQA	1UPQA	1URRA	1USCA	1UUYA	1UV7A	1UXOA	1UXZA	1UZ3A	1UZKA	1V2BA	1V2XA	1V2ZA
1V30A	1V37A	1V4PA	1V70A	1V77A	1V8CA	1V8HA	1V96A	1V9YA	1VCAA	1VE4A	1VH5A	1VHTB	1VI0B
1VIAA	1VIMA	1VJ2A	1VKEA	1VKFA	1VKIA	1VKKA	1VL7A	1VMBA	1VMHA	1VP2A	1VP8A	1VPMA	1VQSA
1VR7A	1VR9A	1VSRA	1VYIA	1VYKA	1W0HA	1W0NA	1W1HA	1W2WA	1W2WB	1W4SA	1W9EA	1WBAA	1WBHA
1WBIA	1WC2A	1WC9A	1WDDS	1WEHA	1WHIA	1WJXA	1WKAA	1WKCA	1WKOA	1WKQA	1WL8A	1WLZA	1WMZA
1WN2A	1WNAA	1WNYA	1WOJA	1WOLA	1WOUA	1WPAA	1WPNA	1WPUA	1WRDA	1WS8B	1WUBA	1WVHA	1WVQA
1WWIA	1WWZA	1WZDA	1X0TA	1X3KA	1X6OA	1X6ZA	1X8QA	1X91A	1XAUA	1XBIA	1XE0A	1XE7A	1XFSA
1XG0C	1XI3A	1XJUA	1XKIA	1XKPB	1XKPC	1XKRA	1XLQA	1XMTA	1XODA	1XPPA	1XRKA	1XS0A	1XS1A

1XSVA	1XT5A	1XTEA	1XTTA	1XW3A	1Y02A	1Y0BA	1Y0HA	1Y0KA	1Y43B	1Y5HA	1Y63A	1Y7RA	1Y93A
1Y9IA	1Y9LA	1YACA	1YB3A	1YD9A	1YE8A	1YGTA	1YJ7A	1YKIA	1YLLA	1YLXA	1YM3A	1YNBA	1YO3A
1YPHC	1YPYA	1YQHA	1YRKA	1YSQA	1YSRA	1YTLA	1YU6C	1YUKA	1YUZA	1Z0JA	1Z0WA	1Z2UA	1Z4RA
1Z67A	1Z6MA	1Z9LA	1ZCEA	1ZD0A	1ZHSA	1ZHVA	1ZK5A	1ZKIA	1ZLDA	1ZMAA	1ZPSA	1ZUOA	2A0BA
2A15A	2A2RA	2A35A	2A4VA	2A4XA	2A5DB	2A9IA	2A9SA	2AALA	2ACFA	2AG4A	2AH5A	2AJ6A	2AJ7A
2ANXA	2AP3A	2AQ6A	2AQPA	2AR5A	2ARCA	2ASFA	2ASKA	2AWGA	2AXWA	2AYHA	2B06A	2B0AA	2B0VA
2B1YA	2B7KA	2B82A	2B8MA	2BBAA	2BBRA	2BCMA	2BDRA	2BFWA	2BJDA	2BJNA	2BK9A	2BKMA	2BL8A
2BMOB	2BNMA	2BRFA	2BZ1A	2BZVA	2C2IA	2C2QA	2C2UA	2C3VA	2C4JA	2C60A	2C6UA	2C71A	2C8SA
2C92A	2CARA	2CCMA	2CCVA	2CE0A	2CE2X	2CF7A	2CHCA	2CIOB	2CIUA	2CJ4A	2CJTA	2CKKA	2CO3A
2COVD	2CVEA	2CWRA	2CX1A	2CX7A	2CXHA	2CXYA	2CYJA	2CZQA	2D3YA	2D48A	2D4PA	2D59A	2D5MA
2DC4A	2DKOA	2DKOB	2DOKA	2DPFA	2DQLA	2DT4A	2DTCA	2DTJA	2DVKA	2DXAA	2DXQA	2DY0A	2E85A
2E8BA	2E8EA	2EBEA	2ECUA	2EGJA	2EGZA	2EH3A	2EHGA	2EHPA	2EJNA	2EJXA	2ENDA	2ENGA	2ERFA
