Prediction of protein stability based on structural motifs in naturally occurring protein structures

Masters Thesis

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

by
Swastik Mishra
Reg: 20141051

Supervisor: Dr M. S. Madhusudhan
Department of Biology, IISER Pune

Indian Institute of Science Education and Research, Pune

Abstract

Proteins are important building blocks of life. Proteins play a vital role by performing a wide variety of functions inside the cell. The structure of a protein is an important determinant of its function, and is largely dependent on its amino acid sequence. Therefore, structure prediction from the sequence can help us design novel proteins that may be useful in medicine (e.g. therapeutic proteins) as well as in industry (e.g. antibodies with lower aggregation propensity). Prediction of protein structures from sequence is a major challenge and methods for modelling protein structures require a good structure evaluation criteria both for evaluating initial models as well as for refining them further.

In this study, we discuss the development of a novel protein structure evaluation method that evaluates local regions in structures by comparing them to known regions in the Protein Data Bank (PDB). It then calculates how well represented in the PDB, is the amino acid environment of the region being evaluated, and the conformation of its atoms in 3D. We have demonstrated here that the method may be used to differentiate between the local regions from obsolete structures in the PDB, and their refined versions, with a high level of confidence. We also compared proteins from thermophilic and mesophilic organisms and could successfully differentiate between them approximately 70% of the time. We noted a significant correlation between our evaluation of the protein structures and their melting temperatures. Since the method directly compares against known native structures and evaluates local regions, it may be used for identifying regions that need to be targetted first for structure refinement.

Certificate

This is to certify that this dissertation entitled "Prediction of protein stability based on structural motifs in naturally occurring protein structures" towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by "Swastik Mishra at IISER Pune" under the supervision of "Dr M. S. Madhusudhan, Associate Professor, Department of Biology" during the academic year 2018-2019

Swastik Mishra

Student

Swastik Mishra
BS-MS 20141051
IISER Pune

Supervisor

Dr M. S. Madhusudhan
Associate Professor
IISER Pune

Declaration

I hereby declare that the matter embodied in the report entitled “Prediction of protein stability based on structural motifs in naturally occurring protein structures” are the results of the work carried out by me at the Department of Biology, Indian Institute of Science Education and Research, Pune, under the supervision of Dr M. S. Madhusudhan and the same has not been submitted elsewhere for any other degree.

Swastik Mishra

Student

Swastik Mishra
BS-MS 20141051
IISER Pune

Supervisor

Dr M. S. Madhusudhan
Associate Professor
IISER Pune

Acknowledgements

I would like to express my profound gratitude to my supervisor Dr. M. S. Madhusudhan for being a such a terrific mentor, and a constant source of motivation and inspiration. The brainstorming sessions in his office shall forever be among my fondest memories. It'd be a crime to not recognise how important is a good lab environment for your scientific endeavours. In my case, both my supervisor and my colleagues in the lab have been really supportive, so a big thank you to all of them. I would also like to thank Dr. Chetan Gadgil for being a member of the thesis advisory committee for this project, and for providing valuable inputs.

Over the years, several people have had an impact on my scientific temperament, and therefore are inseparable from my work. In light of this, I would like to thank Dr. Richa Rikhy, Dr. Harinath Chakrapani, and Dr. Bhas Bapat.

I shall forever be in debt to my parents who take credit for my existence and for building me into what I am today. I should also thank my friends Vishrut, Danish, Reema, Harsha, Farhan, Chinmay, Upasana, Smita, and many more who are understanding enough to know that I cannot name them all in here. Developing creativity for your work requires one to be a bit crazy – to engage in seemingly absurd activities elsewhere in life, and my friends have been instrumental in this regard. Neil Gaiman also takes some credit for the same. I thank all of them for being there through the good and bad times of this past year. This work would not have been the same without them.

Contents

1 Background 1

2 Introduction and Objectives 3

3 Methods 5
- 3.1 Chemical Groups Instead of Atoms 5
- 3.2 Datasets 6
  - 3.2.1 Identification of Native Models in a Decoy Set 6
  - 3.2.2 Prediction of Regions of Refinement 7
  - 3.2.3 Correlation with Thermal Stability 8
- 3.3 Scoring Methods 10
  - 3.3.1 Scoring stars in terms of RMSD of best-match in the PDB 10
  - 3.3.2 Scoring stars in terms of distribution of similar geometries (CASPER) 12
  - 3.3.3 Calculation of Chemical Group Temperature Factors 14
  - 3.3.4 Scoring parameters 14
- 3.4 Conflicting sequences between predecessor and successor structures . . . 15
- 3.5 Different residue environment upon refinement 15
  - 3.5.1 Consideration of homologous structures in the PDB 16
  - 3.5.2 Contact Order 16

4 Results 17
- 4.1 Prediction of Regions of Refinement 17
- 4.2 Correlation with Thermal Fluctuations 22
  - 4.2.1 Differentiating thermophilic and mesophilic homologs 22
  - 4.2.2 Correlation with Tm values 23
  - 4.2.3 Correlation with difference between Tm 29
  - 4.2.4 Correlation with Temperature Factors 29

5 Discussion and Further Work 31
- 5.1 Summary 31
  - 5.1.1 Obsolete PDB refinement study 31
  - 5.1.2 Correlation with Thermal Fluctuations 31
- 5.2 The approximation of additivity 32
- 5.3 Topology independence 33
- 5.4 Structure refinement 34
- 5.5 The CASPER metrics 34
- 5.6 Knowledge based potentials 35
- 5.7 Related work 35
  - 5.7.1 Charge dynamics dependent force field development 35
  - 5.7.2 PackPred mutation predictor 36

Appendices 40

A Method and Framework Developed Earlier 40

B Identification of Native Model In A Decoy Set 43

C Cleaning up and culling of obsolete PDB records 44
C.1 Completely obsolete records 44
C.2 Outdated records and multiple successors 44
C.3 Culling parameters 44

D Tm Correlation 44

List of Figures

1 Packing of atoms in a typical protein 1

2 The widening of the gap between sequence and structure databases, and the saturation of unique folds 2

3 stars and centre-matching superimposition 3

4 Illustration of Chemical Groups 5

5 Correlation between logarithm of folding rate and contact order 17

6 Star-wise CASPER query-star scores for obsolete vs successor structures 18

7 Star-wise CASPER badness scores for obsolete vs successor structures . . 19

8 Star-wise CASPER one-sided badness scores for obsolete vs successor structures 20

9 Outlier structures with CASPER Badness 21

10 Percentage of refined structures identified 21

11 Rank ordered scores for thermophilic vs mesophilic proteins; ProTherm$_{lt60plus10}$ set 24

12 Rank ordered scores for thermophilic vs mesophilic proteins; ProTherm$_{lt60plus10}$ set 25

13 Rank ordered scores for thermophilic vs mesophilic proteins from Kumar's set 26

14 Correlation between T_m difference and CASPER badness 27

15 Correlation between T_m and CASPER scores 27

16 CASPER Query-star score versus T_m 27

17 Correlation between T_m difference and CASPER scores 28

18 Example of a thermophilic protein (orange) and mesophilic homolog (blue); Frequency of stars versus Query-star score 33

19 Native vs decoys chemical group wise score profiles 42

List of Tables

1 List of amino acids and their constituent chemical groups 7
2 Szilágy's set [Szilágyi and Závodszky, 2000] 10
3 ProTherm dataset 11
4 Culling parameters for dataset used for B-factor study 12
5 Comparison of number of higher scoring structures of successor versus predecessor structures 22
6 Mean scores comparison of thermophilic vs mesophilic structures 23
7 Correlation between T_m and CASPER scores; mesophilic proteins 28
8 Correlation between T_m and CASPER scores; themophilic proteins 29
9 Correlation between T_m difference and CASPER scores; mesophilic proteins 29
10 SCC of chemical group wise scores vs chemical group wise B-factors . . . 30
11 Native model rankings out of decoy set size with 300 decoys. Penalty P=5 Å; Clique size N=8; RMSD as score metric 41
12 Native model ranking in Moulder decoy set 43
13 Correlation between T_m and CASPER scores 45

1 Background

Protein molecules are essential parts of organisms and participate in virtually every process within a biological cell. The three dimensional structure of a protein dictates the biological function, and the structure is determined largely by the protein sequence [Anfinsen, 1972, Tanaka and Scheraga, 1976]. Hence, knowing the structure of a protein helps us to better understand how the protein works. It can provide control over how to affect it or modify it [Worth et al., 2011, Pack and Yoo, 2004]. For example, *stable* site-directed mutations in a known protein structure can help us change the function of the protein completely, or tweak the original function to our needs [Nowlin et al., 1988]. Such protein engineering can have far reaching consequences, which include finding possible cures to Alzheimer's, Prion diseases, etc [Boyle, 2008].

To solve a protein structure or to refine an existing structure, one may use either experimental or computational methods. One may contend that solving protein structures experimentally gives you a more accurate model of the structure. However, given the rate at which we are discovering protein sequences versus the rate at which structures are being deposited in the Protein Data Bank(PDB) [Schwede, 2013, Berman Helen M, 2000] (Figure 2a), plus the time, effort and cost requirements of experimental methods of solving protein structures; finding alternative and quick computational ways of predicting protein structure has become important.

4.7Å
2.3Å
2.5Å
3.4Å
ILE 2.A
VAL 3.A
r1
VAL 3.A
r8
r12
r8
r1
r8
ILE 2.A
r2
r12
r8

(a)

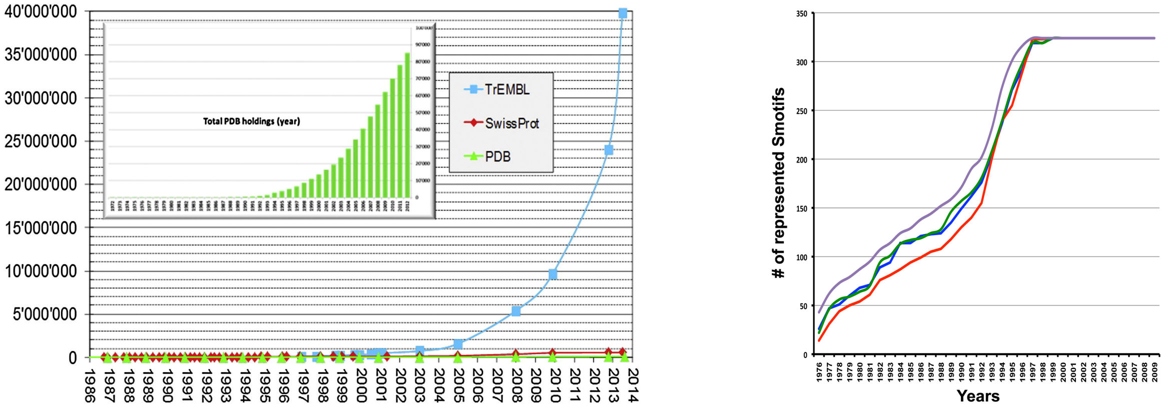
(b)

Figure 1: Packing of atoms in a typical protein; PDB ID: 2mta
(a) Illustration of a protein structure and the local packing of atoms. (b) Illustration of packing in terms of chemical groups(right) versus atoms(left) for two adjacent residues from a protein. The centroid of the atoms in a chemical group defines its position.

Computational methods involve *ab initio* modelling [Perez et al., 2016, Shaw et al., 2009] or the use of template based (homology) modelling (which are based on previously known protein structures) [Webb and Sali, 2016]. These methods require an *evaluation* method [Feig, 2017, Shen and Sali, 2006, Zhou and Zhou, 2002] to know whether the *in silico* built model, is accurate enough or not. The evaluation method is essentially a stand-in for a free energy function under the solution conditions usually encountered in cells, a minima of which is what corresponds to the native structure of the folded protein.

Consequently, given an efficient evaluation scheme, one can reverse-engineer to build the protein structure itself [Webb and Sali, 2016, Shen and Sali, 2006]. Also, if one can sample the local packing of residues in the protein (Figure 1a), one assumes that this packing should correspond to the free energy minimum that the atoms in those residues could attain, when the whole protein has reached a global free energy minimum (although it may not reach the global minimum [Chen and Kihara, 2011]). Studies have shown that the number of unique folds being discovered in the PDB has saturated over time

[Fernandez-Fuentes et al., 2010] (see Figure 2b). This lays the foundation for this study, where the packing of residues in naturally occurring protein structures has been sampled to evaluate the deviation of a structural model from known natural ones.



(a)

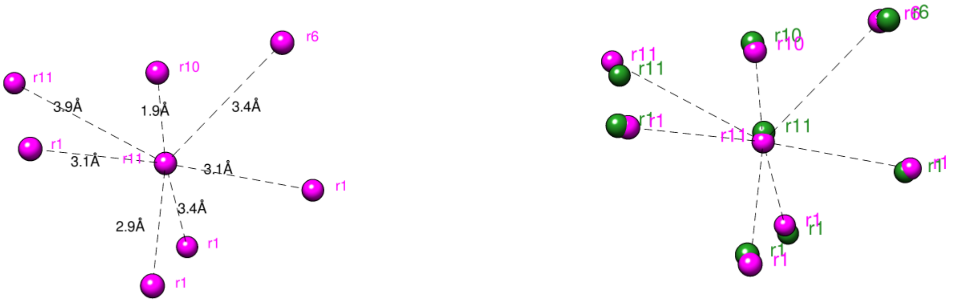
(b)

Figure 2: The widening of the gap between sequence and structure databases, and the saturation of unique folds.
(a) Number of protein sequences in TrEMBL or SwissProt and number of structures in PDB, over the years [Schwede, 2013]. (b) Saturation of number of unique folds discovered over the years, where folds are characterised as *Smotifs* which are super-secondary structures. [Fernandez-Fuentes et al., 2010]

2 Introduction and Objectives

In this study, we have departed from the use of conventional knowledge-based methods. We score protein structures based on the structural similarity of their *three-dimensional motifs*, using known protein structures in the PDB. Further, instead of amino-acid residue packing, we are looking at the packing of *chemical groups*. A chemical group is a group of atoms with a certain arrangement in three-dimensional space, such that it is one of the 16 different arrangements possible as defined along the lines of previous work done by Akash Bahai. (See Appendix 3.1 for more details about chemical groups. Figure 1b for an illustration) One can therefore use just these 16 chemical groups to define the structure and orientation of any of the 20 naturally occurring amino acids. The usage of a chemical group allows us to look at the packing of parts of a protein with a higher resolution than that of residues. However, it is at a resolution higher than that of atom-wise definition of the protein structure, wherein many of the atoms are covalently, strongly bound, to not be independent enough in their interactions with neighbouring atoms.

We define a three-dimensional motif called a **star**. A star is a representation of how chemical groups are packed in a protein. More importantly, the definition that we use is a strategy to sample what we assume are patterns of the local packing of atoms, repeating across naturally occurring proteins.



(a)

(b)

Figure 3: stars and centre-matching superimposition
(a) An illustration of an 8-body star. The central chemical group is of type/identity r11, and the rest are its nearest neighbours (b) An illustration of superimposition of two star-stars of very similar geometry, and the mapping between different chemical groups.

Let S_n be the set of all possible sets of n chemical groups in a protein structure. A **star** is defined as follows: Given the i^{th} chemical group $A_i \in S_n$, pick first n-1 nearest-neighbours $A_j, (j = 1, 2...n-1)$ such that Euclidean distance $D[A_i, A_j] < d_{thr}$, where d_{thr} is some known optimal distance cut-off[1]. See fig. 3a for an illustration of a typical star, and 3b for an example of a superimposition. Note the mapping of the central chemical groups of the stars to each other. Since there are two r11 chemical groups in that star, but one was mapped at the centre, the number of permutations is halved compared to the case where the centres are not mapped to each other.

Now, given any *query* star (defined in terms of chemical groups), we superimpose it on the known stars (*target* stars) of similar composition of chemical groups from the PDB

[1] this is important so that chemical groups from too far away in space are not picked up as part of the star since that would mean that chemical groups with negligible interaction with each other were considered to be part of the same residue environment

database, to find out the root mean square deviation (RMSD) of the groups from each other.

The method for parsing to chemical group format, finding stars, and superimposing the stars on stars from PDB database, was developed earlier during semester projects done by the author. See Appendix A to read about what has been done earlier.

The objectives of this project are broadly three fold:

1. **Prediction of poorly packed regions in proteins:** These regions presumably require structural refinement. We expect native-like stars to be able to find good (ie. low RMSD) matches in the PDB (which has stars from native structure models), and therefore, we should be able to identify these stars.

2. **Correlating net-score of a structure with melting temperature:** The reasoning made in the previous point can be extended to this, since a more flexible structure with lower melting point should find lower RMSD matches, and therefore worse scores. The scoring scheme is discussed in the next section.

3. **Correlating scores of the stars with B-factor values:** The packing of side chains of amino acids seems to have a bearing on thermal stability [Meruelo et al., 2012]. The flexibility of residues is correlated to B-factors/temperature-factors in a region. But a more flexible star is expected to have more distant variants (in terms of geometry) in the PDB, and worse scores by our scheme. There may be a way to find a proxy to B-factor values which will allow us to compare B-factors across structures, which is currently not possible.

Note that in this text, whenever *native model* of a protein structure is mentioned, it refers to the structural model that was obtained by fitting the protein sequence to experimental data (X-ray crystallography data, unless noted otherwise; structure obtained as PDB file from RCSB-PDB, unless noted otherwise). The more *native-like* a structure is, the lower is its RMSD from the native structure.