2ET1A	2EW0A	2F01A	2F1FA	2F22A	2F23A	2F46A	2F5GA	2F62A	2F9HA	2FA1A	2FA5A	2FB6A	2FBNA
2FCJA	2FCKA	2FCLA	2FCOA	2FCWA	2FD4A	2FD5A	2FEXA	2FG1A	2FHPA	2FHZA	2FHZB	2FI1A	2FI9A
2FJ8A	2FKKA	2FLHA	2FLIA	2FOMB	2FP1A	2FPHX	2FPRB	2FQ3A	2FQ4A	2FR5A	2FRGP	2FSRA	2FSUA
2FSXA	2FTRA	2FUJA	2FULA	2FUPA	2FURA	2FVVA	2FYGA	2G1UA	2G2CA	2G2SB	2G3RA	2G5RA	2G64A
2G6YA	2G7SA	2G9WA	2GAXA	2GDMA	2GEYA	2GFFA	2GHTA	2GIBA	2GIYA	2GJ3A	2GJ8A	2GKGA	2GKPA
2GLZA	2GMQA	2GMYA	2GRRB	2GS5A	2GU3A	2GU9A	2GUDA	2GUIA	2GWMA	2GX5A	2GXGA	2GXQA	2GYQA
2GZ4A	2GZBA	2GZQA	2GZVA	2H1CA	2H1TA	2H2BA	2H3LA	2H5CA	2H88C	2H88D	2H8EA	2HA8A	2HALA
2HAZA	2HD9A	2HDOA	2HEWF	2HFNA	2HFTA	2HHGA	2HIYA	2HKVA	2HLRA	2HLYA	2HNGA	2HPLA	2HQSC
2HSJA	2HU9A	2HUHA	2HW2A	2HX0A	2HX5A	2HXSA	2HY5A	2HY5B	2HY5C	2HYTA	2HZFA	2HZQA	2I02A
2I5FA	2I5HA	2I6HA	2I74A	2I8DA	2I8TA	2I9WA	2IA1A	2IA7A	2IAYA	2IBDA	2IBLA	2IC2A	2ICGA
2IDLA	2IEQA	2IF6A	2IGIA	2IGPA	2IH3C	2IIAA	2IIHA	2IKBA	2IKKA	2ILKA	2IMFA	2IMJA	2IMLA
2IMSA	2IN0A	2INWA	2IP6A	2IQYA	2ISBA	2IT2A	2IT9A	2IU1A	2IU5A	2IVYA	2IWRA	2IYVA	2IZ6A
2J1AA	2J1VA	2J2JA	2J43A	2J6AA	2J6BA	2J73A	2J8KA	2J97A	2J9WA	2JCBA	2JDAA	2JDCA	2JE3A
2JEKA	2JJUA	2JK9A	2JLIA	2LISA	2LTNA	2MCMA	2MHRA	2NLVA	2NMLA	2NNUA	2NP5A	2NPTA	2NR7A
2NRKA	2NRRA	2NS9A	2NSAA	2NSZA	2NUHA	2NV0A	2NW2A	2NWFA	2NX4A	2NYIA	2NYUA	2O0QA	2O1QA
2O2XA	2O30A	2O3FA	2O6LA	2O6PA	2O70A	2O7AA	2O8QA	2O90A	2O99A	2OA2A	2OB5A	2OD4A	2OD5A
2OEBA	2OFCA	2OFKA	2OFZA	2OH1A	2OHWA	2OIKA	2OIXA	2OIZD	2OJ5A	2OKFA	2OKMA	2OLMA	2OMLA
2OMZB	2ONFA	2OOCA	2OOKA	2OPCA	2OPLA	2OQBA	2OQGA	2ORWA	2OS0A	2OU5A	2OU6A	2OV0A	2OVJA
2OX6A	2OX7A	2OXGB	2OYAA	2OYOA	2OZHA	2OZJA	2OZNA	2OZNB	2P0BA	2P0KA	2P0NA	2P0SA	2P12A
2P14A	2P1MA	2P25A	2P38A	2P39A	2P3HA	2P3PA	2P58C	2P65A	2P6WA	2P7OA	2P8GA	2P8IA	2P97A
2PA7A	2PAGA	2PFIA	2PIEA	2PLNA	2PLRA	2PMUA	2PN0A	2PN6A	2PNDA	2PQ7A	2PQVA	2PR7A	2PRVA
2PRXA	2PTTB	2PU3A	2PU9A	2PV2A	2PVBA	2PXRC	2PXXA	2PYQA	2Q03A	2Q0SA	2Q3TA	2Q3VA	2Q3WA
2Q3XA	2Q48A	2Q4MA	2Q5CA	2Q6MA	2Q9KA	2QEUA	2QF4A	2QFEA	2QGUA	2QH9A	2QHQA	2QIPA	2QIYA
2QJWA	2QJZA	2QKPA	2QL8A	2QLWA	2QMCB	2QMLA	2QMOA	2QNGA	2QNLA	2QNTA	2QR3A	2QS9A	2QSIA
2QSWA	2QSXA	2QT1A	2QTDA	2QUDA	2QUOA	2QVGA	2QVKA	2QWUA	2QZCA	2R01A	2R0XA	2R16A	2R1IB
2R25A	2R2CA	2R2YA	2R4IA	2R4QA	2R5OA	2R78A	2RA9A	2RASA	2RB8A	2RBDA	2RBGA	2RC3A	2RCIA
2RDCA	2RDGA	2RDMA	2RFFA	2RG8A	2RGQA	2RH3A	2RIQA	2RJ2A	2RK3A	2RK9A	2RKQA	2RL8A	2RLDA
2SAKA	2SICI	2SPCA	2TGIA	2TNFA	2UV4A	2UVOA	2UVPA	2V1MA	2V4XA	2V6KA	2V6VA	2V75A	2V76A
2V7FA	2V8FA	2V9VA	2VB1A	2VBUA	2VFKA	2VH3A	2VHKA	2VIFA	2VKJA	2VLGA	2VLQB	2VMHA	2VN6A
2VNGA	2VNLA	2VO9A	2VOKA	2VOOA	2VPAA	2VPTA	2VQ8A	2VQYA	2VRSA	2VSMB	2VXTI	2VXZA	2VY8A
2VZCA	2VZPA	2W0IA	2W15A	2W1JA	2W1RA	2W31A	2W3GA	2W3XA	2W43A	2W50A	2W6KA	2W72B	2W7AA
2W7ZA	2W9YA	2WAGA	2WAWA	2WB9A	2WCRA	2WCWA	2WDSA	2WFIA	2WFOA	2WFWA	2WH6A	2WH7A	2WJ5A
2WJ9A	2WJRA	2WKBA	2WLVA	2WM8A	2WNPF	2WNVC	2WQ4A	2WQFA	2WTGA	2WTPA	2WUHA	2WVFA	2WVIA
2WWXB	2WY3B	2WY4A	2WZ1A	2WZ9A	2WZOA	2X1QA	2X2SA	2X32A	2X3GA	2X46A	2X4JA	2X5CA	2X5NA
2X5PA	2X5YA	2XBLA	2XDPA	2XEVA	2XFGB	2XGUA	2XHFA	2XLKA	2XODA	2XOLA	2XOMA	2XOVA	2XPPA

2XPVA	2XRHA	2XSKA	2XTSB	2XTYA	2XU3A	2XVMA	2XVSA	2XW6A	2XWSA	2XXNA	2XZEA	2Y0OA	2Y28A
2Y44A	2Y6XA	2Y71A	2Y78A	2Y8GA	2Y8YA	2YBYA	2YD6A	2YF4A	2YFUA	2YG2A	2YH5A	2YH6A	2YJLA
2YK9A	2YKZA	2YOGA	2YVEA	2YWIA	2YWLA	2YXBA	2YZJA	2YZKA	2YZYA	2Z08A	2Z0TA	2Z0XA	2Z10A
2Z14A	2Z3HA	2Z51A	2Z5WA	2Z6OA	2Z84A	2Z98A	2ZCAA	2ZCMA	2ZDPA	2ZEXA	2ZFDB	2ZK9X	2ZNRA