3 Methods

3.1 Chemical Groups Instead of Atoms

Based on previous work by Akash Bahai

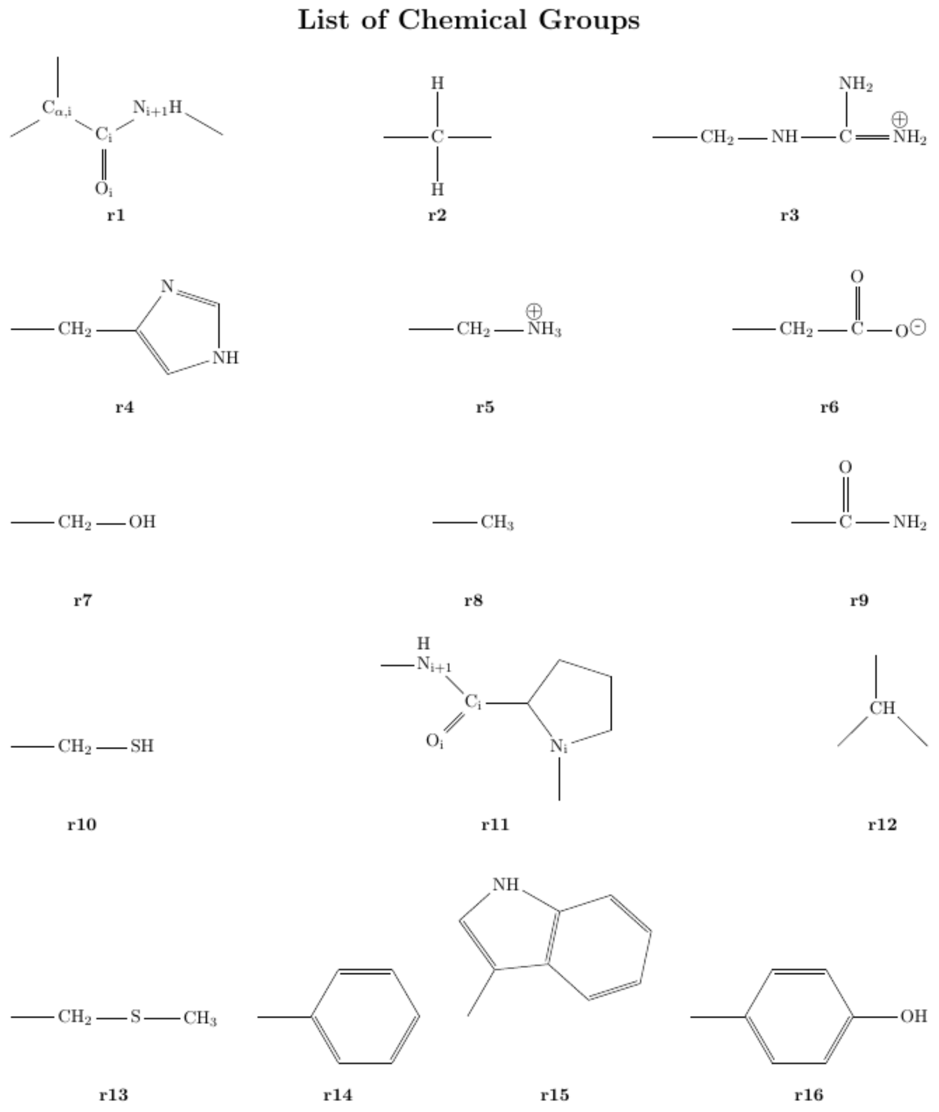


Figure 4: The 16 Chemical Groups

Sixteen *chemical groups* were defined, whose combinations can form all the 20 naturally occurring amino acids. The hydrogen atoms in the residues have not been included in these groups, since PDB data (mostly X-ray data) is used, without information of the light atoms such as that of hydrogens. These definitions are based on previous work done by Akash Bahai as part of his Masters Thesis project [2]. See Appendix A for a discussion on the implementation of chemical groups, stars, as well as scoring methodology.

The basis for differentiating between different chemical groups is as follows:

[2] Link to the thesis at IISER Pune library: http://idl.iiserpune.ac.in:8080/jspui/handle/123456789/570

1. Main chain atoms correspond to chemical group *r1*, and each r1 consists of CA_i, C_i, O_i, N_{i+1} for the i^{th} residue being parsed.

2. For side chain chemical groups,

   (a) should be able to move with a large range of motion w.r.t each other, and therefore

   (b) shouldn't be defined in a way such that there are resonance/mesomeric effects *across* chemical groups

   (c) the partial charges within a certain chemical group should be approximately an integer value. This criteria is added not for the structural studies described here, but because of a separate project in the research group. In brief, we intend on building a force-field using known geometry and dynamics of stars, with the dynamics of partial charges of atoms taken into account. See Section 5.7 for further discussion.

Chemical groups are also differentiated based on the atom they are covalently bonded to e.g. primary, secondary and tertiary Carbon atoms are treated differently as *r8, r2*, and *r12*. Proline is treated as a special case, since the Nitrogen atom of the amino acid under consideration is part of the proline ring. We don't parse the r1 group normally. Instead, we skip the r1 for Proline, and parse the whole Proline residue into only a single r11 group. Figure 4 illustrates how the groups have been named.

The distance between the groups is defined as the distance between their centroids. These chemical groups are sets of covalently bonded atoms within the residues and a combination of these groups can form all the 20 residues. For purposes of this project, we are not considering the relative orientation of atoms within the groups with respect to each other.

The algorithm for converting from atoms to groups *for* i^{th} *residue* is as follows:

1. Parse main chain atoms

   (a) If [starting residue] then group is r1 using $N_{i-1}, CA_i, C_i, O_i, N_{i+1}$; If N_{i-1} (N-terminal amine group) is absent, use the rest of the atoms

   (b) If [residue is not terminal residue] then group is r1 using CA_i, C_i, O_i, N_{i+1}

   (c) If [residue is ending residue] and then group is r1 using CA_i, C_i, O_i, OXT; if OXT absent, use the rest of the atoms

2. Parse side chain atoms using Table 1, and chemical group definitions described in Figure 4

3.2 Datasets

3.2.1 Identification of Native Models in a Decoy Set

A good structure evaluation method should be able to identify a native model in a set of decoy models. In this study, the goal was for the scoring scheme to rank order all the structural models in a decoy set, wherein lower rank of the native model implies better performance of the scoring scheme. The reason for doing this is primarily to find an estimate of the optimal star-size.

Table 1: List of amino acids and their constituent chemical groups

Amino Acid	Chemical Groups	Amino Acid	Chemical Groups
Arginine	r1 + r2 + r2 +r3	Glycine	r1
Histidine	r1 + r4	Proline	r11
Lysine	r1 + r2 + r2 + r2 + r5	Alanine	r1 + r8
Aspartic Acid	r1 + r6	Valine	r1 + r12 + r8 + r8
Glutamic Acid	r1 + r2 + r6	Isoleucine	r1 + r12 + r2 + r8 + r8
Serine	r1 + r7	Leucine	r1 + r2 + r12 + r8 + r8
Threonine	r1 + r7 + r8 + r8	Methionine	r1 + r13
Asparagine	r1 + r2 + r9	Phenylalanine	r1 + r2 + r14
Glutamine	r1 + r2 + r2 + r9	Tyrosine	r1 + r2 + r16
Cysteine	r1 + r10	Tryptophan	r1 + r2 + r15

The star-size is an important parameter to optimise for the following reason: Two stars can be superimposed only if the composition in terms of the chemical groups is the same. This ensures that there is a one-to-one mapping between the chemical groups of the two stars being superimposed. If there are multiple chemical groups with the same identity, we permute over all possible combinations of one-to-one mapping between the stars. The higher the star-size is, the more is the number of permutations that need to be done if there are multiple chemical groups of the same kind in a star. See Appendix A for more details. See Section 3.3 for details of the parameters and optimisation strategy.

The Moulder set[John, 2003] was taken as the set of decoys to evaluate our scoring scheme. It consists of 20 decoy sets, where each decoy set consists of one native model and 300 computationally modelled decoys. The Moulder set of decoys was constructed using iterative target-template alignment and comparative model-building to produce suboptimal models for the following 20 proteins:

```
1BBH 1C2R 1CAU 1CEW 1CID 1DXT 1EAF 1GKY 1LGA 1MDC 1MUP 1ONC 2AFN 2CMD 2FBJ
2MTA 2PNA 2SIM 4SBV 8I1B
```

The native model for 2PNA is the only NMR structure in the set, the rest of the proteins being X-ray crystallographic structures.

Since the scoring scheme is computationally very time intensive, 30 decoys out of the 300 for each set were arbitrarily[3] chosen as a smaller decoy subset for the study. The results for this study have been used to optimise certain scoring parameters relevant to the objectives of this dissertation, but this study was not one of those objectives. The results have been added to the Appendix for reference. See Appendix B for details of how well the native ranks w.r.t the decoys in the Moulder set.

3.2.2 Prediction of Regions of Refinement

A list of obsolete structures in the PDB and their replacement structures, is available online in the PDB database. The replacement structures are referred to as *successor/refined* structures in the rest of this text, while the obsolete ones may be referred to as

[3] with the sole criteria that the 30 decoys should be uniformly sampled across the set of 300, in terms of their RMSD w.r.t the native model

predecessors. Some of these structures have been solved again, while others are refined versions of their predecessors. While being blind to the changes made to get the successor structures, a good evaluation method should be able to predict the regions where changes were supposed to be made, and these should correlate well with the regions that actually got changed. It should be able to differentiate between the stars from the obsolete and the refined structures.

The list of obsolete records downloaded on 19th June 2018 has 3795 records. Some of these records are outdated and some have undergone multiple rounds of replacement. The data was cleaned up and culled for a non-redundant (NR) set (<30% sequence similarity using PISCES server [Wang and Dunbrack, 2003]). The resulting 115 records were used for this study. See Appendix C for details of how the records were cleaned and culled.

3.2.3 Correlation with Thermal Stability

Comparison of thermophilic and mesophilic proteins

1. Kumar's set [Kumar et al., 2000]: Contains a non-redundant set of 18 high quality thermophilic protein structures, all of which are of less than 2.5 Å resolution. The dataset was created using the 1998 version of PDB. The `source.idx` file in the PDB was searched for the keywords THERM and PYRO and cleaned and culled later. The structures are dissimilar (seqence identity <= 20% and RMSD >= 2.00 Å), and don't include NMR structures or theoretical models. 5 out of the 18 thermophilic proteins have known T_m values mentioned in the article itself, and 3 out of these have the mesophilic T_m mentioned as well. For all the 18 structures, high quality homologous mesophilic structures have been culled in a similar manner from the PDB.

   The objective is to see if our scoring scheme can differentiate between the thermophilic and the mesophilic proteins[4]. The objective is the same in case of ProTherm sets as well.

2. Szilágyi's set [Szilágyi and Závodszky, 2000]: Contains 25 protein families with 64 mesophilic and 29 thermophilic protein subunits, all of them being of high quality just like Kumar's set. The differentiation between thermophilic and mesophilic proteins is based on optimal living temperatures (T_{opt}) of the organisms from where the proteins were extracted from.

   The objective is to see if any one of the thermophilic homologs scores the best among all the homologs for any of the 25 proteins in the set. Ideally one would like to make sure that there is a correlation between the T_{opt} and the scores. The objective here is a crude approximation of this, since the mesophilic homologs in the set do not have T_{opt} mentioned.

3. ProTherm set [Bava et al., 2004]: ProTherm is a thermodynamic database for proteins and mutants, with ~25800 entries mapping proteins to available structures in the PDB, sequences in SWISS-PROT, thermodynamic data such as melting temperature (T_m), as well as additional data such as source organism for any of the entries. The ProTherm database was culled for entries for which the wild-type structure is available. Since most of the data is from differential scanning

[4] see end of Section 3.3.2 for a discussion on what is a better score

calorimetry experiments, the database consists of multiple T_m entries for each PDB entry. These correspond to the different phase transitions of the protein during the experiment, and in this study, the T_m corresponding to the last phase transition (highest T_m) was considered as the T_m of the protein.

The homolog with the minimum and maximum T_m are referred to as mesophilic and thermohilic proteins respectively, for this dataset. Note that this naming convention is such for the sake of clarity, and not what is usually found in literature. Mesophilic may be shortened to *meso* and thermophilic to *thermo* in this text.

Two different variants of this set were considered:

(a) ProTherm$_{lt60plus10}$ (19 proteins): mesophilic protein has $T_m < 60°C$ and difference between thermophilic and mesophilic proteins is at least 10°C [5]

(b) ProTherm$_{lt60gt70}$ (13 proteins): mesophilic protein has $T_m < 60°C$ and thermophilic protein has $T_m > 70°C$. This set is a subset of the ProTherm$_{lt60plus10}$ set.

	Family name	**PDB ID, source organism, Topt, resolution**
1	Transcription initation factor IIb (TIF-2B)	1volA (Human, meso) 2.7 Å
2		1aisB (Pyrococcus woesei, 100 °C) 2.1 Å
3	Superoxide dismutase (Mn- or Fe-dependent) (SOD)	1abmA (Human, meso) 2.2 Å
4		1ar4A (Propionibacterium freudenreichii, meso)...
5		1idsA (Mycobacterium tuberculosis, meso) 2.0 Å
6		1isaA (Escherichia coli, meso) 1.8 Å
7		1vewA (Escherichia coli, meso) 2.1 Å
8		3mdsA (Thermus thermophilus, 75°C) 1.8 Å
9	Glutamate dehydrogenase (Glu-DH)	1hrdA (Clostridium symbiosum, meso) 1.96 Å
10		1gtmA (Pyrococcus furiosus, 100°C) 2.2 Å
11	Malate dehydrogenase (MDH)	4mdhA (Pig heart, meso) 2.5 Å
12		1bmdA (Thermus flavus, 72.5°C) 1.9 Å
13	Phycocyanin alpha chain (Phyc-a)	1cpcA (Fremyella diplosiphon, meso) 1.66 Å
14		1liaA (Polysiphonia urceolata, meso) 2.8 Å
15		1allA (Spirulina platensis, meso) 2.3 Å
16		1phnA (Cyanidium caldarium, 45°C) 1.65 Å
17	Signal recognition particle (receptor) (SRP)	1fts (Escherichia coli, meso) 2.2 Å
18		1ffh (Thermus aquaticus, 72.5°C) 2.05 Å
19	Ferredoxin	1fxd (Desulfovibrio gigas, meso) 1.7 Å
20		1fxrA (Desulfovibrio africanus, meso) 2.3 Å
21		1vjw (Thermotoga maritima, 80°C) 1.75 Å
22	Subtilisin	1sup (Bacillus amyloliquefaciens, meso) 1.6 Å
23		1cseE (Bacillus subtilis, meso) 1.2 Å
24		1bh6 (Bacillus licheniformus, meso) 1.75 Å
25		1svn (Bacillus lentus, meso) 1.4 Å
26		2pkc (Tritirachium album limber, meso) 1.5 Å
27		1sbnE (Bacillus subtilis, meso) 2.1 Å
28		1meeA (Bacillus mesentericus, meso) 2.0 Å
29		1thm (Thermoactinomyces vulgaris, 60°C) 1.37Å
30	Neutral protease (thermolysin) (NPR)	1npc (Bacillus cereus, meso) 2.0 Å
31		1lnfE (Bacillus thermoproteolyticus, 52.5°C) 1...
32	Rubredoxin	1iro (Clostridium pasteurianum, meso) 1.1 Å
33		1rdg (Desulfovibrio gigas, meso) 1.4 Å
34		6rxn (Desulfovibrio desulfuricans, meso) 1.5 Å
35		8rxnA (Desulfovibrio vulgaris, meso) 1.0 Å
36		1caa (Pyrococcus furiosus, 100°C) 1.8 Å
37	Cyclodextrin glycosyltransferase (CGTase)	1cdg (Bacillus circulans strain 251, meso) 2.0 Å
38		1cgt (Bacillus circulans strain 8, meso) 2.0 Å
39		1pamA (Bacillus sp. 1011, meso) 1.8 Å
40		1ciu (Thermoanaerobacterium thermosulfurigenes...
41		1cyg (Bacillus stearothermophilus, 52.5°C) 2.5 Å
42	Phycocyanin beta chain (Phyc-b)	1allB (Spirulina platensis, meso) 2.3 Å
43		1cpcB (Fremyella diplosiphon, meso) 1.66 Å
44		1liaB (Polysiphonia urceolata, meso) 2.8 Å
45		1phnB (Cyanidium caldarium, 45°C) 1.65 Å
46	3-Phosphoglycerate kinase (PGK)	1qpg (Yeast, meso) 2.4 Å
47		1php (Bacillus stearothermophilus, 52.5°C) 1.65 Å
48		1vpe (Thermotoga maritima, 80°C) 2.0 Å

[5] The mesophilic T_m value is taken as less than 60°C because most similar studies in literature take a cutoff of $T_m \sim 50 - 60°C$ to differentiate between mesophilic and thermophilic protein structures. Instead of taking a hard–cutoff, a difference of 10°C was taken. This ensures that the mesophilic and thermophilic homologs are not very similar in terms of T_m