2ZOUA	2ZQOA	2ZS0B	2ZS0D	2ZS1B	2ZVYA	2ZX2A	2ZYZB	3A07A	3A0YA	3A2ZA	3A35A	3A57A	3A6RA
3A8GA	3A8GB	3ACHA	3AG3D	3AG3E	3AG7A	3AGNA	3AGTA	3AIAA	3AJ4A	3AK8A	3AKSA	3ALUA	3AQ2A
3AWUB	3B08A	3B49A	3B5MA	3B64A	3B6EA	3B79A	3B7CA	3B7YA	3BA3A	3BAMA	3BB9A	3BCWA	3BCYA
3BDIA	3BDVA	3BEDA	3BEMA	3BFQG	3BHOA	3BHQA	3BHWA	3BJNA	3BKRA	3BLNA	3BLZA	3BMZA	3BO6A
3BODA	3BOEA	3BPKA	3BPVA	3BPZA	3BQXA	3BRCA	3BT5A	3BVFA	3BWUD	3BWUF	3BWVA	3BWYA	3BWZA
3BY8A	3BYQA	3C1QA	3C5KA	3C7MA	3C7XA	3C8LA	3C97A	3CANA	3CAOA	3CBNA	3CCGA	3CE7A	3CECA
3CG4A	3CG6A	3CH4B	3CHBD	3CHMA	3CI3A	3CI6A	3CJDA	3CJEA	3CK1A	3CLAA	3CM3A	3CNHA	3CP5A
3CP7A	3CRNA	3CRYA	3CT5A	3CT6A	3CVOA	3CWRA	3CYPB	3CZ1A	3CZXA	3D01A	3D06A	3D0FA	3D0JA
3D1BA	3D1PA	3D32A	3D33A	3D3BA	3D4EA	3D5PA	3D79A	3D7IA	3D7JA	3D9NA	3D9XA	3D9YA	3DA0A
3DA8A	3DALA	3DB7A	3DCZA	3DD7A	3DDCB	3DEWA	3DFGA	3DFRA	3DLCA	3DLUA	3DM8A	3DMCA	3DMNA
3DN7A	3DO8A	3DOUA	3DQPA	3DQYA	3DSBA	3DXYA	3DZAA	3E05A	3E0ZA	3E11A	3E23A	3E3UA	3E4GA
3E4VA	3E7HA	3E8MA	3E8OA	3E8TA	3E9FA	3E9TB	3E9VA	3EA6A	3EBTA	3EBYA	3EC6A	3ECHA	3EDOA
3EEAA	3EERA	3EF4A	3EF8A	3EFYA	3EGVB	3EINA	3EJFA	3EK3A	3ELKA	3EMIA	3EO6A	3EOIA	3ER7A
3ES4A	3ESLA	3ESMA	3EURA	3EXNA	3EXVA	3EYEA	3EZIA	3F0DA	3F0PA	3F13B	3F14A	3F1PB	3F2EA
3F2ZA	3F40A	3F43A	3F44A	3F4AA	3F4MA	3F5OA	3F5RA	3F6CA	3F6VA	3F7EA	3F7XA	3F8XB	3F95A
3F9SA	3F9XA	3FBUA	3FCNA	3FD3A	3FDEA	3FDQA	3FDXA	3FEAA	3FG8A	3FG9A	3FGEA	3FGRA	3FGVA
3FH1A	3FIAA	3FK8A	3FKAA	3FKCA	3FKEA	3FL2A	3FM2A	3FN5A	3FOJA	3FP5A	3FPNB	3FPRA	3FRQA
3FSAA	3FSOA	3FT1A	3FTTA	3FW2A	3FWZA	3FXAA	3FYBA	3FYMA	3FZEA	3G0KA	3G0MA	3G16A	3G46A
3G48A	3G9KF	3GA3A	3GA4A	3GAGA	3GAXA	3GBWA	3GBYA	3GE3E	3GFAA	3GGYA	3GHAA	3GIUA	3GK6A
3GKNA	3GLAA	3GMFA	3GMGA	3GMXA	3GNZP	3GP6A	3GPKA	3GQHA	3GRDA	3GUZA	3GWIA	3GWNA	3GX8A
3GXHA	3GYKA	3GZBA	3GZRA	3GZXB	3H05A	3H0NA	3H3HA	3H51A	3H5JA	3H6QA	3H79A	3H7HA	3H7HB
3H87A	3H8TA	3H93A	3HA9A	3HF5A	3HM4A	3HMCA	3HMSA	3HMZA	3HN5A	3HNXA	3HNYM	3HOIA	3HP4A
3HPCX	3HQXA	3HRAA	3HTNA	3HTSB	3HU5A	3HUHA	3HUPA	3HV2A	3HVWA	3HWUA	3HX8A	3HY3A	3HYNA
3HZ8A	3HZPA	3I2VA	3I57A	3I7MA	3I96A	3IBZA	3IC3A	3IDBB	3IDUA	3IE4A	3IEZA	3IHSA	3IHTA
3IISM	3IJMA	3IKBA	3IM6A	3IMKA	3IP0A	3IQ1A	3IQTA	3IR4A	3IRBA	3IT4A	3IT4B	3ITFA	3ITQA
3IU6A	3IUOA	3IV4A	3IVVA	3IWFA	3IX3A	3IXSA	3JRVA	3JSRA	3JSYA	3JU0A	3JUDA	3JUMA	3JYZA
3K05A	3K12A	3K1HA	3K2ZA	3K3CA	3K3VA	3K5JA	3K67A	3K6QA	3K6ZA	3K7IB	3K7PA	3KANA	3KB5A
3KBBA	3KBGA	3KBYA	3KD3A	3KE7A	3KEOA	3KEVA	3KF6A	3KF6B	3KFFA	3KGKA	3KH1A	3KKFA	3KKGA
3KM5A	3KMAA	3KMVA	3KORA	3KQ0A	3KSNA	3KTAA	3KTAB	3KU3B	3KUUA	3KUVA	3KVHA	3KWEA	3KYJA
3KYZA	3L00A	3L0FA	3L41A	3L42A	3L46A	3L4AA	3L4EA	3L4HA	3L4RA	3L51A	3L51B	3L6NA	3LAAA
3LATA	3LAXA	3LB2A	3LBFA	3LD7A	3LFJA	3LFKA	3LFRA	3LGBA	3LHCA	3LHEA	3LHNA	3LJTA	3LJWA
3LLOA	3LLUA	3LLVA	3LPWA	3LQBA	3LQNA	3LS0A	3LTJA	3LUCA	3LUIA	3LURA	3LW3A	3LWAA	3LWCA
3LWXA	3LX3A	3LXRF	3LYDA	3LYGA	3LYHA	3LZQA	3M0FA	3M1XA	3M3MA	3M7AA	3M7OA	3M86A	3M8JA
3M9LA	3M9QA	3M9ZA	3MAOA	3MC3A	3MCWA	3ME7A	3MEAA	3MF7A	3MH9A	3MKOA	3MMHA	3MN2A	3MNLB
3MNMA	3MPCA	3MQ2A	3MQQA	3MQZA	3MR0A	3MVCA	3MVSA	3MWZA	3MXNA	3MXNB	3MXOA	3MXZA	3N08A
3N0UA	3N10A	3N1EA	3N1FC	3N4JA	3N6YA	3N72A	3N79A	3NBCA	3NBMA	3NDUA	3NE0A	3NECA	3NEUA
3NFKA	3NFWA	3NJCA	3NJNA	3NKEA	3NL9A	3NOHA	3NPDA	3NR5A	3NRFA	3NRHA	3NRWA	3NRXA	3NSWA
3NTKA	3NUFA	3NZNA	3O0LA	3O22A	3O2HA	3O3XA	3O48A	3O7IA	3O94A	3OBLA	3OBQA	3OCMA	3OGHA
3OGNA	3OHEA	3OIOA	3OJ0A	3OKXA	3OMDA	3ON9A	3ONHA	3OOPA	3OOUA	3OQGA	3OQPA	3OR5A	3OSEA