49	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	1a7kA (Leishmania mexicana, meso) 2.8 Å
50		1gadO (Escherichia coli, meso) 1.8 Å
51		1szjG (Palinurus versicolor, meso) 2.0 Å
52		1gd1O (Bacillus stearothermophilus, 52.5°C) 1.8 Å
53		1hdgO (Thermotoga maritima, 80°C) 2.5 Å
54	Xylanase (I) (Xyl-1)	1enxA (Trichoderma reesei, meso) 1.5 Å
55		1ukrA (Aspergillus niger, meso) 2.4 Å
56		1xnb (Bacillus circulans, meso) 1.49 Å
57		1xnd (Trichoderma harzianum, meso) 1.8 Å
58		1xyn (Trichoderma reesei, meso) 2.0 Å
59		1yna (Thermomyces lanuginosus, 45°C) 1.55 Å
60	Xylanase (II) (Xyl-2)	1clxA (Pseudomonas fluorescens, meso) 1.8 Å
61		2exo (Cellulomonas fimi, meso) 1.8 Å
62		1xyzA (Clostridium thermocellum, 60°C) 1.4 Å
63	TATA box binding protein (TATA-BP)	1cdwA (Human, meso) 1.9 Å
64		1vokA (Arabidopsis thaliana, meso) 2.1 Å
65		1pczA (Pyrococcus woesei, 100°C) 2.2 Å
66	Adenylate kinase (ADK)	1ak2 (Bovine, meso) 1.92 Å
67		2ak3A (Bovine, meso) 1.85 Å
68		1aky (Yeast, meso) 1.63 Å
69		1ukz (Yeast, meso) 1.9 Å
70		1akeA (Escherichia coli, meso) 1.9 Å
71		3ukd (Dictyostelium discoideum, meso) 1.9 Å
72		1zip (Bacillus stearothermophilus, 52.5°C) 1.85 Å
73	Carboxypeptidase (CP)	2ctc (Bovine, meso) 1.4 Å
74		1nsa (Pig, meso) 2.3 Å
75		1pca (Pig, meso) 2.0 Å
76		1obr (Thermoactinomyces vulgaris, 55°C) 2.3 Å
77	Ornithine carbamoyltransferase (OCT)	2otcA (Escherichia coli, meso) 2.8 Å
78		1a1s (Pyrococcus furiosus, 100°C) 2.7 Å
79	Pyrophosphatase (PPase)	1obwA (Escherichia coli, meso) 2.15 Å
80		2prd (Thermus thermophilus, 72.5°C) 2.0 Å
81	CheY protein (CheY)	3chy (Escherichia coli, meso) 1.66 Å
82		2chf (Salmonella typhimurium, meso) 1.8 Å
83		1tmy (Thermotoga maritima, 80°C) 1.9 Å
84	Glutathione / trypanothione reductase (G/T re...	1aogA (Trypanosoma cruzi, meso) 2.3 Å
85		1febA (Crithidia fasciculata, meso) 2.0 Å
86		1gerA (Escherichia coli, meso) 1.86 Å
87		3grs (Human, meso) 1.54 Å
88		1ebdA (Bacillus stearothermophilus, 52.5°C) 2.6 Å
89	Phosphofructokinase (PFK)	1pfkA (Escherichia coli, meso) 2.4 Å
90		4pfk (Bacillus stearothermophilus, 52.5°C) 2.4 Å
91	Triacylglycerol acylhydrolase (TAGAH)	1lgyA (Rhizopus niveus, meso) 2.2 Å
92		1tib (Humicola lanuginosa, 50°C) 1.84 Å
93		3tgl (Rhizomucor miehei, 45°C) 1.9 Å

Table 2: Szilágy's set [Szilágyi and Závodszky, 2000]

Correlating scores with B-factor values:

A non-redundant high-resolution subset of the PDB database was extracted. This set has 243 structures. The culling for the non-redundant set was done using PISCES server [Wang and Dunbrack, 2003] using parameters laid out in Table 4

3.3 Scoring Methods

Since the structure of the protein is being defined in terms of chemical groups, and every chemical group defines a star around it, no region in the model is left un-evaluated.

The regions across an obsolete protein model in this case can be evaluated quantitatively by scoring the stars in the model – the stars being representatives of the different regions in the model.

Two different scoring schemes have been tried out, and are discussed below.

3.3.1 Scoring stars in terms of RMSD of best-match in the PDB

Consider a query star that is from a native-like model. Since we are assuming that most native stars have been recorded in the PDB already, we should be able to find a star in the PDB which is similar to the query star. The similarity is in terms of both the composition of the star (which is in terms of the identities of the chemical groups present in it) and in terms of geometry of the star, ie. relative arrangement of these chemical groups in 3D.

	Protein	PDB ID	Source	T_m in °C
1	3-isopropylmalate dehydrogenase*	1CM7	Escherichia coli	34.00
2		1OSI	Thermus thermophilus	87.40
3		1WPW	Sulfolobus sp. strain 7	96.00
4	Acylphosphatase*	1APS	Human	56.70
5		1Y9O	Sulfolobus solfataricus	100.80
6	Aldolase	1ADO	Rabbit	61.20
7		1DHN	Staphylococcus aureus	44.00
8	Alpha-amylase*	1AQH	Pseudoalteromonas haloplanktis	44.00
9		1BPL	Bacillus licheniformis	104.30
10		1JAE	Tenebrio molitor	66.40
11		1PPI	Pig	65.60
12		1SMD	Human	70.40
13		3KWX	Aspergillus oryzae	86.00
14	Alpha-lactalbumin*	1HFY	Goat	71.20
15		1HFZ	Bovine	71.30
16		1HML	Human	43.00
17	Beta lactamase	1BLC	Staphylococcus aureus	41.60
18		1BMC	Bacillus cereus	51.03
19		3BLS	Escherichia coli	54.60
20		4BLM	Bacillus licheniformis	68.90
21	Cel12A	1H8V	Trichoderma reesei	54.40
22		1OA2	Gliocladium roseum	45.90
23		1OA3	Hypocrea schweinitzii	49.20
24		1OLR	Humicola grisea	68.70
25	Cytochrome c*	1AKK	Horse	83.00
26		1I5T	Rat	60.00
27		1YCC	Saccharomyces cerevisiae	51.70
28		2B4Z	Bovine	78.00
29	Cytochrome c oxidase	1AR1	Paracoccus denitrificans	67.00
30		1OCC	Bovine	57.00
31	Frataxin	1EKG	Human	69.30
32		1EW4	Escherichia coli	64.10
33		2GA5	Saccharomyces cerevisiae	53.60
34	Lipase*	2FX5	Pseudomonas mendocina	53.00
35		3D2A	Bacillus subtilis	71.20
36	Lysozyme*	1AM7	Lambda phage	52.30
37		1EL1	Canine	90.00
38		1H09	Bacteriophage Cp-1	52.00
39		1LZ1	Human	80.10
40		2EQL	Horse	70.00
41		2LZM	Bacteriophage T4	68.00
42		4LYZ	Chicken	91.90
43	Myoglobin*	1BVC	Sperm whale	82.20
44		1YMB	Horse	84.00
45		2FAL	Aplysia limacina	52.00
46	Prion protein*	1AG2	Mouse	71.00
47		1QLX	Human	60.00
48		1UW3	Sheep	70.00
49	Pyrophosphatase*	1FAJ	Escherichia coli	93.00
50		1K23	Bacillus subtilis	50.00
51		1QEZ	Sulfolobus acidocaldarius	98.00
52		2PRD	Thermus thermophilus	99.00
53	Ribonuclease A*	1DZA	Human	53.70
54		1RTB	Bovine	90.00
55	Triose-phosphate isomerase*	1BTM	Bacillus stearothermophilus	102.00
56		1TPE	Trypanosoma brucei	57.00
57		1YPI	Saccharomyces cerevisiae	59.00
58		3TIM	Trypanosoma brucei brucei	52.20
59	Tropomyosin*	1IC2	Chicken	74.00
60		2T_mA	Rat	54.20
61	Tryptophan synthase alpha-subunit	1WQ5	Escherichia coli	62.40
62		2WSY	Salmonella typhimurium	47.60

Table 3: Proteins that are part of the ProTherm$_{lt60plus10}$ dataset. Only the structures corresponding to the minimum and maximum T_m values were considered for analysis.
* marked proteins refer to the subset of proteins in the ProTherm$_{lt60gt70}$ set

Total number of structures	**243**
Pairwise sequence similarity	<= 30%
Resolution	<= 1.0
R-factor	<= 0.3
Non X-ray entries	Excluded
CA-only entries	Excluded

Table 4: Culling parameters for dataset used for B-factor study

To score a certain query star, the metric in this method, is the RMSD of superimposition with the best-match star in the PDB. If there are no matches within an RMSD cutoff, we penalise the star with a penalty score. While scoring, n and d_{thr} are specified for star-size and distance-threshold respectively. Star-size is the number of chemical groups in a star, and distance-threshold is the distance from the central chemical group to the farthest one in the star.

This *best-match RMSD* method doesn't allow us to directly answer how bad is a bad score. A star which is usually buried because it has hydrophobic chemical groups, can probably find low RMSD matches because there isn't much flexibility and consequently better scores. However, a star which is supposed to be exposed, will have more flexibility. This star will therefore find higher RMSD matches and have a higher (worse) score. We have no means to say how high a score is bad enough to say that this star needs refinement. In general, since, RMSD values can go from zero to arbitrarily high numbers, there is no reasonable way to say below what RMSD cut-off should the RMSD of superimposition imply a *good match*, so that one may infer that the query is a good star.

Because of the above mentioned reason, we tried an alternate scoring scheme that takes the distribution of geometries in the PDB into account, and allows us to differentiate between good and bad packings.

The usage of this RMSD based method was limited to the decoy-set study and for optimising parameters such as star-size. The scoring scheme in the following section is what was used for all other studies discussed in this text.

3.3.2 Scoring stars in terms of distribution of similar geometries (CASPER)

Ideally, we would like to cluster all the stars of a given composition in the PDB, in terms of geometry, and because of redundancy in the PDB, we would like to try matching our query star against one, or a few representatives from each cluster. However, clustering is a computationally expensive job. To perform a clustering, an all-against-all superimposition needs to be done for all the stars of any composition. Then the RMSDs could be used as linkage distances to cluster the stars in RMSD space. With current computational capabilities this is impractical. Instead an approximate method has been tried here to get similar results.

Consider a scenario where we had such clusters already, and representatives of each cluster to test our query star against. Once we find a matching representative for our query star, we can compare the query star against the cluster that the representative belongs to, and see how close the query is to the cluster. An approximation to this scheme can be made in the following manner: Given a query star, we match it against a sample S number of target stars of the same composition from the PDB [6] and rank order these target stars based on RMSD. The top 100 or top matches with less than 2Å RMSD[7] matches are taken as what we call *sisterly set*.

If the query star is a native-like star, this sisterly-set is supposed to be of mostly similar geometries and would have belonged to the same cluster of stars, had we done a proper clustering. Since the query star is close to this set of stars in RMSD space, it should be *closer* or almost as *close* as the top ranked star (otherwise referred to as *Superstar* in this text, for clarity). To quantify the distribution of geometries in the sisterly set, with respect to the query star, we use the following scheme:

$$score = \sum_{x=a+kn}^{b} \frac{f(x)}{N} \tag{1}$$

i.e, in steps of k, we sum $f(x)$ from a to b, and $N = \frac{b-a}{k} + 1$. $f(x)$ is the fraction of stars in the sisterly-set that matched with the query star with RMSD *lesser than* xÅ. For this study, we have used k=0.1, a=1.0Å and b=2.0Å. i.e.

$$score = \frac{(f(1.0) + f(1.1) + f(1.2) + ... + f(2.0)}{11} \tag{2}$$

a is chosen as 1.0 since the resolution of these amino-acids is ~1Å, so we assume that less than 1Å RMSD matches are too close to be considered as different from the query star at all.

With this scheme in mind, consider a query star which is extremely non-native-like. This star will not find close matches in the sisterly-set and therefore, the first term in the numerator will be ~0 and thus the total score will tend to 0, since the rest of the terms are populated using subsets of the first term. On the other hand, a native-like query star will have a score closer to 1, since there will be many target stars with RMSD low enough to contribute to the later terms in the scoring function. Since the *Superstar* is native-like, being part of the PDB already, we expect it to have a low score too (with respect to the sisterly set). A native-like query star will have a score less than or equal to the Superstar's score, since the sisterly-set is computed with respect to the query star. For this reason, this kind of a scoring scheme although is better than the previous method, provides an advantage to the query-star. A variant of this method wherein the sisterly set is computed w.r.t to the Superstar, shall be tested in the future.

This scheme will be referred to as the **CASPER scheme** (Cumulative Average distribution of Stars as Protein local Environment Representatives) in the rest of this text. If q is the score obtained for the query-star and s is that of the superstar/best-match star, then:

- q may be referred to as CASPER query-star score. A higher score is better in this case, since that would mean that the query-star finds more representatives that are similar in geometry, in the PDB.
- s - q is the CASPER-badness score. A higher badness score is worse, since it implies that the superstar has a geometry very different from that of the query-star, which

[6] comparing against all the stars of the same composition is computationally expensive. The number of stars of any given composition is of the order of $\sim 10^6$. We sample in the order of $\sim 10^3$ to get decent results in reasonable time. See Appendix refsec:appD for more details

[7] *ad hoc* estimate based on resolution of the structures being scored; to be optimised in the future

allows it to find better matches in the sisterly set which was constructed w.r.t to the query-star[8].

- CASPER one-sided badness equals s - q if $s > q$ else it is 0. One may argue that if the query-star scores better than the superstar, it is a trivial solution, since the sisterly set was constructed w.r.t to the querystar. Therefore by construction, any query-star where the query-star score itself is higher than the superstar score can be given a baseline score of zero, and this one-sided badness scores can only be positive.

3.3.3 Calculation of Chemical Group Temperature Factors

Note that B-factor values are assigned to atoms, based on their thermal fluctuation w.r.t the larger structure itself, quantified in terms of structure factor during any X-ray crystallography experiment. To translate it to chemical group wise B-factor values following methods were tried out:

- Percentile rank of an atomic B-factor, averaged over all the atoms in the chemical group
- Z-score of the atomic B-factors, averaged over all the atoms in the chemical group
- Depth based Z-score of the atomic B-factors. The distribution of B-factors is bimodal normal in terms of depth of the atoms. We calculated the z-score of the atomic B-factors w.r.t to one of the two normal distributions that it belongs to be based on the atomic depth. Atomic depth was calculated using Depth software (stand-alone version) [Tan et al., 2013], with default parameter values.

In this study, star-wise scores are best-match RMSDs/ Superstar RMSD for each stars. Chemical group-wise scores are average of all the scores of the stars that a chemical group participates in.

3.3.4 Scoring parameters

For scoring any of the PDB files that are discussed in this text, the scoring parameters are as follows:

1. Star-size = 7
   (a) Star-sizes of 7, 8, 9, and 10 were used for the decoy-set study (Section 3.2.1) with the following subset of the proteins in the decoy set: `1BBH 1EAF 1GKY 1MDC 1ONC 2AFN 2CMD 2FBJ 2SIM`, which were arbitrarily picked from the set of 20 proteins. Since star-size 7 was the lowest size for which all 9 had the native ranked 1, star-size was set at 7 as a rough estimate to save computational time.
   (b) RMSD-cutoff for structural overlap calculation = 2Å(the inter-chemical group distance is approx. the same, on an average); Only complete matches (structure overlap of 100% are counted as *matches*, if at all

[8] which is why it is called *badness* and not *goodness*

(c) Penalty value=5 (arbitraily high penalty value; Stars that don't count as matches are given this value for RMSD of superimposition, and a value of zero is assigned for the CASPER query-star score)

(d) Sampling size=2000 (unless a different value is mentioned for any specific study; This is the maximum number of stars that are sampled as *target stars* for any query-star, among all the stars in the PDB)

3.4 Conflicting sequences between predecessor and successor structures

While comparing stars between predecessor and successor structures[9] it is important to have a one-to-one mapping between the ones in the predecessor and the ones in the successor structure, so that there is a fair comparison between the two versions of any given star (obsolete and refined). This means that there needs to be a one-to-one mapping between the amino acids in a similar way, since that would allow a mapping for chemical groups and hence for the corresponding stars as well. However, sometimes certain amino acids in the predecessor may not be the same as that in the successor, for various reasons for example because of poorly resolved experimental data. More often, amino acids near the ends of a chain may be present in one of the two (predecessor/successor) but absent in the other. This may happen because the terminal stretches are usually floppy and poorly packed w.r.t to the rest of the structure. In such a scenario it becomes difficult to resolve the coordinates of those atoms out of the experimental data and those atoms are sometimes left out of the final structure deposited in the PDB.

To deal with such cases, a sequence alignment was performed between the predecessor and successor sequence (Smith-Waterman local sequence alignment, using MODELLER [Webb and Sali, 2016]). Only the amino acids which were in consensus at a certain position in the sequence were selected. This also provided a list of equivalences between the chemical groups [10].

3.5 Different residue environment upon refinement

When a structure is tagged obsolete and is replaced by a successor structure, there are modifications made in the geometry of the structure. The stars in the successor structure may be different in terms of their chemical group composition[11]. Therefore, some of the stars in the predecessor don't have a counterpart with the same composition in the successor, and we can't compare these two stars directly. They represent different residue environments in the structures. However, if the composition of the star stays the same, the residue environment probably hasn't changed much, and therefore need have had extensive modifications during refinement. Because of this difference, and because two stars with differing compositions can't be superimposed by our method, we considered

[9] see Section 3.2.2 for definitions of *predecessor* and *successor*

[10] Note that since every chemical group defines a star around it, if a one-to-one mapping exists between predecessor and successor chemical groups, this mapping can be extended to say that the stars around these chemical groups are also one-to-one mapped. The identity of a star is essentially the central chemical group's number itself.

[11] For example, for a 5-body star with identities of chemical group members as: {r1, r2, r2, r5, r8} may be a certain composition. Replacing the r8 with say an r10 chemical group as the fifth member of the star, leads to a change in composition

the two cases separately. Stars that have changed in composition were considered at various levels of similarity.

3.5.1 Consideration of homologous structures in the PDB

As mentioned earlier, a non-redundant subset of the PDB database is used for comparison with the PDB. This subset has pairwise sequence similarity less than 30% between the structures. While scoring a structure from the PDB (for example, in case of 3.2.3), all structures with sequence similarity greater than 30% are excluded from the sampling space of stars. The usage of a non-redundant PDB database makes the computation more tractable in terms of time taken. Further, since the structures compared have very less similarity, they are likely to be non-homologous. Thus, only the structure and fold of the protein becomes a major factor during the comparison, instead of finding matches between homologous structures with similar sequences which is a trivial solution anyway.