3OSTA	3OSXA	3OU2A	3OUIA	3OWRA	3OXPA	3P0YA	3P1GA	3P2TA	3P48A	3P4HA	3P4LA	3P9AA	3P9VA
3PD7A	3PDDA	3PFYA	3PG6A	3PHXA	3PIWA	3PJPA	3PJVD	3PKZA	3PLWA	3PM2A	3PMCA	3PN3A	3PO1C
3PO8A	3POJA	3PP2A	3PP9A	3PQHA	3PR6A	3PROC	3PT8B	3PVEA	3PVHA	3PVIA	3PYWA	3Q20A	3Q46A
3Q49B	3Q4OA	3Q62A	3Q64A	3Q6AA	3Q6BA	3Q7RA	3QB8A	3QBMA	3QC7A	3QF2A	3QH6A	3QHBA	3QHPA

3QL9A	3QM9A	3QO0A	3QORA	3QP4A	3QPAA	3QQIA	3QR7A	3QRAA	3QRLA	3QS2A	3QU3A	3QZBA	3QZMA
3QZRA	3QZXA	3R0NA	3R2QA	3R5GA	3R5ZA	3R62A	3R6AA	3R72A	3R87A	3R8JA	3R90A	3R9FA	3RETA
3RF0A	3RFEA	3RG8A	3RHBA	3RJVA	3RKCA	3RLKA	3RLOA	3RLSA	3RNQA	3RNQB	3RO3A	3ROBA	3ROFA
3RPEA	3RQIA	3RRIA	3RT2A	3RT4B	3RTLA	3RVCA	3RWNA	3RX9A	3RY4A	3RYPA	3RZNA	3S0AA	3S4EA
3S57A	3S6EA	3S6FA	3S6MA	3S8IA	3S8KA	3S8SA	3S9XA	3SAOA	3SD2A	3SE5A	3SGWA	3SHGA	3SHOA
3SJ5A	3SK2A	3SK7A	3SNFA	3SO6A	3SOEA	3SOJA	3SONA	3SP4A	3SP7A	3SQFA	3SU6A	3SUKA	3T3LA
3T7ZA	3T8KA	3T92A	3T9GA	3TC2A	3TD3A	3TE8A	3TH2L	3TIOA	3TIPA	3TIWA	3TJ8A	3TOWA	3TRDA
3TS3A	3TU8A	3TUOA	3TVKA	3TVQA	3TYTA	3U1CA	3U1DA	3U1UA	3U2AA	3U2SC	3U3ZA	3U4VA	3U5SA
3U6GA	3U7ZA	3U97A	3U99A	3U9JA	3UB6A	3UE2A	3UFEA	3UI4A	3UIDA	3ULTA	3UPSA	3UPVA	3URRA
3USHA	3UTKA	3UWSA	3UX2A	3UZQB	3V43A	3V46A	3V4GA	3V4KA	3V7BA	3V7QA	3V9OA	3VBCA	3VBJA
3VCXA	3VE9A	3VFIA	3VG7A	3VGPA	3VI6A	3VORA	3VPPA	3VQJA	3VTUA	3VTXA	3VU9B	3VUBA	3VURA
3VV1A	3VVVA	3VWCA	3VWIA	3VYGA	3VYGE	3VZ9B	3W0TA	3W15B	3W42A	3W61A	3W77A	3WDCA	3WDNA
3WEBA	3WFWA	3WG3A	3WGXA	3WH1A	3WH2A	3WHJA	3WHTB	3WIIH	3WJDA	3WJTA	3WMVA	3WURA	3WUZA
3WW8A	3WZSA	3ZBDA	3ZFPA	3ZH5A	3ZHNA	3ZHOA	3ZJAA	3ZN4A	3ZQKA	3ZQUA	3ZRIA	3ZRXA	3ZSJA
3ZSUA	3ZT9A	3ZUCA	3ZUIA	3ZY7A	3ZYPA	3ZZYA	4A02A	4A2VA	4A3PA	4A3ZA	4A5NA	4A6HB	4A6QA
4A7UA	4ABLA	4AC7A	4AC7B	4ACIA	4ACJA	4ADZA	4AE4A	4AE7A	4AFFA	4AFMA	4AGHA	4AIWA	4AJYV
4AL0A	4ALZA	4ANNA	4ANOA	4AP9A	4AVDA	4AVSA	4AXOA	4AY0A	4B0HA	4B0MA	4B1MA	4B4CA	4B5OA
4B5QA	4B62A	4B8EA	4B8XA	4B9GA	4BFOA	4BGCA	4BJ0A	4BJIA	4BJTA	4BK7A	4BOQA	4BOUA	4BPYA
4BVXA	4BVXB	4BYZA	4BZPA	4C0NA	4C24A	4C6AA	4C6SA	4C9XA	4CA1A	4CAYA	4CBUG	4CDJA	4CE8A
4CFRA	4CGSA	4CHEA	4CHSA	4CICA	4CK4A	4CNLA	4CO8A	4COGA	4CRHA	4CRWB	4CSSA	4CT3A	4CUAA
4CV7A	4CVRA	4CYBA	4D0QA	4DAMA	4DB5A	4DB6A	4DDPA	4DGFA	4DKKA	4DM5A	4DMIA	4DN2A	4DNYA
4DOKA	4DOLA	4DOVA	4DQ9A	4DQJA	4DT5A	4DUIA	4DVCA	4DYQA	4DZOA	4DZZA	4E0AA	4E2UA	4E6FA
4E74A	4E9FA	4EA9A	4EAEA	4EBGA	4EBYA	4EDHA	4EETB	4EGUA	4EHSA	4EKFA	4EKXA	4EL6A	4EMTA
4ENFA	4EO0A	4EO7A	4EP4A	4EPZA	4EQ8A	4EQAC	4EQPA	4ERCA	4ES1A	4ESKA	4ESQA	4ESUA	4ETNA
4EVXA	4EW7A	4EZGA	4F1JA	4F2FA	4F54A	4F8LA	4F8YA	4FBSA	4FDBA	4FDVA	4FGQA	4FH0A	4FLBA
4FR9A	4FTFA	4FVGA	4G08A	4G0XA	4G29A	4G2EA	4G3VA	4G4KA	4G6IA	4G6TA	4G78A	4G79A	4G7XA
4G7XB	4G92C	4G9SA	4G9SB	4GA2A	4GAIA	4GB5A	4GCIA	4GEIA	4GELA	4GGFC	4GGJA	4GGTA	4GJ4A
4GLQA	4GLSE	4GMQA	4GNEA	4GOSA	4GQMA	4GQZA	4GS3A	4GT8A	4GUCA	4GV3A	4GWBA	4GYMA	4GZCA
4H08A	4H0CA	4H2DA	4H5SA	4H6CA	4H6JA	4H7WA	4H7YA	4H87B	4HATB	4HBQA	4HBZA	4HC5A	4HC9A
4HCIB	4HCJA	4HDEA	4HFQA	4HFSA	4HGNA	4HHVA	4HIKA	4HJIA	4HKHA	4HLBA	4HLSA	4HMWA	4HOJA
4HQSA	4HS2A	4HTUA	4HU2A	4HUAA	4HVOA	4HW4A	4HWMA	4HX4A	4HY4A	4HY9A	4HYLA	4HZ4A	4I1KA
4I4KA	4I4OA	4I4TE	4I66A	4I6XA	4IABA	4IAUA	4ICIA	4IEUA	4IGAA	4IGIA	4IGKA	4IGVA	4IGZA
4IHZA	4IKDA	4IL7A	4IM6A	4IMQA	4IN0A	4INKA	4INWA	4IPUA	4IRFA	4ITKA	4IUMA	4IVVA	4IWBA
4IWKA	4IX7A	4IYJA	4IYSA	4IZXA	4J1PA	4J2CA	4J2KA	4J32A	4J32B	4J39A	4J4ZA	4J5RA	4J7ZA
4J8SA	4J9YB	4JDUA	4JE1A	4JEAA	4JEMA	4JERA	4JF3A	4JF8A	4JG2A	4JGIA	4JGLA	4JHLA	4JHTA