3.5.2 Contact Order

It is known that there is a statistically significant relationship between protein folding kinetics and the contact order (CO) of different amino acids in a protein sequence [Plaxco et al., 1998, Grantcharova et al., 2001]. CO is defined to be the average *sequence separation* of residues that form contacts in the 3D structure of the protein. A set of residues with higher contact order will have contribute to the protein folding more slowly than others with lower CO, since it costs more in terms of entropy to bring high CO residues together in space. These residues are usually found at lower residue depths and are the first to contribute to an unfolding process.

With this logic, for the study regarding the melting temperature of proteins, we tried scoring the protein structures only in terms of stars that have a high CO. This was performed at contact order cutoffs of 5%, 10%, and 20%. CO cutoff in this text implies that stars were ranked in decreasing order of their CO values and the fraction of top ranking stars within the percentage cutoff are taken for scoring the whole protein structure.

The CO cutoff for a star is defined here as the maximum CO between any of the chemical groups in the star. Since chemical groups are numbered serially, [12] the CO values can be obtained for chemical groups by taking the difference between the chemical group numbers for any pair of chemical groups.

[12] even though they are not in a linear sequence and therefore this CO calculation is an approximation; the main chain for a residue is numbered first and then the side chains are numbered in terms of increasing distance of connectivity from the main chain chemical group

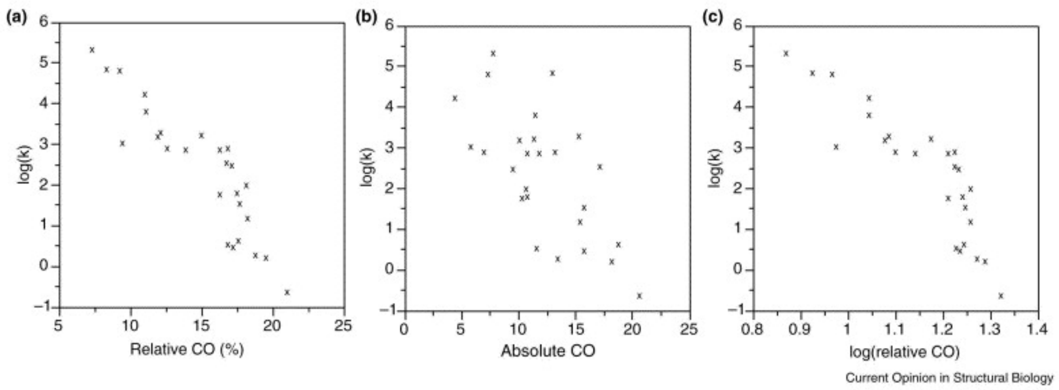


Figure 5: Correlation between logarithm of folding rate and (a) relative CO (i.e. absolute CO divided by chain length),(b) absolute CO and (c) log(relative CO)
[Grantcharova et al., 2001]

4 Results

4.1 Prediction of Regions of Refinement

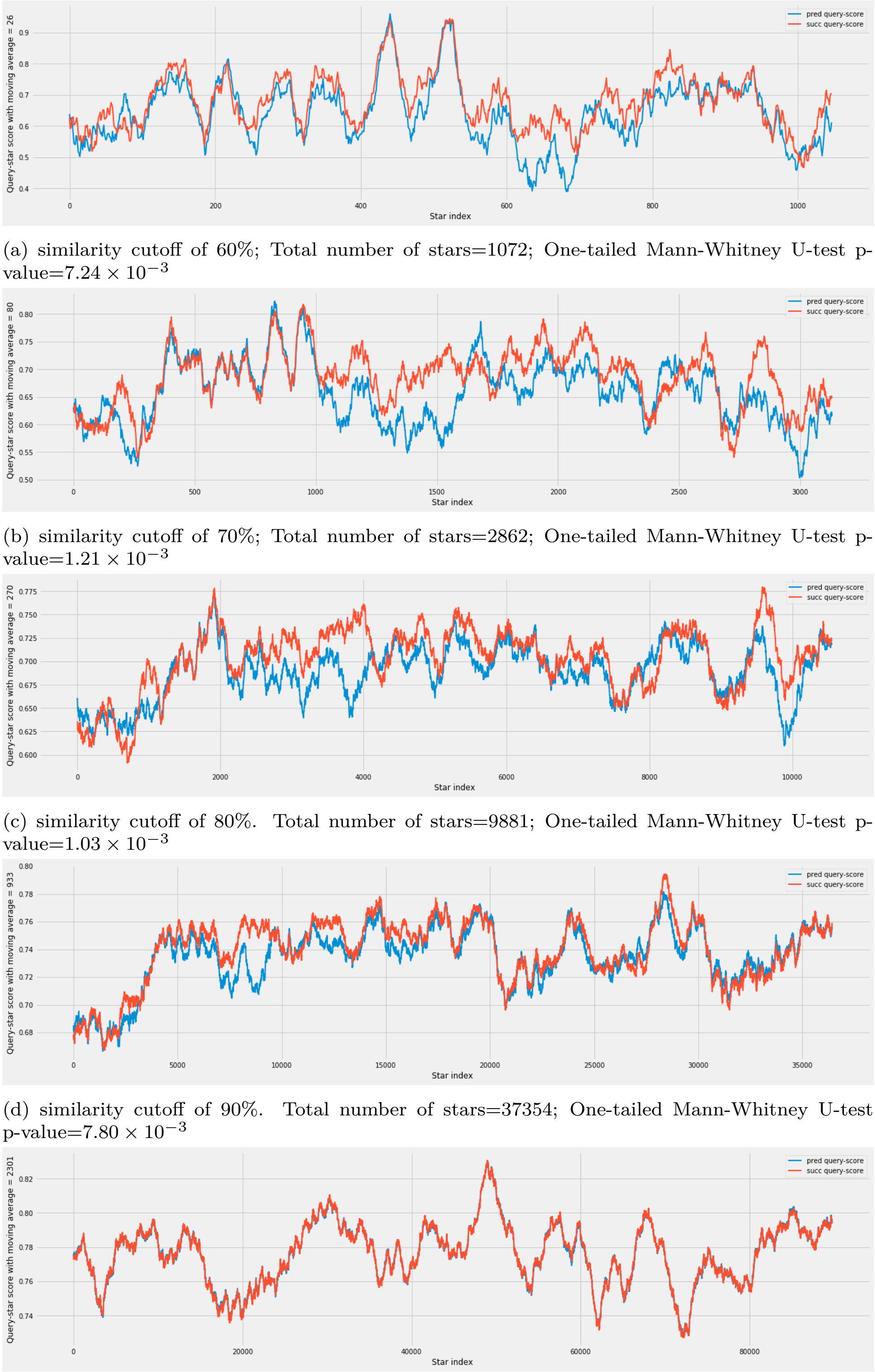
For each record in the obsolete structures list in the PDB, a pair of *obsolete* (otherwise referred to as *predecessor* here) and its corresponding successor structure is present. The set of predecessor structures and the successor structures were scored and the scores were compared for the two sets.

Overall trends in star-wise scores: With a good local structure evaluation method, one can modify different parts of a structure and be able to judge whether there was a refinement in the structure or not. This means that our evaluation method should at least be able to differentiate between the refined and the predecessor structures.

When the composition of stars is changed upon refinement, we notice the successor scores are better[13] than those for the predecessor using any of the CASPER score metrics. We performed a one-sided Mann-Whitney U-test with the null hypothesis that it is equally likely that a randomly selected successor star-score is higher than or lower than that of the predecessor. The alternative hypothesis is that the predecessor star-score is *worse* than that of the successor. Note that when the composition of stars change, there are only minor modifications. It's less likely to find half of the star to have a different composition, compared to finding that one of the chemical groups has been replaced by another. Therefore, we checked the trends for different levels of similarity cutoff. Similarity here is the ratio of chemical groups in the star that have stayed the same between the predecessor and the successor versions of it. In all the cases that we checked (similarity values of at least 40% to 90%, in intervals of 10%), the null hypothesis can be rejected. See Figures 6–8 for moving average profiles and details of Mann-Whitney U-test results. The stars are one-to-one mapped across the two profiles using methods described in Section 3.4. Outliers have been removed from the figures for CASPER Badness or CASPER one-sided badness score profiles. Very few stars are selected when maximum similarity between the stars is less than 40%, with only 39 stars in the 30% case, and 1 star in the 20% case.

Outliers: The mean CASPER badness scores for predecessors was less than -0.1 in

[13] see end of Section 3.3.2 for a discussion on what is a better score based on the way the scoring scheme is constructed



(a) similarity cutoff of 60%; Total number of stars=1072; One-tailed Mann-Whitney U-test p-value=7.24×10^{-3}

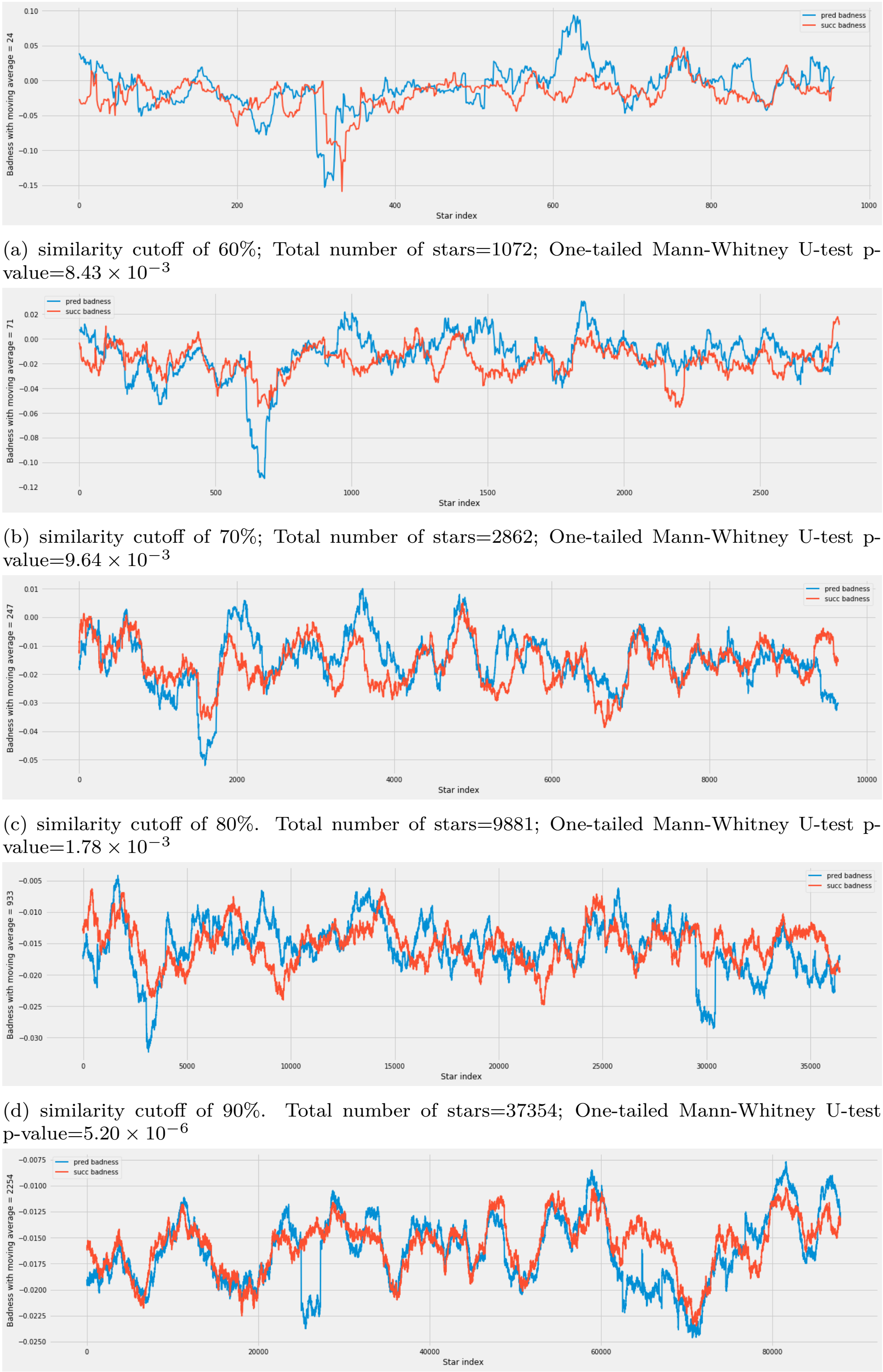
(b) similarity cutoff of 70%; Total number of stars=2862; One-tailed Mann-Whitney U-test p-value=1.21×10^{-3}

(c) similarity cutoff of 80%. Total number of stars=9881; One-tailed Mann-Whitney U-test p-value=1.03×10^{-3}

(d) similarity cutoff of 90%. Total number of stars=37354; One-tailed Mann-Whitney U-test p-value=7.80×10^{-3}

(e) stars that didn't change in composition after refinement; Total number of stars=90181; One-tailed Mann-Whitney U-test p-value=0.395

Figure 6: Star-wise CASPER query-star scores for predecessor vs successor structures. Moving average window is 2.5% of the total number of stars. Note the difference in number of stars evident from the x-axis.



(a) similarity cutoff of 60%; Total number of stars=1072; One-tailed Mann-Whitney U-test p-value=8.43×10^{-3}

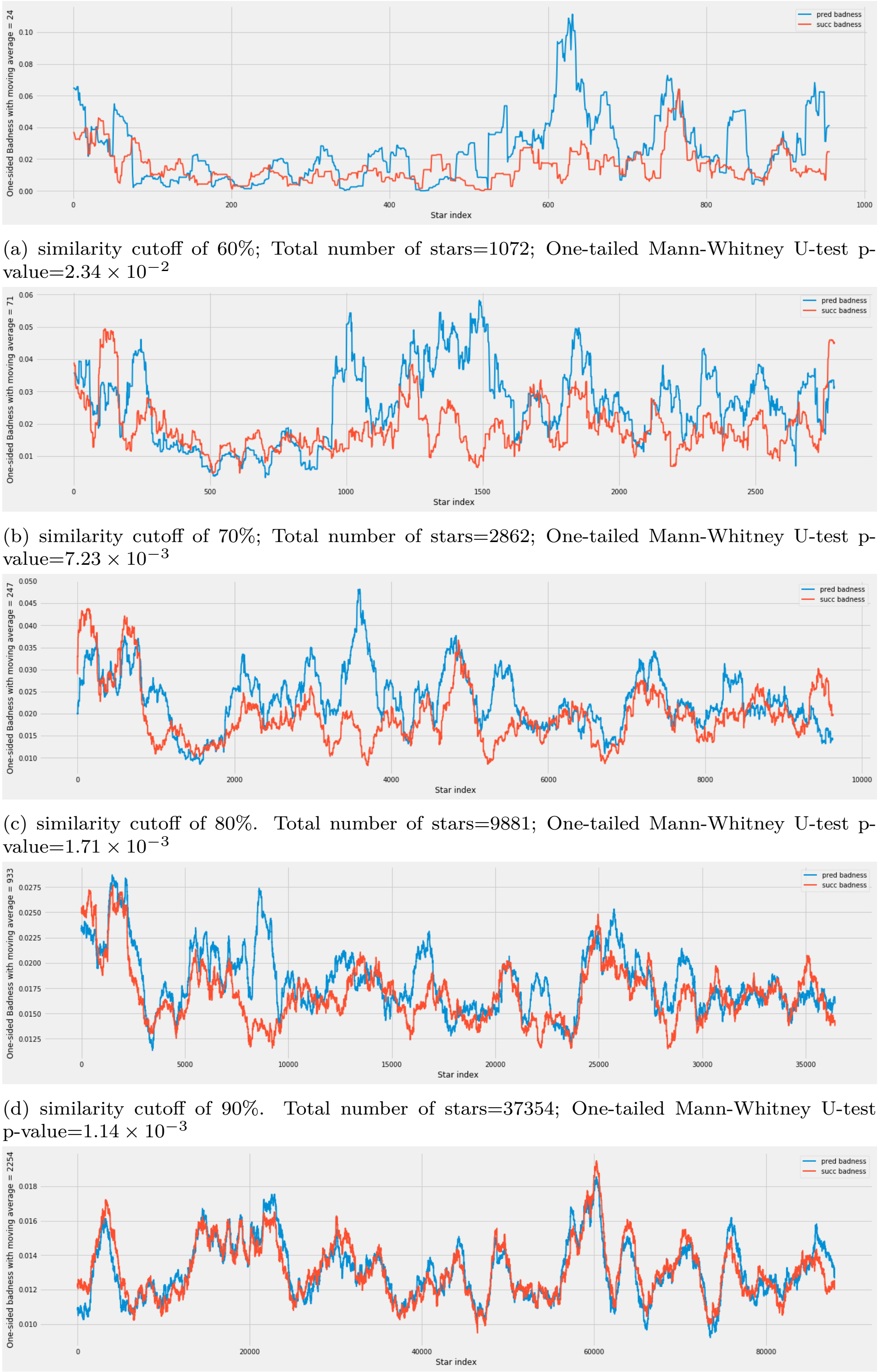
(b) similarity cutoff of 70%; Total number of stars=2862; One-tailed Mann-Whitney U-test p-value=9.64×10^{-3}

(c) similarity cutoff of 80%. Total number of stars=9881; One-tailed Mann-Whitney U-test p-value=1.78×10^{-3}

(d) similarity cutoff of 90%. Total number of stars=37354; One-tailed Mann-Whitney U-test p-value=5.20×10^{-6}

(e) stars that didn't change in composition after refinement; Total number of stars=90181; One-tailed Mann-Whitney U-test p-value=0.339

Figure 7: Star-wise CASPER badness scores for predecessor vs successor structures. Moving average window is 2.5% of the total number of stars.



(a) similarity cutoff of 60%; Total number of stars=1072; One-tailed Mann-Whitney U-test p-value=2.34×10^{-2}

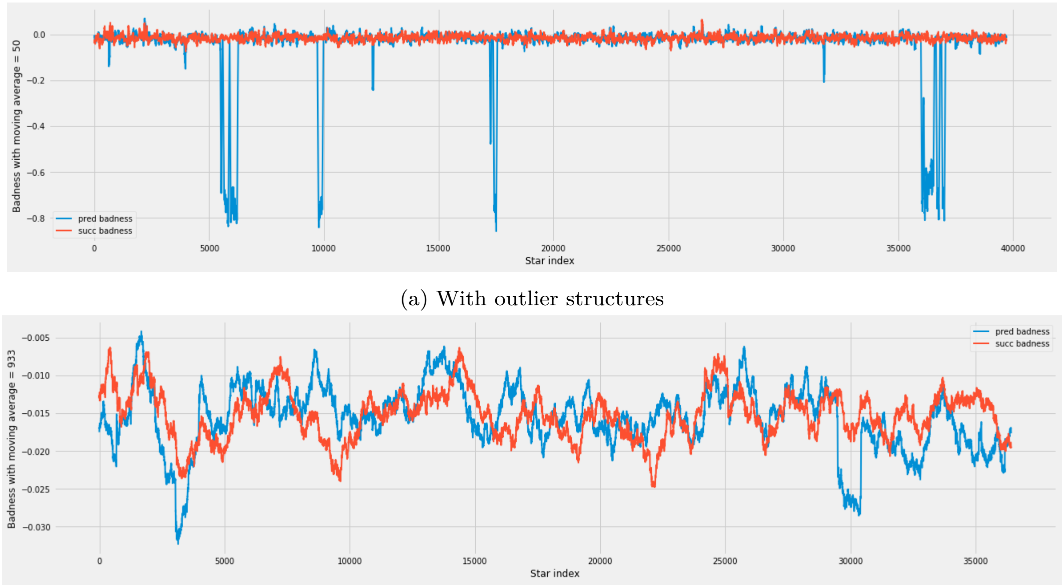
(b) similarity cutoff of 70%; Total number of stars=2862; One-tailed Mann-Whitney U-test p-value=7.23×10^{-3}

(c) similarity cutoff of 80%. Total number of stars=9881; One-tailed Mann-Whitney U-test p-value=1.71×10^{-3}

(d) similarity cutoff of 90%. Total number of stars=37354; One-tailed Mann-Whitney U-test p-value=1.14×10^{-3}

(e) stars that didn't change in composition after refinement; Total number of stars=90181; One-tailed Mann-Whitney U-test p-value=0.339

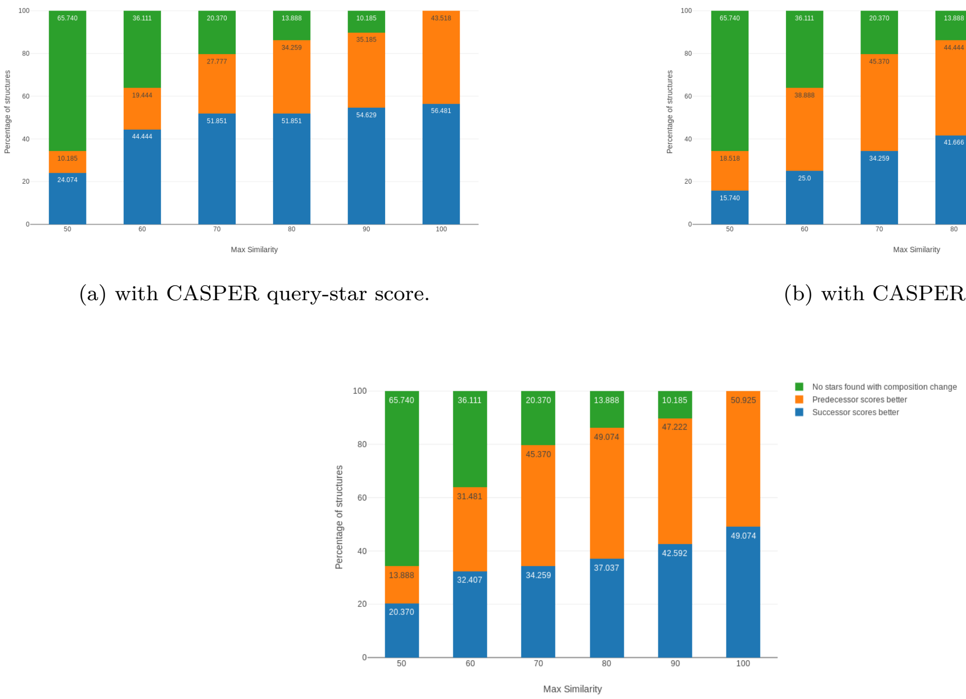
Figure 8: Star-wise CASPER one-sided badness scores for predecessor vs successor structures. Moving average window is 2.5% of the total number of stars.



(a) With outlier structures

(b) Without outlier structures

Figure 9: Outlier structures with CASPER Badness
Note the y-axis range. Since this is a moving average plot for illustration purposes, this is not the actual range of CASPER Badness scores, but an approximation of it.



(a) with CASPER query-star score.

(b) with CASPER badness

(c) with CASPER one-sided badness

Figure 10: Percentage of successor/predecessor structures that score better than their counterpart; plots shown at different levels of maximum similarity between stars compared from the structures. The three subfigures show the distributions for the three different CASPER scoring metrics. See Table 5 for absolute numbers.

7 out of the 115 pairs of structures. These turn up as outliers among the rest when badness is plotted against star-index. See Figure 9. Note that once the 7 outlier pairs are removed, the rest of the scores fall within a much narrower range of values (approximately 10% of the range of values that we get with the outliers included). There was no common feature or pattern across these outliers that could be found. However, further work needs to be done to make a conclusive statement as to why these pairs turned up as outliers.

Figure 10 shows how many of the refined (successor) structures were identified correctly, differentiated from the predecessor structure. The best performance was with CASPER query-star score with 70.27% of structures identified, at 50% maximum similarity (26 out of 37 structures). One-tailed Mann-Whitney U-test was also performed for each of the 108 structures. In less than 10% of the structures was the p-value for rejecting the null hypothesis less than 0.1. The comparison shown in Figure 10 and Table 5 are based on the mean score of the structure, using the CASPER scoring metrics.

Max Similarity %	CASPER Score Metric	% of structures where successor scores better	% of structures where predecessor scores better	No. of proteins where there was no change in star composition
100	QSS	61	47	0
	Badness	57	51	0
	One-sided Badness	53	55	0
90	QSS	59	38	11
	Badness	54	43	11
	One-sided Badness	46	51	11
80	QSS	56	37	15
	Badness	45	48	15
	One-sided Badness	40	53	15
70	QSS	56	30	22
	Badness	37	49	22
	One-sided Badness	37	49	22
60	QSS	48	21	39
	Badness	27	42	39
	One-sided Badness	35	34	39
40 or 50	QSS	26	11	71
	Badness	17	20	71
	One-sided Badness	22	15	71

Table 5: Comparison of number of higher scoring structures of successor versus predecessor structures. Total number of pairs of structures=108 pairs (one pair is one predecessor and corresponding successor structure). QSS=Query-Star Score

4.2 Correlation with Thermal Fluctuations

4.2.1 Differentiating thermophilic and mesophilic homologs

Correctly differentiating thermophilic proteins from mesophilic homologs can help us better design thermostable proteins. For each of the datasets, the PDB files for the structures mentioned in the set were downloaded. They were then scored and the results have been tabulated in Table 6. Note that 40% similarity and 50% similarity are the same, since they correspond to 60% and 50% *dissimilarity*. Since the star-size is 7, these refer to 4 out of the 7 chemical groups being dissimilar, in both the similarity cutoffs.

The net protein score was calculated at various contact order cutoffs. See Section 3.5.2 for the definitions of contact order and contact order cutoff. Note that Szilágy's set has multiple mesophilic structures with no mention of T_{opt} of source organism, and multiple

Method (CO cutoff)	Kumar's set	Szilágy's set[1]	ProTherm $lt60plus10$	ProTherm $lt60gt70$
QSS	13	13	9	7
QSS (5%)	8	10	12	8
QSS (10%)	12	12	11	9
QSS (20%)	12	12	12	8
badness	10	10	9	7
badness (5%)	8	7	12	8
badness (10%)	8	7	10	7
badness (20%)	10	9	7	5
one-sided badness	10	13	9	6
one-sided badness (5%)	9	8	8	5
one-sided badness (10%)	5	13	11	9
one-sided badness (20%)	8	10	8	5
Size of dataset	**18**	**25**	**19**	**13**

Table 6: Mean scores comparison of thermophilic vs mesophilic structures; Number of proteins where thermophilic structure scored better than mesophilic structure.
[1] Note that Szilágy's set has multiple mesophilic structures with no T_{opt} mentioned. The objective was to see **if *a* thermophilic protein scores the highest CASPER query-star score, or the lowest, in the badness metrics**. Only if one of the thermophilic proteins scored as rank 1 among all the structures for a specific protein, the protein was counted, otherwise it wasn't counted for this table.

thermophilic homologous structures for the same protein. Kumar's dataset doesn't mention T_m values (except for three of the proteins). T_{opt} of source organism need not correlate with T_m of the protein in consideration. This is perhaps the reason why CASPER metrics fare badly in case of Szilágy's set, compared to the other two datasets.

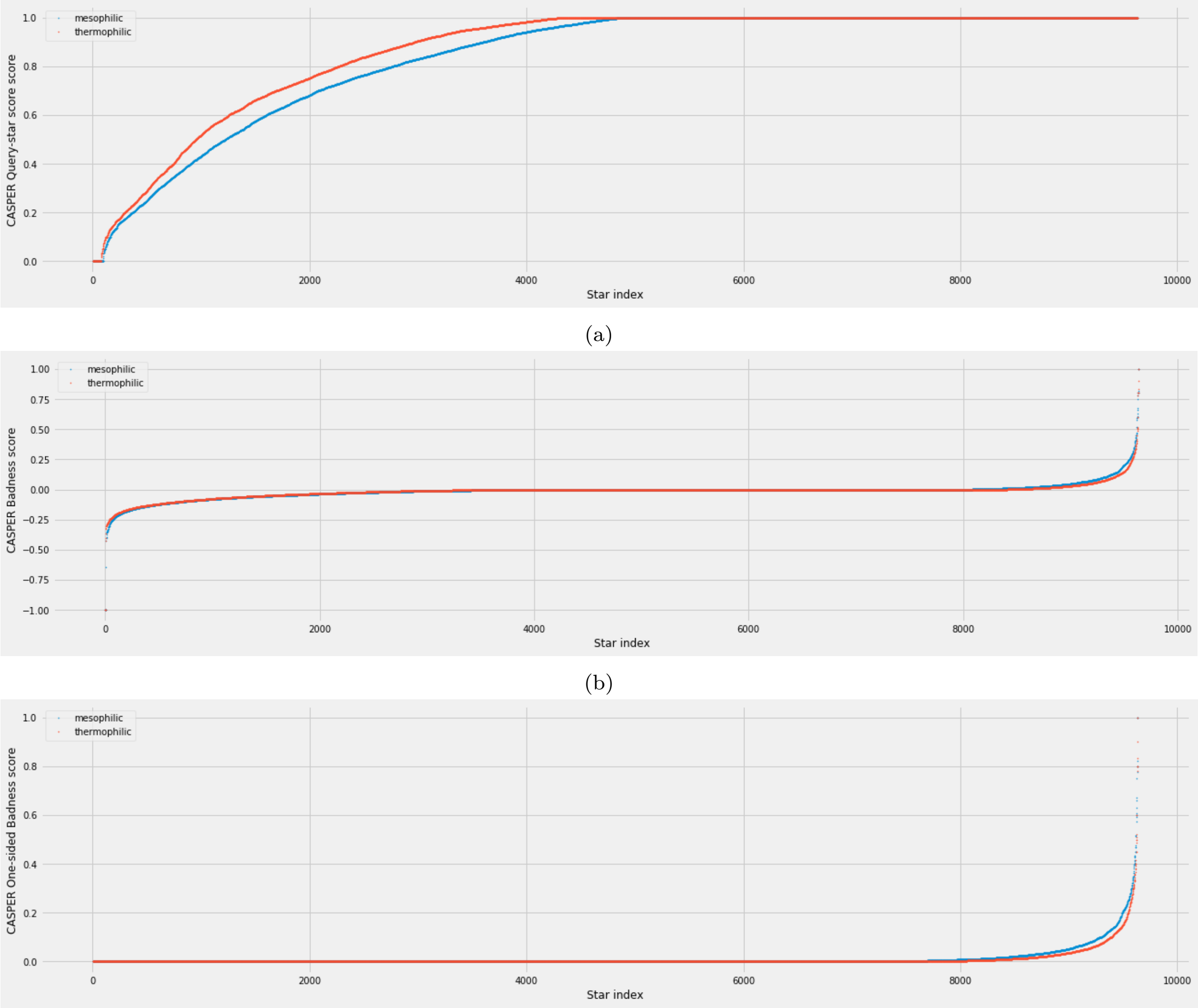
A one-tailed Mann-Whitney test was conducted to compare the set of stars from thermophilic proteins to that of the mesophilic proteins. With all three CASPER metrics, the thermophilic stars score significantly *better* than the mesophilic stars in the ProTherm sets. See Figure 11 for details.

4.2.2 Correlation with Tm values

When all the stuctures in ProTherm sets were taken together (mesophilic as well as thermophilic), there was very little correlation[14] found between any of the CASPER metrics and the scores. The maximum correlation among any of the metrics was -0.24 (with CASPER badness as a metric, and contact order cutoff of 5%)

However, when the mesophilic and thermophilic proteins are separated, the CASPER scores show correlation with the T_m values. Spearman rho for mesophilic query-star scores versus melting temperature is as high as -0.828 (p-value $= 4 \times 10^{-4}$, contact order cutoff=5%, see Table 7a). However, the same is not true for the set of thermophilic proteins. On the other hand, thermophilic badness scores correlate well with melting temperature, with Spearman rho as high as -0.791 (p-value = 0.001, contact order cutoff=5%, see Table 7b). For a detailed table of correlation between various CASPER metrics tested with the T_m, or with the difference between thermophilic and mesophilic T_m, see Appendix D

[14] Spearman rank correlation coefficient

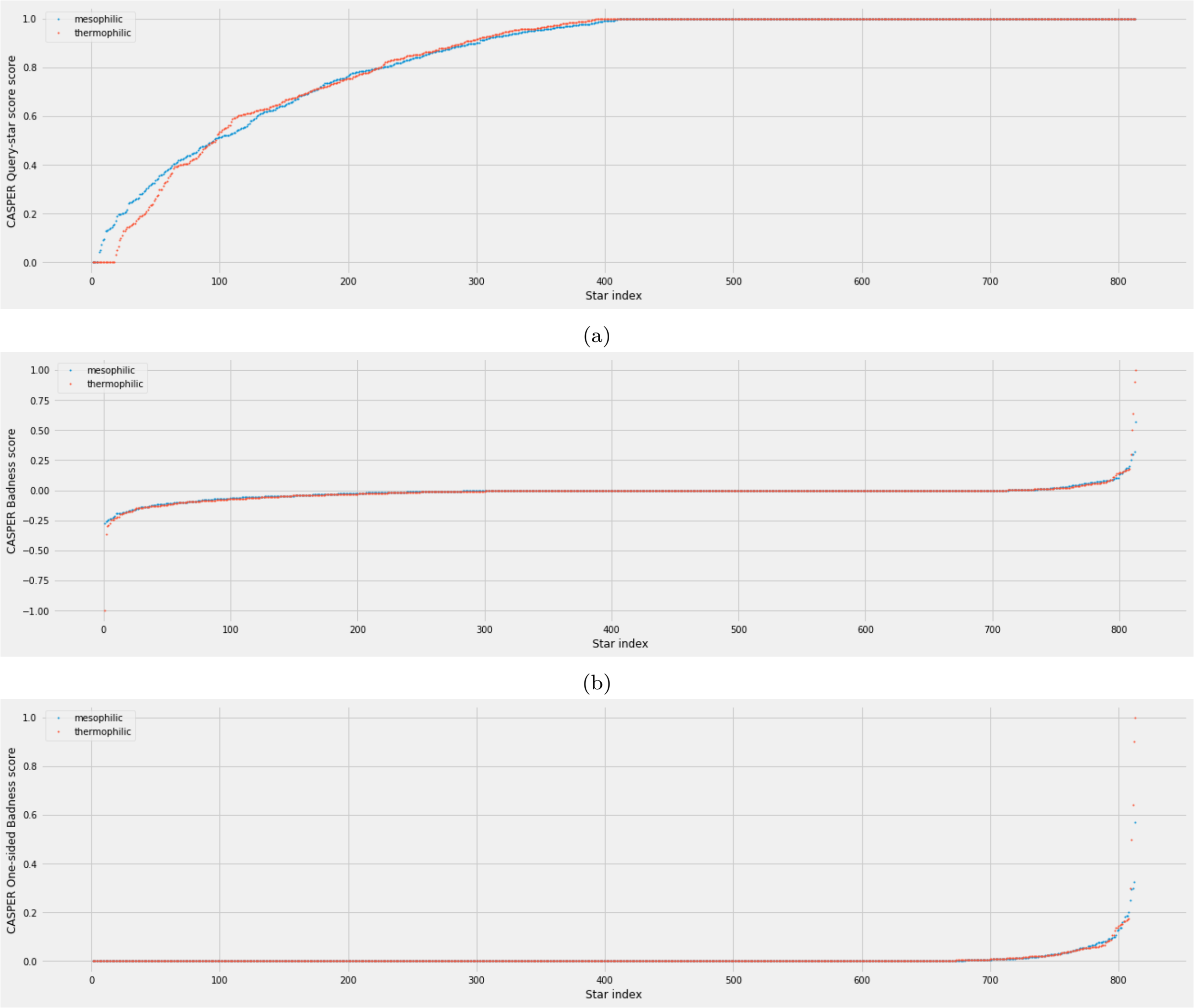


(a)