4JIFA	4JIUA	4JJOA	4JL5A	4JP6A	4JQFA	4JV8B	4JVUA	4JW0A	4JWJA	4JXRA	4JYKA	4K2EA	4K45A
4K4OA	4K6WB	4K7BA	4K82A	4K8WA	4K90B	4KDWA	4KE2A	4KEFA	4KG3A	4KIAA	4KM6A	4KOPA	4KQDA
4KT3A	4KT3B	4KT6B	4KTWA	4KWAA	4KYQA	4KYUA	4L07A	4L0VA	4L57A	4L8AA	4L9EA	4L9NA	4LA2A
4LBAA	4LBHA	4LD1A	4LD6A	4LFLB	4LJHA	4LLDA	4LLDB	4LLEA	4LLMA	4LLQA	4LMIA	4LMYA	4LN7A
4LPIA	4LPLA	4LPQA	4LQ4A	4LQ6A	4LQBA	4LRJA	4LUAA	4LUKA	4LUPA	4LUQC	4LWRA	4LX3A	4LXOA
4LZLA	4M2PA	4M62S	4M7RA	4M91A	4MAIA	4MAXA	4MD5B	4MDXA	4MG3A	4MGQA	4MI7A	4MIWA	4MJDA
4MLMA	4MN9A	4MNNA	4MNOA	4MQ3A	4MTMA	4MTUA	4MU3A	4MURA	4MUVA	4MXTA	4MYPA	4MZ2A	4MZCA
4MZJA	4N0KA	4N0PA	4N1VA	4N2PA	4N5UA	4N6CA	4N6QA	4NB5A	4NBPA	4NBXA	4NF1A	4NI6A	4NKPA
4NN2A	4NO4A	4NOAA	4NOBA	4NOHA	4NPFX	4NPNA	4NT1A	4NTKA	4NUHA	4NV4A	4NXIA	4NXYA	4NYHA
4NYQA	4NZUL	4O06A	4O0AA	4O1RA	4O65A	4O6GA	4O6UA	4O7JA	4O7KA	4OA3A	4ODKA	4OE8B	4OE8C
4OF6A	4OFKA	4OKEA	4ONRA	4OO4A	4OQ9A	4OSNA	4OUNA	4OUSA	4OX6A	4OXXA	4OY6A	4OY7A	4OYDB
4P09A	4P0TA	4P0ZA	4P3FA	4P3HA	4P5EA	4P5MA	4P82A	4P9IA	4PBDA	4PH8A	4PHJA	4PJ2A	4PJ2C

4PMQA	4PO5A	4PQQA	4PS6A	4PSFA	4PSYA	4PT1A	4PTBA	4PUXA	4PWQA	4PZ3A	4Q0YA	4Q14A	4Q29A
4Q2SA	4Q2UB	4Q53A	4Q5WA	4Q7EA	4Q7OA	4Q9WA	4QAGA	4QASA	4QB0A	4QBBA	4QC6A	4QDNA	4QFTA
4QGPA	4QKDA	4QM6A	4QMAA	4QN8A	4QNDA	4QP5A	4QPWA	4QRHA	4QRLA	4QSPA	4QT6A	4QU6A	4QUSA
4QW0A	4QX5A	4QXLA	4QYTA	4R03A	4R1JA	4R4KA	4R6RC	4R81A	4RBRA	4RCJA	4RI6A	4RLEA	4RMBA
4RN7A	4RO3A	4RS2A	4RYRA	4S1PA	4S2XA	4TJVA	4TKBA	4TKZA	4TPSA	4TXRA	4TYZA	4U5RA	4U8FA
4U9BA	4U9VB	4UC1A	4UDSA	4UE8A	4UMGA	4UNUA	4UONA	4UOSA	4UQWA	4UQZB	4UUUA	4UXUA	4UZ3A
4V0KA	4V2YA	4V4M0	4W6YA	4W78A	4W78B	4W79A	4WANA	4WEEA	4WH9A	4WHIA	4WHSA	4WILA	4WKSA
4WQKA	4WSFA	4WUIA	4WWFA	5NLLA	7FD1A								