(b)

(c)

Figure 11: Rank ordered scores for thermophilic vs mesophilic proteins; proteins from ProTherm$_{lt60gt70}$ set. Scoring parameters: star-size=7, sampling-size=10000 stars for each composition; moving average window is 5% of the total number of stars.
(a) CASPER Query-star score on y-axis; One-tailed Mann-Whitney U-test (for alternative hypothesis being mesophilic scores less than thermophilic structure) p-value $=\sim 0$; (value below floating point threshold for computational calculation)
mean for mesophilic profile=0.8409, and that for thermophilic profile=0.8640
(b) CASPER Badness on y-axis; One-tailed Mann-Whitney (for alternative hypothesis being mesophilic scores greater than thermophilic structure) U-test p-value$=\sim 8.42 \times 10^{-13}$;
mean for mesophilic profile=-0.0135, and that for thermophilic profile=-0.0145
(c) CASPER one-sided badness on y-axis; One-tailed Mann-Whitney U-test (for alternative hypothesis being mesophilic scores greater than thermophilic structure) p-value$=\sim 2.11 \times 10^{-26}$;
mean for mesophilic profile=0.01160, and that for thermophilic profile=0.00940

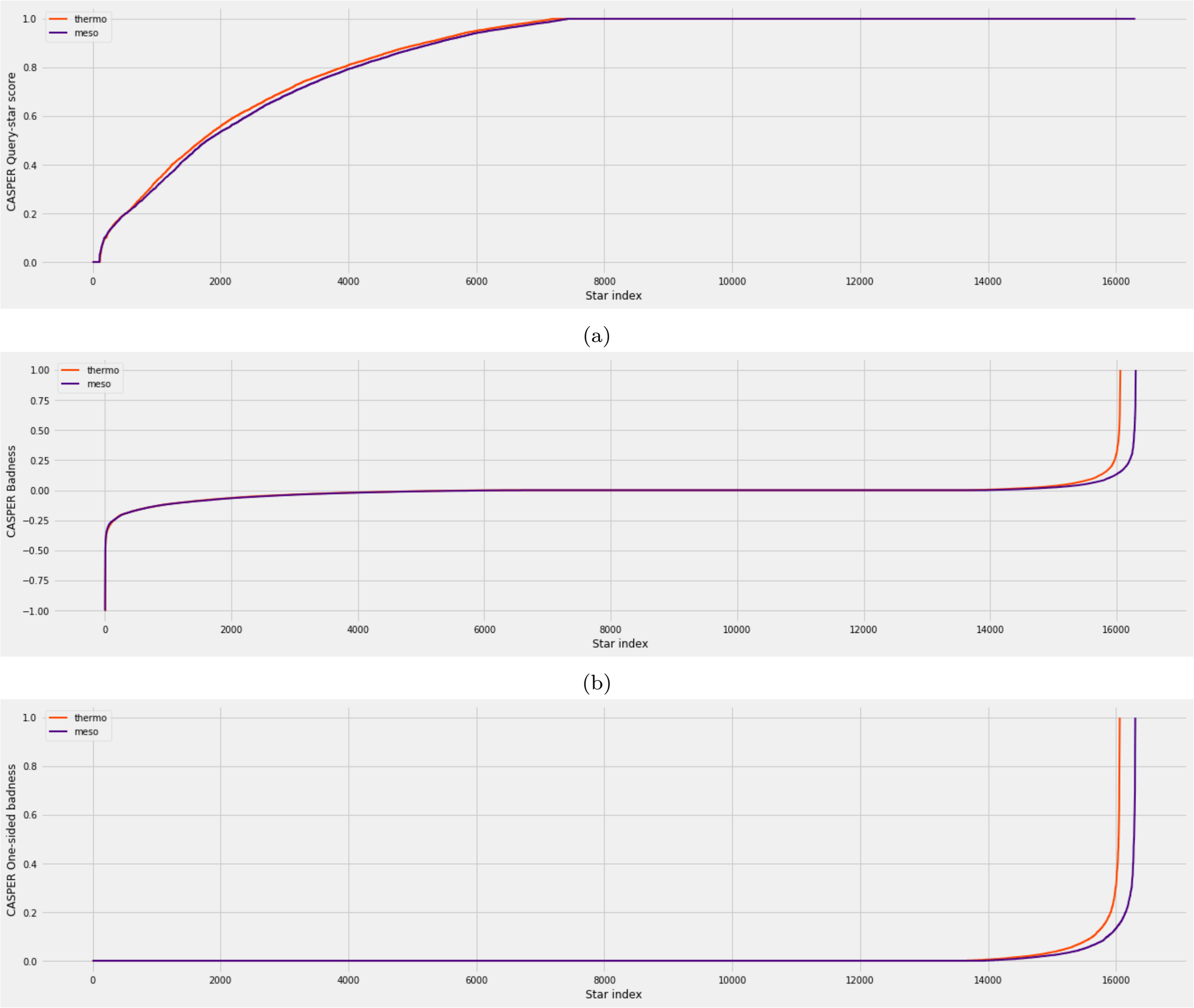


(a)

(b)

(c)

Figure 12: Rank ordered scores for thermophilic vs mesophilic proteins; proteins from ProTherm$_{lt60plus10}$ set. Scoring parameters: star-size=7, sampling-size=10000 stars for each composition; moving average window is 5% of the total number of stars.
(a) CASPER Query-star score on y-axis; One-tailed Mann-Whitney U-test (for alternative hypothesis being mesophilic scores less than thermophilic structure) p-value $=\sim 0$; (value below floating point threshold for computational calculation)
mean for mesophilic profile=0.84305, and that for thermophilic profile=0.85832
(b) CASPER Badness on y-axis; One-tailed Mann-Whitney (for alternative hypothesis being mesophilic scores greater than thermophilic structure) U-test p-value$=\sim 4.169 \times 10^{-117}$;
mean for mesophilic profile=-0.01253, and that for thermophilic profile=-0.01424
(c) CASPER one-sided badness on y-axis; One-tailed Mann-Whitney U-test (for alternative hypothesis being mesophilic scores greater than thermophilic structure) p-value$=\sim 4.07 \times 10^{-07}$;
mean for mesophilic profile=0.01005, and that for thermophilic profile=0.00942



(a)

(b)

(c)

Figure 13: Rank ordered scores for thermophilic vs mesophilic proteins; proteins from Kumar's set [Kumar et al., 2000]. Scoring parameters: star-size=7, sampling-size=10000 stars for each composition; moving average window is 5% of the total number of stars. Total number of stars in thermohilic set = 16061, and in mesophilic set=16034.
(a) CASPER Query-star score on y-axis; One-tailed Mann-Whitney U-test (for alternative hypothesis being mesophilic scores less than thermophilic structure) p-value =0.049
mean for mesophilic profile=0.85661, and that for thermophilic profile=0.86121
(b) CASPER Badness on y-axis; One-tailed Mann-Whitney (for alternative hypothesis being mesophilic scores greater than thermophilic structure) U-test p-value=0.974
mean for mesophilic profile=-0.01447, and that for thermophilic profile=-0.01410
(c) CASPER one-sided badness on y-axis; One-tailed Mann-Whitney U-test (for alternative hypothesis being mesophilic scores greater than thermophilic structure) p-value=0.883
mean for mesophilic profile=0.00883, and that for thermophilic profile=0.00905

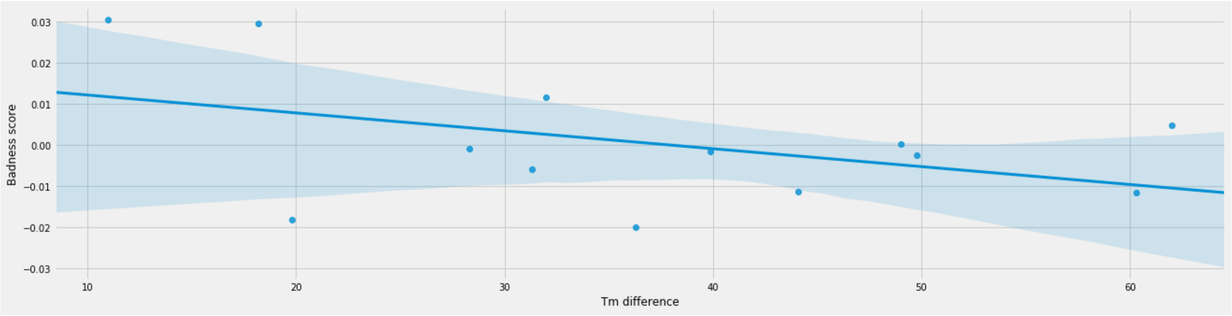
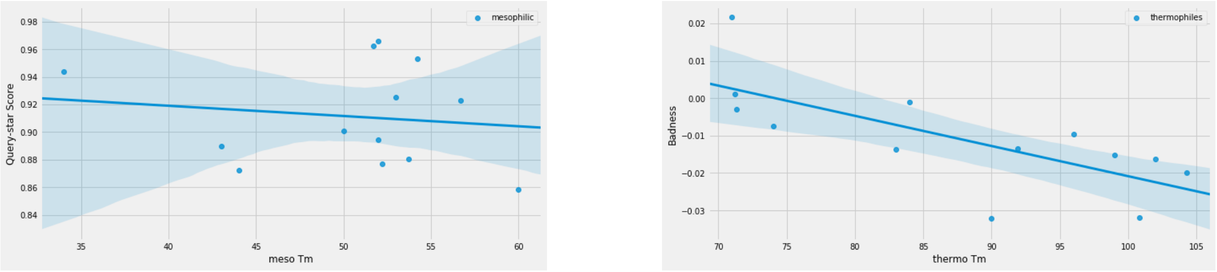


Figure 14: Correlation between T_m difference and CASPER badness scores for all proteins in ProTherm$_{lt60gt70}$ set. Spearman correlation coefficient = -0.274; p-value = 0.363
Translucent bands show 95% CI



(a) (b)

Figure 15: Correlation between T_m and CASPER scores in ProTherm$_{lt60gt70}$ set
(a) with CASPER query-star score for mesophilic proteins; Spearman rho = -0.828 (p-value = 0.0004, contact order cutoff=5%); Pearson correlation coefficient = -0.7179 (p-value = 0.0057)
(b) with CASPER badness for thermophilic proteins; Spearman rho = -0.791 (p-value = 0.001)); Pearson correlation coefficient = -0.7172 (p-value = 0.0058)

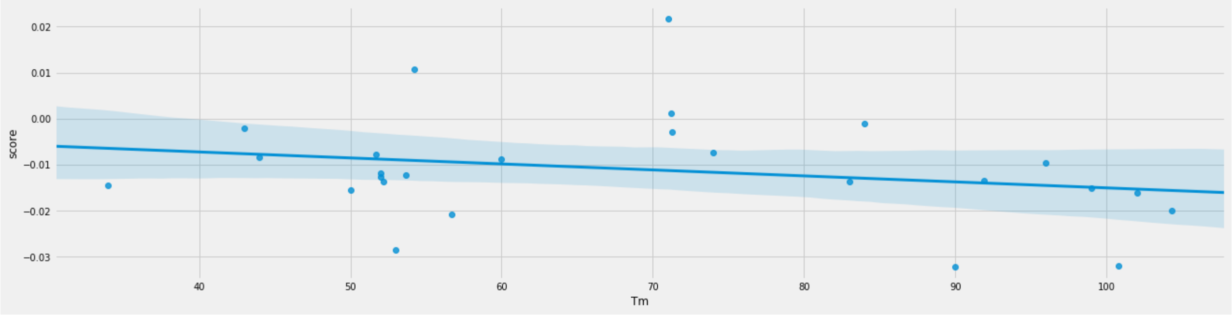
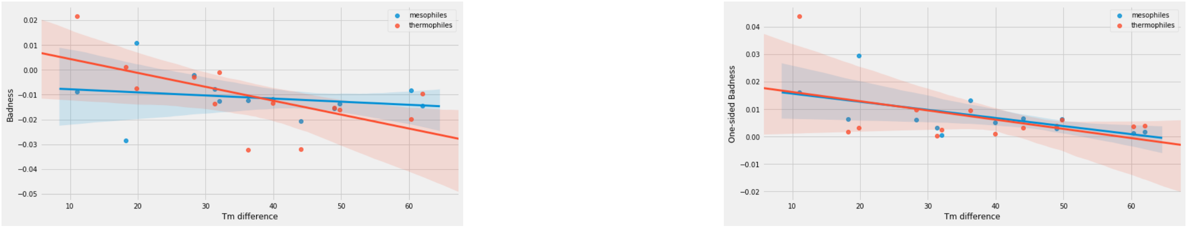


Figure 16: CASPER Query-star score versus T_m for ProTherm$_{lt60plus10}$ set. Spearman correlation coefficient = -0.249; p-value = 0.219



(a)

(b)

Figure 17: Correlation between T_m difference and CASPER scores in ProTherm$_{lt60gt70}$ set, with contact order cutoff of 5%
(a) with CASPER badness scores; Pearson correlation coefficient for thermophilic: -0.622 (p-value = 0.023)
mesophilic proteins:-0.214 (p-value = 0.482)
(b) with CASPER one-sided badness; Pearson correlation coefficient for thermophilic = -0.463 (p-value = 0.110)
mesophilic = -0.584 (p-value = 0.035)
Translucent band shows 95% CI; contact order cutoff=5% for all subfigures.

	Score Metric	meso SCC	p-value	meso PCC	p-value	Contact Order Cutoff
1	Query-star Score	-0.539	0.057	-0.317	0.292	100.0
2	Query-star Score	**-0.828**	0.0004	**-0.718**	0.006	5.0
3	Query-star Score	-0.823	0.001	-0.610	0.027	10.0
4	Query-star Score	-0.666	0.013	-0.425	0.147	20.0

(a) CASPER metric: Query-star score

	Score Metric	meso SCC	p-value	meso PCC	p-value	Contact Order Cutoff
1	One-sided Badness	0.575	0.040	0.244	0.422	100.0
2	One-sided Badness	**0.806**	0.001	0.465	0.109	5.0
3	One-sided Badness	0.699	0.008	0.353	0.237	10.0
4	One-sided Badness	0.627	0.022	0.374	0.208	20.0

(b) CASPER metric: One-sided Badness

Table 7: Correlation between T_m and CASPER scores for mesophilic proteins in ProTherm$_{lt60gt70}$ set. The extrema for the correlation values are in bold. P-value is for null hypothesis that there is no correlation based on the coefficient. For calculating p-values, the correlation coefficient was transformed into a t-statistic and the p-value was calculated by using a t-test. Correlation coefficents which are not significant enough to reject the null hypothesis at 90% CI (i.e p-value>0.1) have been marked in gray.
PCC = Pearson Correlation Coefficient
SCC = Spearman Correlation Coefficient
For other correlation coefficients and details see Appendix D

	Score Metric	thermo SCC	p-value	thermo PCC	p-value	Contact Order Cutoff
1	Badness	-0.555	0.049	-0.549	0.052	100.0
2	Badness	**-0.791**	0.001	**-0.717**	0.006	5.0
3	Badness	-0.560	0.046	-0.613	0.026	10.0
4	Badness	-0.566	0.044	-0.619	0.024	20.0

Table 8: Correlation between T_m and CASPER scores for thermophilic proteins in ProTherm$_{lt60gt70}$ set. The extrema for the correlation values are in bold. P-value is for null hypothesis that there is no correlation based on the coefficient.
PCC = Pearson Correlation Coefficient
SCC = Spearman Correlation Coefficient

	Score Metric	meso SCC	p-value	meso PCC	p-value	Contact Order Cutoff
3	One-sided Badness	-0.357	0.231	-0.429	0.143	100.0
6	One-sided Badness	**-0.566**	0.044	**-0.585**	0.036	5.0
9	One-sided Badness	-0.407	0.168	-0.427	0.145	10.0
12	One-sided Badness	-0.269	0.374	-0.511	0.075	20.0

Table 9: Correlation between T_m difference and CASPER scores for mesophilic proteins in ProTherm$_{lt60gt70}$ set. The extrema for the correlation values are in bold

4.2.3 Correlation with difference between Tm of mesophilic and thermophilic homologs

Similar to the correlation with T_m values, the difference in T_m values (i.e. thermophilic T_m minus mesophilic T_m) was also compared to the CASPER scores. A good correlation between the T_m difference and the difference in scores, may allow us to predict how much of a T_m change may occur with a certain tweak in the structure (e.g. in case of a mutation). There is little correlation between the T_m difference and the scores differences. The most correlated with the T_m difference were the Badness scores at contact order cutoff of 5%: Spearman rho = -0.274, but with a p-value of 0.363 (implying very low confidence in the correlation).

However, the badness scores for thermophilic proteins correlates well with the difference in T_m; Spearman rho is -0.670 (p-value = 0.012) at contact order cutoff of 5%. For mesophilic proteins, the one-sided badness scores have a Spearman rho of -0.566 (p-value = 0.044) at contact order cutoff of 5%. See Figures 17a, 17b for regression plots for the same.

4.2.4 Correlation with Temperature Factors

As mentioned before, a flexible region of a protein will have higher temperature factors compared to more rigid, buried regions. Further, it's also known that the B-factors follow a bi-modal Gaussian distribution, corresponding to buried vs exposed residues[Parthasarathy and Murthy, 2008]. A more flexible region in a protein is expected to have more geometrical variants present in the PDB, compared to a less flexible region. This means that finding a good, low RMSD match, for stars from such region is tougher. With this reasoning, we tried to look for a correlation between the B-factor values of chemical groups and the corresponding scores that they get. There is no method described in literature about how to compare structures, and therefore, if we find a correlation, we can use our scores as a proxy for B-factors, and therefore be able to compare B-factors across structures.

The depth cutoff is to differentiate between the two parts of the bimodal B-factor distribution. Z-score of atomic B-factor was calculated w.r.t the normal distribution that an atom is a part of, and the mean of these Z-scores was taken as the B-factor for the chemical group.

The Spearman rank-correlation coefficients (SCC) for the different strategies are described in Table 2. See 3.3.3 for a discussion on the strategies used for converting atomic B-factor values to corresponding values for chemical groups. In none of the strategies could we find a good SCC.

Strategy	**SCC**
Mean of atomic B-factors	0.017
Percentile rank of atomic B-factors	0.009
Z-score	-0.012
Depth based Z-score (depth cutoff = 6Å)	-0.004
Depth based Z-score (depth cutoff = 7Å)	0.003
Depth based Z-score (depth cutoff = 8Å)	-0.006

Table 10: SCC of chemical group wise scores vs chemical group wise B-factors, for different strategies of calculating chemical group wise B-factor values

5 Discussion and Further Work

5.1 Summary

5.1.1 Obsolete PDB refinement study

Our primary objective was to find a correlation between RMSD of refinement and local scores of the stars. This would allow us to determine which regions in a protein need to be targetted first for refinement, compared to others. However, we weren't able to find a good correlation in this regard.

However, when all the stars from the whole dataset are pooled together, the CASPER metrics were able to distinguish between the stars taken from obsolete and successor structures with very good confidence. Note that the statistical test employed for the same is a one-tailed Mann-Whitney test. Since this is a non-parametric test, the mean or the parameters of the distribution of the scores don't matter. What does matter however, is whether the results have more favorable outcomes for the alternative hypothesis, than unfavorable ones[15], when the scores are rank-ordered.

It's also important to interpret the test results correctly. We cannot conclude that the scores are *significantly higher* for one compared to the other. All we can say is that if a random pair of stars is picked up, one each from the refined and the obsolete sets, the refined has a very high probability of scoring better with any of the CASPER metrics.

The other key result is that the percentage of structures identified correctly as refined structures vs. the percentage identified incorrectly (Figure 10), increases as we decrease the maximum similarity of composition between the stars being compared. This means that perhaps the residue environments change drastically (captured in terms of composition of stars) as we go about progressively refining a structure. This is also noticeable in the score profiles in Figures 6-8. As the similarity increases, the score profiles become increasingly hard to distinguish. In fact when the composition doesn't change at all (similarity = 100%), even the Mann-Whitney test result is not significant, with a p-value greater than 0.1.

5.1.2 Correlation with Thermal Fluctuations

Three different datasets were used for comparing how thermophilic structures score compared to their mesophilic counterparts. Of the three, we contend that Kumar's set and Szilágy's set are unreliable at least for the purpose of the study conducted here. In case of Kumar's set, there is no objective metric to distinguish the proteins themselves, and in case of Szilágy's set, T_{opt} of the source organism is taken as the metric. This is understandable, since there are relatively few studies where thermodynamic data for both mesophilic and thermophilic homologs are available. More importantly, even if they are available, the corresponding structures may not be available.

We culled relavant records from the ProTherm database in terms of T_m values, for this purpose. Although this results in a much smaller dataset, it's perhaps more reliable since the structural stability of the protein would determine its unfolding rate, and consequently its melting temperature. T_{opt} of the source organism, in comparison, is a bad metric since the protein may be very thermostable but still exist in a mesophilic organism with low T_{opt}. This is perhaps the reason why the scoring scheme performs badly in case of Szilágy's set. In the other sets, the efficiency is higher. With Kumar's set, in as much as ~72%

[15] i.e. whether the refined structures score better than the unrefined ones or not

(13 out of 18 pairs) of the pairs, the thermophilic structure scored better according to our expectations (Table 6).

In this regard, the expectation was that with lower CO cutoffs, the efficiency[16] of identifying the thermophilic structures would increase. Although the best efficiency in case of ProTherm datasets was when the CO was taken into consideration, there doesn't seem to be any trend as such as we go about changing the CO cutoff.

However, the best correlations between CASPER scores and mesophilic or thermophilic T_m is actually at lower CO cutoffs (e.g. Table 7a). In fact in Table 13, note that even though different CASPER metrics correlate better or worse with mesophilic and thermophilic T_m, the best correlation is when the CO cutoff was the lowest among all the values that were tried. Perhaps a weighted mean (instead of an absolute mean) makes more sense for the net-score of a structure. The star-wise scores are probably not additive in a direct manner, and a better approximation would be to weight the scores in terms of CO of the star, depth of the star, etc. See Section 5.2 for a discussion on additivity of scores.

Ideally, we would like to find a good correlation between the net score of the structure and the T_m for the same. We can then perhaps use the information to predict T_m using wild-type or mutated structures, and be able to design thermostable mutants for known structures. Such a correlation was not found when all the structures from the ProTherm Datasets were taken into consideration (e.g. Figure 14).

However, when the mesophilic and thermophilic structures are considered separately, the scores correlate extremely well with T_m values. Although this is a useful and non-trivial result, it can't be used at least at this stage for predicting T_m values. We need to find a threshold score value, or some other method of determining protein thermophilicity. In light of this purpose, the fact that stars from thermophilic structures score better than mesophilic structures in ProTherm datasets, may prove to be useful. We expect thermophilic structures to contain specific residue environments that can stabilise them at environments with high temperatures. In fact, the studies from which Kumar's set and Szilágy's set were extracted [Kumar et al., 2000, Szilágyi and Závodszky, 2000] provide ample evidence of differing structural elements between thermophilic and mesophilic structures. We believe that these structural elements have implicitly been added to the stars that have been sampled in the PDB database, and we intend on discovering these stars and their chemical group compositions in the future.

5.2 The approximation of additivity

One may argue that such an evaluation method is similar to a knowledge based *potential energy function* or a pseudo-free energy function, since the closer a local atomistic packing is to that found in experimentally solved native structures, the lower is the energy for that conformation of atoms. Keeping this in mind, we have made the approximation of additivity, by taking the mean of the star-wise scores as a measure of the score of the whole protein. This may be a bad approximation, since the free energy contribution of two or more phenomena (in this case, the occurrence of stars w.r.t to each other) can be added up only if they are independent events [Dill, 1997]. This is obviously not true, and it's possible that a bad star has multiple good stars (in terms of CASPER scores) as neighbours, which contribute to stabilising the bad star.

[16] number of thermophilic structures identified correctly divided by size of dataset

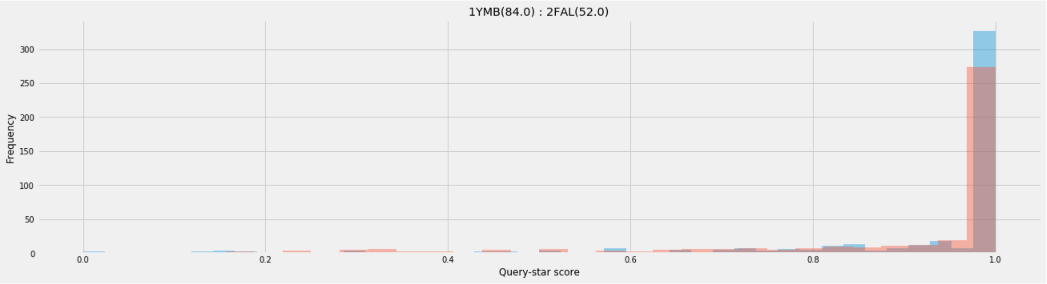


Figure 18: Example of a thermophilic protein (Myoglobin) and mesophilic homolog; Frequency of stars versus Query-star score. The T_m values are in $^{\circ}C$, mentioned in brackets for each of the structures, along with their PDB IDs. The trend for a large majority of stars having a score of 1.0 is there across all structures that were compared.

The above mentioed problem is faced by most knowledge-based potentials in general [Dill, 1997]. The advantage in CASPER's case however, is that the evaluation is local in nature. As a result, in the scenario wherein CASPER is used for structure refinement or structure prediction, one may work on increasing the number of stars that score well in the whole protein structure, instead of striving to increase the net score of the protein.

However, we do need to find a better way to find net-score of a protein structure model, since not all stars contribute equally to the stability (evident from the change in results when CO is taken into consideration). The simplest example is the existence of a hydrophobic rigid core in case of most globular proteins. The stars found at higher residue depths for example, are known to be involved more in stabilising the protein [Chakravarty and Varadarajan, 1999, Tan et al., 2013], and should be given more weight than the ones which are exposed.

5.3 Topology independence

The work done here takes inspiration from previous work on topology independent structural superimposition [Nguyen et al., 2011, Nguyen and Madhusudhan, 2011], which suggests that a superimposition performed without explicitly adding information of the primary or secondary structure folds of a protein structure may provide better superimposition, and also provide valuable insights into the kind of amino acid residue environments in the protein. Two proteins with completely different folds may have similar residue environments atomistic packing in terms of geometry, which is usually overlooked by traditional structure superimposition methods. CASPER is therefore different from other structure evaluation methods that consider structural motifs because it can compare regions of protein structures that may not be part of similar secondary or super-secondary structures.

The usage of a non-redundant PDB database prevents trivial solutions such as finding matches for the query-star in homologous structures. Thus, for whatever representation that is found for any of the query-stars, it can be stated that the match is because of the representation of similar residue environments in unrelated protein structures. Quite often one finds very similar matches for most stars in any given PDB. This could either mean that the parameters for CASPER need to be optimised further, and they may be very lenient and are therefore able to find close matches, or that the PDB is extremely redundant in terms of the representation of structural motifs.

5.4 Structure refinement

A major problem with current structure refinement methods, is the problem of initial direction of refinement. Without the help of good local evaluation metrics, it is hard to decide which regions require refinement more than others [Feig, 2017, Park and Seok, 2012]. This is the reason behind the conception of the CASPER *badness* metrics, since these may allow us to answer the question of *how bad is a badly packed region*. With a simple RMSD metric for example, since RMSD can be any positive number, there is no way to know how much of an RMSD value is a bad score.

Current structure refinement methods usually use a combination of molecular dynamics (MD) simulations and knowledge-based statistical potentials. The statistical potential helps in providing restraints for the MD, so that the protein structure doesn't unfold away from the native state. Additionally, in case of hybrid refinement methods that sample conformational space using Monte Carlo or Normal Mode Analysis (NMA) based sampling, the potential energy function can provide direction in which the sampling may be performed, with a suitable set of moves at every iteration [Feig, 2017]. However, such restraints may prevent the initial models from sampling conformations that are farther away in RMSD space. One may simulate an *annealing* process for this purpose, but annealing is often a random process. The conformational sampling that one is actually trying to simulate in such a scenario, is the effect of chaperone-mediated protein folding, which is not a very random process, and is well controlled. We believe that CASPER may provide a reasonable solution to this problem. For a star to move away from it's residue environment, one may look for similar stars with a certain degree of similarity cutoff that may be formed by the use of other chemical groups in the vicinity. This not only allows for a simulation of the annealing process, but may also be guided by native-like stars that have been observed in the PDB already.

At this stage, to a certain extent we are able to distinguish between refined and obsolete PDB structures. There is certainly a difference in trends of the scores when refined structures are compared to their predecessor counterparts. However, whether there is a correlation between the extent of refinement and the badness of the obsolete structure (or any part of it), remains to be seen. We intend on developing a method for structure refinement based on the evaluation of regions provided by the CASPER metrics.

5.5 The CASPER metrics

The CASPER query-star score consistently has shown good performance, but lacks the objective threshold that CASPER badness provides. We cannot say conclusively whether a certain region requires refinement. All we can do is point out which regions have stars that score bad, and therefore need to be targetted first. Nonetheless, the fact that the query-star score metric is able to discern the trends even in Kumar's set (which doesn't have T_m values), encourages us to work further using that metric.

One of the reasons why CASPER badness may be performing poorly is perhaps the way the sisterly set is constructed. The set is constructed w.r.t the query-star, and this gives it an advantage in terms of score, compared to the superstar. Perhaps a reverse metric of constructing the sisterly set w.r.t the superstar may be a better idea. This is because the CASPER metric in general is an attempt at approximating the selection of stars which are similar in geometry. Since the superstar is already native-like, it should provide a better approximation since it is closer to the rest of the stars in terms of geometry. However, this is computationally more time-consuming (almost twice as much). Nevertheless, such a modified method is currently being implemented for testing.

5.6 Knowledge based potentials

The protein structure evaluation method discussed in this text is an example of a knowledge-based method, although it is unlike other knowledge-based methods. Most knowledge-based methods can be categorised as what are known as statistical potential energy functions. Statistical potentials are attractive because they usually consume less resources in terms of computational power and in terms of time, in comparison to *ab initio* physics-based methods. A typical statistical potential would have a component of evaluating what was the probability of occurrence of a certain arrangement of atoms or amino acids in 3D space, otherwise referred to as an *expected probability*. This is then compared to an *observed probability*, which is derived from known protein structures. Work on empirical methods such as these are therefore focussed on optimising parameters related to observed structural features, or the reference system for calculating expected probabilities [Shen and Sali, 2006, Zhou and Zhou, 2002].

Although improved reference states have been quite successful in protein structure evaluation (and as a result, for protein structure prediction), they are still approximations to the actual expected probabilities of the conformations in the protein structure. The method that we have presented here overcomes this difficulty by getting rid of estimating such an expected probability measure. Instead, we are looking at how well represented in the PDB, is a certain structural region from any protein. This is close to a probability measure of the same. However, we are also considering how close is the query structure to any previously known structures deposited in the PDB. In doing so, we have made the assumption that the PDB is almost complete [Fernandez-Fuentes et al., 2010] in terms of our knowledge local structural features of proteins[17], and that they contain close to native structures of proteins.

5.7 Related work

5.7.1 Charge dynamics dependent force field development

Classical MD force fields have fixed atomic parameters even for charges. An alternative is the usage of polarisable force fields, which are semi-empirical and are quite nascent and inefficient for simulation of macromolecules [Lopes et al., 2013, Ponder et al., 2010]. The concept of a star of chemical groups allows us to look at topology independent charge dynamics of local regions in the protein, based on the geometry of the atoms in space as well as their residue environment. This is the basis for related work being carried out, wherein we are trying to simulate charge dynamics in stars of chemical groups, but represented in terms of constituent atoms. With the knowledge of how charges change during the dynamics of protein folding, unfolding, or partial folding will not only allow us to improve existing force fields, but also allows us a way to predict pKa of amino acids better (since pKa can change with change in partial charges for the atoms in an amino acid).

[17] which means that they contain all possible structural motifs that could have been explored

5.7.2 PackPred mutation predictor

Packpred is a software developed earlier in the research group. It predicts the functional consequences of point mutations in proteins, starting with a 3D structure, or by modelling the structure. It combines residue environment information using DEPTH [Chakravarty and Varadarajan, 1999, Tan et al., 2013] and CLICK [Nguyen et al., 2011]. The underlying concepts in this study and Packpred are similar, especially the usage of structural motifs (in CLICK's case as *cliques* of atoms) for creating a multibody statistical measure.

Benchmarking of Packpred server (http://cospi.iiserpune.ac.in/packpred/) with other mutation predictors/evaluators were also carried out as part of the fifth year project. This includes testing the Packpred server and standalone softwares, and ensuring that they run smoothly.

References

[Anfinsen, 1972] Anfinsen, C. B. (1972). The formation and stabilization of protein structure. Biochemical Journal *128*, 737–749.

[Bava et al., 2004] Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. and Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic acids research *32*, 120D–121.

[Berman Helen M, 2000] Berman Helen M, W. J. F. Z. G. G. B. T. N. W. H. S. I. N. B. P. E. (2000). The protein data bank. Nucleic acids research *28*, 235–242.

[Boyle, 2008] Boyle, J. (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 *36*, 317–318.

[Chakravarty and Varadarajan, 1999] Chakravarty, S. and Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. Structure *7*, 723–732.

[Chen and Kihara, 2011] Chen, H. and Kihara, D. (2011). Effect of using suboptimal alignments in template-based protein structure prediction. Proteins *79*, 315–334.

[Dill, 1997] Dill, K. A. (1997). Additivity principles in biochemistry. The Journal of Biological Chemistry *272*, 701–704.

[Feig, 2017] Feig, M. (2017). Computational protein structure refinement: almost there, yet still so far to go. Wiley Interdisciplinary Reviews: Computational Molecular Science *7*, e1307.

[Fernandez-Fuentes et al., 2010] Fernandez-Fuentes, N., Dybas, J. M. and Fiser, A. (2010). Structural Characteristics of Novel Protein Folds. PLOS Computational Biology *6*, 1–11.

[Grantcharova et al., 2001] Grantcharova, V., Alm, E. J., Baker, D. and Horwich, A. L. (2001). Mechanisms of protein folding. Current Opinion in Structural Biology *11*, 70–82.

[John, 2003] John, B. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Research *31*, 3982–3992.

[Kearsley, 1989] Kearsley, S. K. (1989). On the orthogonal transformation used for structural comparisons. Acta Crystallographica Section A *45*, 208–210.

[Kumar et al., 2000] Kumar, S., Tsai, C.-J. and Nussinov, R. (2000). Factors enhancing protein thermostability. Protein Engineering, Design and Selection *13*, 179–191.

[Leopold et al., 1992] Leopold, P. E., Montal, M. and Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proceedings of the National Academy of Sciences of the United States of America *89*, 8721–8725.

[Lopes et al., 2013] Lopes, P. E. M., Huang, J., Shim, J., Luo, Y., Li, H., Roux, B. and Mackerell, A. D. (2013). Force Field for Peptides and Proteins based on the Classical Drude Oscillator. Journal of Chemical Theory and Computation *9*, 5430–5449.

[Meruelo et al., 2012] Meruelo, A. D., Han, S. K., Kim, S. and Bowie, J. U. (2012). Structural differences between thermophilic and mesophilic membrane proteins. Protein Science *21*, 1746–1753.

[Nguyen and Madhusudhan, 2011] Nguyen, M. N. and Madhusudhan, M. S. (2011). Biological insights from topology independent comparison of protein 3D structures. Nucleic Acids Research *39*, e94—-e94.

[Nguyen et al., 2011] Nguyen, M. N., Tan, K. P. and Madhusudhan, M. S. (2011). CLICK—topology-independent comparison of biomolecular 3D structures. Nucleic Acids Research *39*, W24–W28.

[Nowlin et al., 1988] Nowlin, D. M., Bollinger, J. and Hazelbauer, G. L. (1988). Site-directed mutations altering methyl-accepting residues of a sensory transducer protein. Proteins: Structure, Function, and Genetics *3*, 102–112.

[Pack and Yoo, 2004] Pack, S. P. and Yoo, Y. J. (2004). Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. Journal of Biotechnology *111*, 269–277.

[Park and Seok, 2012] Park, H. and Seok, C. (2012). Refinement of Unreliable Local Regions in Template-based Protein Models. Proteins: Structure, Function, and Bioinformatics *80*, n/a–n/a.

[Parthasarathy and Murthy, 2008] Parthasarathy, S. and Murthy, M. (2008). Analysis of temperature factor distribution in high-resolution protein structures. Protein Science *6*, 2561–2567.

[Perez et al., 2016] Perez, A., Morrone, J. A., Brini, E., MacCallum, J. L. and Dill, K. A. (2016). Blind protein structure prediction using accelerated free-energy simulations. Science Advances *2*, e1601274.

[Plaxco et al., 1998] Plaxco, K. W., Simons, K. T. and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins 1 1Edited by P. E. Wright. Journal of Molecular Biology *277*, 985–994.

[Ponder et al., 2010] Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio, R. A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E. and Head-Gordon, T. (2010). Current status of the AMOEBA polarizable force field. The Journal of Physical Chemistry. B *114*, 2549–2564.

[Schwede, 2013] Schwede, T. (2013). Protein Modeling: What Happened to the “Protein Structure Gap? Structure *21*, 1531 – 1540.

[Shaw et al., 2009] Shaw, D. E., Bowers, K. J., Chow, E., Eastwood, M. P., Ierardi, D. J., Klepeis, J. L., Kuskin, J. S., Larson, R. H., Lindorff-Larsen, K., Maragakis, P., Moraes, M. A., Dror, R. O., Piana, S., Shan, Y., Towles, B., Salmon, J. K., Grossman, J. P., Mackenzie, K. M., Bank, J. A., Young, C., Deneroff, M. M. and Batson, B. (2009). Millisecond-scale molecular dynamics simulations on Anton. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC ’09 p. 1, ACM Press, Portland, Oregon.

[Shen and Sali, 2006] Shen, M.-y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Science *15*, 2507–2524.

[Szilágyi and Závodszky, 2000] Szilágyi, A. and Závodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure *8*, 493–504.

[Tan et al., 2013] Tan, K. P., Nguyen, T. B., Patel, S., Varadarajan, R. and Madhusudhan, M. S. (2013). Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Research *41*, W314–W321.

[Tanaka and Scheraga, 1976] Tanaka, S. and Scheraga, H. A. (1976). Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. Macromolecules *9*, 945–950.

[Wang and Dunbrack, 2003] Wang, G. and Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591.

[Webb and Sali, 2016] Webb, B. and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER: Comparative Protein Structure Modeling Using Modeller. In Current Protocols in Bioinformatics, (Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D. and Yates, J. R., eds), pp. 5.6.1–5.6.37. John Wiley & Sons, Inc. Hoboken, NJ, USA.

[Worth et al., 2011] Worth, C. L., Preissner, R. and Blundell, T. L. (2011). SDM–a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Research *39*, W215–222.

[Zhou and Zhou, 2002] Zhou, H. and 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*, 2714–2726.

Appendices

A Method and Framework Developed Earlier

The stars database: The database being used here is a pre-computed database. We have parsed the PDB database to chemical group format and done a search of all stars in all the proteins. We are using the set of all protein structures recorded in the PDB; except the ones which have multiple models (NMR structures for example), CA-only structures, and structures with large number of missing residues and/or atoms (in which case we can't parse it to chemical group representation). The total number of structures under these criteria is 115,915 structures. However, since the PDB database is redundant, we use only a non-redundant (NR) subset of it(culled using list of sequence similarity clusters in PDB database, as of July 20, 2018), such that pairwise sequence similarity is always less than 30%. The NR-30 subset is 25373 structures large in size.

Pre-computation: As described later, the way we store the information of all the stars allows us to not store the coordinates of the stars but look them up whenever necessary. The stars are indexed based on their composition of chemical groups for faster access.

Parallelisation and sampling: While scoring the test protein on the computing cluster, each query star is assigned to one core. For each star, a random sample of stars (sample size is say, S) are picked with the same composition from the star database. The reason for this is because:

1. We don't need to superimpose against all the stars in the database, but only the ones with same composition of chemical groups.

2. It is a practical limitation that the superimposition of the query star against all stars with same composition in the database is not fast enough.

3. In a lot of cases, for example in case of alpha-helices, the stars in the database are redundant (and their population is too large). For stars derived from such folds, it's perhaps unnecessary to superimpose exhaustively. Instead, a smaller random sample serves as a representative for the whole set of stars which have the same composition.

The group-pdb format: The protein structures have been defined in terms of the chemical groups (we call the new format a *.gpdb* format for group-pdb). The formatting of the text is standard *.pdb* itself, so that the lines can be read by programs which use *.pdb* format.

The star-finding job is done by finding the nearest neighbours for each chemical group, using a cKD-Tree algorithm. The program returns a list of stars, when the *.gpdb* file along with the size and distance cut-off are given as arguments; the output file (*.cliq* extension) is formatted such that it has indices of the groups written into each line, and each line is a star. The corresponding structure is also written out as a *.gpdb* file for later superimposition. When coordinates are required for superimposition, one can map the indices of the chemical groups in the star to that in the *.gpdb* file it was extracted from.

Superimposition by centre-to-centre mapping: The superimposition is done by using the Kearsely algorithm [Kearsley, 1989]. We map the central chemical group of query star to the central chemical group of the target star. (Note that there is always one centre in every star and this is the chemical group with respect to which the star was defined.) Thereafter, superimpositions for all possible permutations of mappings of chemical groups of same identity (e.g. r1 to r1,r2 to r2, etc.) is done and lowest RMSD permutation is the one which is taken as the RMSD of superimposition.

Sample lines from a .gpdb file, for illustration purposes:

```
ATOM     40 r2   CYS A  11       9.463   2.028   7.521
ATOM     41 r10  CYS A  11       7.856   2.863   7.659
ATOM     42 r1   ILE A  12      12.778  -2.378   7.626
```

Sample lines from a .cliq file, for illustration purposes:

```
1cxs        4       5       0       6       7       2       1       9
1cxs        5       6       4       10      11      7       68      9
1cxs        6       7       9       5       8       68      4       69
```

PDB ID	Sampling-size S	Rank of native model
1onc	2000	1
1onc	4000	1
1bbh	2000	1
1bbh	4000	1
1c2r	1000	1
1c2r	2000	1
1cau	1000	7
1cau	2000	18
1cau	4000	3
2pna	1000	215

Table 11: Native model rankings out of decoy set size with 300 decoys. Penalty P=5 Å; Clique size N=8; RMSD as score metric

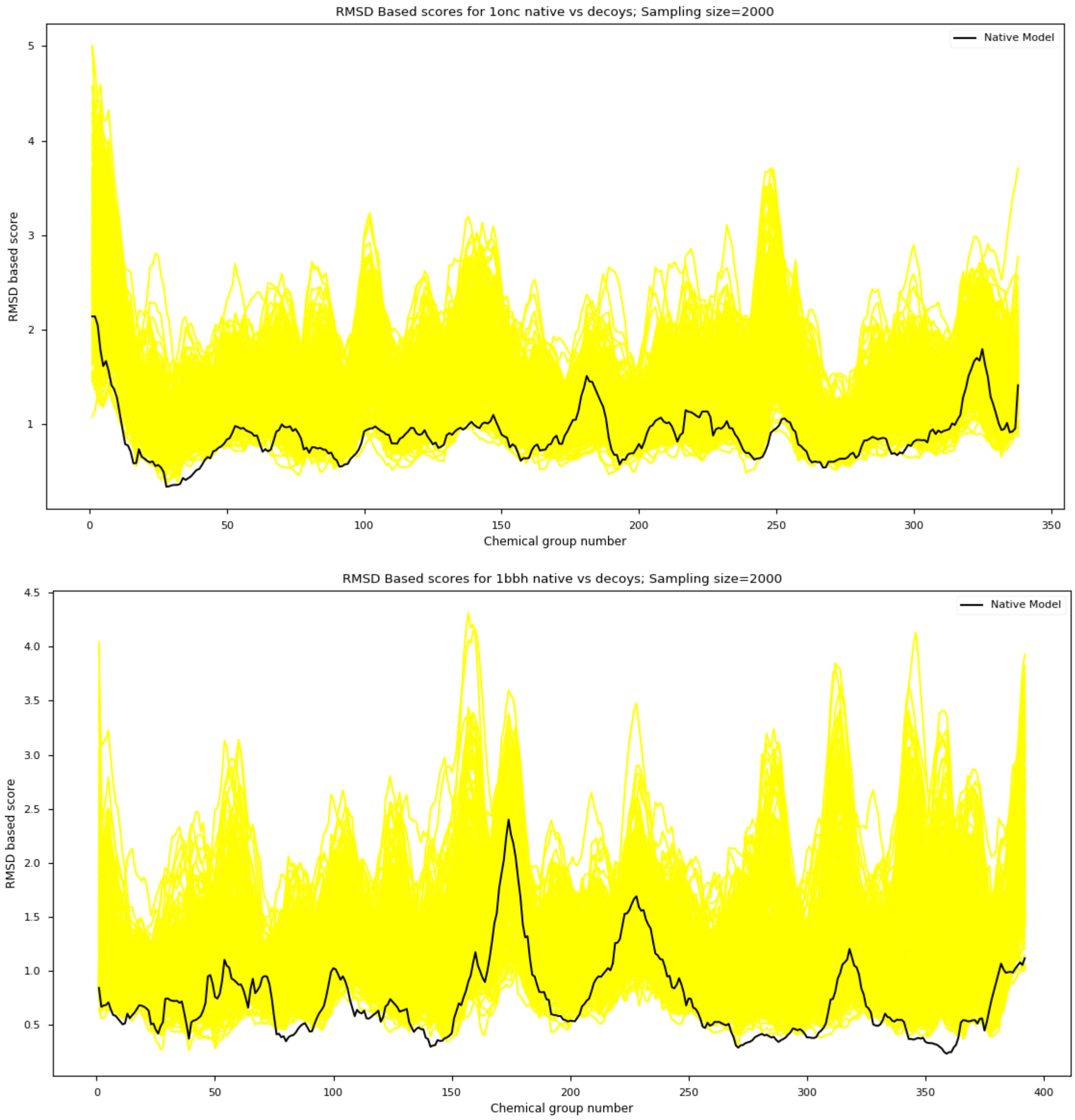


Figure 19: Star-wise best-match RMSD score profiles for 1onc and 1bbh native model(black line) vs decoys (300 for each protein) from Moulder decoy set[John, 2003]. Note how the native scores lower than the decoys in most of the stars

B Identification of Native Model In A Decoy Set

index	PDB ID	CASPER Query-star score	Superstar RMSD	CASPER One-sided badness	CASPER Badness
0	1bbh	1	1	1	2
1	1c2r	1	1	1	9
2	1gky	1	1	1	1
3	1eaf	1	1	1	1
4	2cmd	1	1	1	1
5	1onc	1	1	1	12
6	1mdc	1	1	2	2
7	2sim	1	1	1	9
8	2afn	1	1	1	1
9	2fbj	1	1	1	1
10	1cau	2	1	1	3
11	1cew	1	1	1	2
12	1cid	2	3	3	5
13	1dxt	1	1	1	1
14	1lga	1	1	1	9
15	1mup	3	1	4	3
16	2mta	1	1	1	4
17	2pna[1]	3	17	9	18
18	4sbv	1	1	1	1
19	8i1b	1	1	1	10
	Fraction with native model as rank=1	**16/20**	**18/20**	**16/20**	**7/20**

Table 12: Native model ranking in Moulder decoy set. Rank is out of 31, for each protein.

[1] NMR structure

C Cleaning up and culling of obsolete PDB records

C.1 Completely obsolete records

Records where there is a missing successor for the obsolete PDB entry, were ignored for this study.

For example, there can be entries like:

```
ENTRY     DATE      OBSLTE   SUCCSSR

OBSLTE    24-JUL-07 1F83     3G94
OBSLTE    14-JUL-09 3G94
```

3G94 was removed as obsolete because of paper retraction!

C.2 Outdated records and multiple successors

Sometimes there can be outdated obsolete PDB records,e.g.

```
OBSLTE    30-OCT-78 151C     251C
OBSLTE    02-OCT-81 251C     351C
```

That is, they haven't been updated to be written as:

```
OBSLTE    30-OCT-78 151C     251C     351C
```

At other times, the update has been made, so there are multiple successors for the same obsolete record

```
OBSLTE    18-JUL-84 1HHB     2HHB     3HHB     4HHB
```

In both cases, the final structure is taken as the refined structure, and everything else as obsolete entries paired with it. So if A is replaced by B, which is then replaced by C and then D – we get three obsolete-successor pairs, namely: A-D, B-D and C-D

C.3 Culling parameters

A non-redundant subset with the following characteristics was used for the obsolete PDB refinement study. Culling for this subset of the PDB was done using PISCES server[Wang and Dunbrack, 2003].
Percentage sequence similarity <= 30%
Resolution <1.8 Å
R-factor: 0.25
exclude non-X-ray, and exclude CA-only structures

D Correlation between T_m or T_m difference and CASPER scores

	Score Metric	meso SCC with T_m diff	meso SCC with T_m diff pval	thermo SCC with T_m diff	thermo SCC with T_m diff pval	meso PCC with T_m diff	meso PCC with T_m diff pval	thermo PCC with T_m diff	thermo PCC with T_m diff pval	meso SCC with T_m	meso SCC with T_m pval	thermo SCC with T_m	thermo SCC with T_m pval	meso PCC with T_m	meso PCC with T_m pval	thermo PCC with T_m	thermo PCC with T_m pval	Contact Order Cutoff
1	Query-star Score	0.412	0.162	-0.011	0.972	0.352	0.238	-0.007	0.983	-0.539	0.057	-0.066	0.831	-0.317	0.292	-0.124	0.686	100.0
2	Badness	0.099	0.748	-0.489	0.090	-0.125	0.684	-0.490	0.089	-0.121	0.694	-0.555	0.049	-0.047	0.878	-0.549	0.052	100.0
3	One-sided Badness	-0.357	0.231	0.033	0.915	-0.429	0.143	-0.162	0.596	0.575	0.040	0.049	0.873	0.244	0.422	-0.011	0.973	100.0
4	Query-star Score	0.335	0.263	-0.121	0.694	0.497	0.084	-0.084	0.785	**-0.828**	0.000	-0.192	0.529	**-0.718**	0.006	-0.177	0.563	5.0
5	Badness	-0.330	0.271	**-0.670**	0.012	-0.214	0.482	**-0.623**	0.023	-0.099	0.748	**-0.791**	0.001	-0.049	0.874	**-0.717**	0.006	5.0
6	One-sided Badness	**-0.566**	0.044	0.066	0.831	**-0.585**	0.036	-0.464	0.110	0.806	0.001	-0.011	0.972	0.465	0.109	-0.387	0.191	5.0
7	Query-star Score	0.445	0.128	-0.137	0.655	0.508	0.076	-0.017	0.957	-0.823	0.001	-0.253	0.405	-0.610	0.027	-0.166	0.587	10.0
8	Badness	0.049	0.873	-0.462	0.112	0.102	0.740	-0.534	0.060	-0.195	0.523	-0.560	0.046	-0.211	0.490	-0.613	0.026	10.0
9	One-sided Badness	-0.407	0.168	0.027	0.929	-0.427	0.145	-0.437	0.136	0.699	0.008	0.038	0.901	0.353	0.237	-0.332	0.267	10.0
10	Query-star Score	0.379	0.201	-0.027	0.929	0.330	0.270	-0.035	0.909	-0.666	0.013	-0.071	0.817	-0.425	0.147	-0.137	0.655	20.0
11	Badness	-0.269	0.374	-0.522	0.067	-0.314	0.296	-0.600	0.030	-0.138	0.654	-0.566	0.044	0.019	0.951	-0.619	0.024	20.0
12	One-sided Badness	-0.269	0.374	-0.049	0.873	-0.511	0.075	-0.390	0.188	0.627	0.022	0.011	0.972	0.374	0.208	-0.246	0.418	20.0

Table 13: Correlation between T_m and CASPER scores for proteins in ProTherm$_{lt60gt70}$ set. The extrema for the correlation values are in bold. P-value is for null hypothesis that there is no correlation based on the coefficient. For calculating p-values, the correlation coefficient was transformed into a t-statistic and the p-value was calculated by using a t-test.
PCC = Pearson Correlation Coefficient
SCC = Spearman Correlation Coefficient
column titles ending with *pval* are for p-values of the correlation mentioned to the left of such columns
T_m *diff* = difference between thermophilic and mesophilic T_m values