

Analysis of Tertiary Contact Conservation among Structurally Related Proteins

**TEJENDRA
MS16092**

*A dissertation submitted for the partial fulfilment
of BS-MS dual degree in Science*

Under the guidance of
Dr. Shashi Bhushan Pandit, IISER Mohali



**Indian Institute of Science Education and Research Mohali
April 2021**

Certificate of Examination

This is to certify that the dissertation titled “Analysis of tertiary contact conservation among structurally related proteins” submitted by Mr. Tejendra (Reg. No. MS16092) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.



Dr. Shashi Bhushan Pandit
(Supervisor)



Dr. Santosh B Satbhai

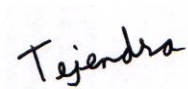


Dr. Kuljeet Singh Sandhu

Dated: April 30, 2021

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Shashi Bhushan Pandit at the Indian Institute of Science Education and Research Mohali. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.



Tejendra
(Candidate)
April 30, 2021

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.



Dr. Shashi Bhushan Pandit
(Supervisor)

Acknowledgements

I cannot be grateful enough to my supervisor Dr. Shashi Bhushan Pandit who supported me in every way possible to help me complete my thesis. He always guided me out of all troubling situations and always motivated me to give my best. I would also like to thank all my lab members for their invaluable support and insights both on and off the field. I am particularly thankful to Paras for his guidance and helpful discussions. I would like to express my gratitude to my parents and close friends, Prerna and Vishal who have been dependable pillars of strength for me all through my time here.

I am thankful to my thesis committee members Dr. Shashi Bhushan Pandit, Dr. Santosh B Satbhai and Dr. Kuljeet Singh Sandhu for their support and guidance. I would also like to thank IISER Mohali library for the invaluable resources and facilities provided by them to the entire IISER community.

Lastly, I would like to acknowledge the support provided by Biology department and IISER Mohali and thank them for providing me an opportunity to pursue my desired project.

— Tejendra

List of Figures

1.1	SCOP hierarchical classification of protein domains	4
2.1	Schematic representation to show contact conservation of structurally equivalent residues.	9
3.1	Histogram showing distribution of contact conservation of structures related at family level in different classes	12
3.2	Histogram showing the distribution of contact conservation of structurally equivalent residues for structures in 4 classes	13
3.3	Box plot showing the distribution of contact conservation of residues of secondary structure type and buried/exposed state for structures related at family level for all four classes	14
3.4	Box plot showing distribution of contact conservation of residues of secondary structure type and buried/exposed state for structures related at superfamily level for all four classes	15
3.5	Box plot showing relation between sequence and contact conservations of family related proteins for all classes	16
3.6	Box plot showing relation between sequence and contact conservations of superfamily related proteins for all classes	16

List of Databases and Software

1. SCOPe database v2.04
2. PALI database v3.0
3. COMPASS database v2.5
4. mTMalign
5. Stride
6. AACon
7. Naccess

Content

Certificate of Examination	i
Declaration	ii
Acknowledgements	iii
List of Figures	iv
List of Databases and Software.....	v
Abstract	viii
Introduction.....	1
1.1 Protein tertiary structure.....	1
1.2 Significance of protein structure	2
1.3 Residue contacts	2
1.4 Structural database SCOP	3
1.5 Structural alignment of proteins	4
1.6 Objective	6
Methodology	7
2.1 Dataset generation	7
2.2 mTMalign structural alignment.....	8
2.3 Contact conservation	8
2.4 Calculation of sequence conservation	10
2.5 Secondary structure and residue accessibility	10
Results and Discussions	11
3.1 Extent of contact conservation in homologous sequences	11

3.2	Contact conservation of buried/exposed and type of secondary structure	13
3.3	Relation between sequence conservation and contact conservation	15
	Conclusions.....	17
	Bibliography	18

Abstract

Protein tertiary structures, despite insignificant sequence similarity, show remarkable structural similarity. However, it is not clearly known whether tertiary contacting residues of structurally equivalent residues are also structurally equivalent. Moreover, whether such residues are conserved in homologous sequences is also unclear. In the present study, we have systematically studied the contact conservation of structurally conserved residues at various levels of structural similarities. Further, we hypothesized that structurally and functionally important residues should have high contact conservation even in distantly related proteins. The analyses showed that there is significant contact conservation, however, coverage of these structurally equivalent residues diminishes from closely to distantly related proteins. We also computed the correlation of contact conservation with various features such as sequence conservation, secondary structure and residue accessibility. This showed that contact conservation is observed mostly in regular secondary structures, as expected, however, coil regions show variable conservation. We found that both buried/exposed residues have comparable contact conservation of structurally equivalent residues. Surprisingly, we observed that structurally equivalent residues do not show residue conservation. This could be mainly because of the co-evolution of interacting residues.

Chapter 1

Introduction

1.1 Protein tertiary structure

The complex biological reactions, which make up processes in living organisms, are dependent upon the presence of proteins for their role as structural molecules. A protein's biological function is defined by the organization of atoms in the three-dimensional space, which is referred to as the tertiary structure of proteins. Traditionally, protein structure can be described based on increasing complexity at four hierarchical levels: primary, secondary, tertiary and quaternary structure. The *primary structure* is the linear representation of amino acid sequence and *secondary structure* is local conformation of protein backbone, secondary structure arrange in spatial orientation to form *tertiary structure*, which refers to a completely folded and energetically stable state of the protein that represents 3-D arrangement of secondary structure (Branden and Tooze, 1999). The structure is stabilized by a number of favorable interactions such as hydrogen bonds, hydrophobic interactions, electrostatic interactions, salt-bridges, and disulfide bonds. The tertiary structure of protein associated with other polypeptide proteins in specific geometry and spatial orientation of these is referred to as *quaternary structure*. Among these, hydrogen bonds are an important force, as they provide

both directionality and specificity of interactions and are known to stabilize secondary structures of proteins.

1.2 Significance of protein structure

Proteins adopt three-dimensional (3-D) structure to perform diverse molecular functions. This was realized with the experimental determination of sperm whale myoglobin structure in the 1960s and followed by structures of many other proteins. This led to the idea of sequence-structure-function paradigm, according to which protein sequence determines its structure and structure determines the protein function. However, recent studies have found aberration to this paradigm wherein proteins, known as intrinsically disordered proteins, without adopting unique three-dimensional structure perform their function. Therefore, knowledge of protein tertiary structure can provide molecular insights into function of proteins such as spatial arrangement of catalytic residues in an active site or how a protein interacts with other proteins for structural or other regulatory activities. Moreover, crucial structural features such as ligand binding sites, protein interaction sites, and flexible regions could be exploited in rational designing of drugs against the lead drug target. Given the significance of tertiary structure determination, a worldwide initiative Structural Genomics Initiative (SGI) of National Institute of Health (NIH) was established to decrease the ever-increasing large gap between the number of protein sequences (Burley and Bonanno, 2002). Moreover, this would allow constructing template libraries of structures, which can be used for modeling of protein structures using sophisticated computational approaches (Yan & Moult, 2005; Levitt, 2009).

With the availability of many protein structures, it was realized that proteins can be structurally similar despite having low or insignificant sequence identity between them (Chothia, 1992). This suggests that evolutionary selection pressure instead of evolving new arrangements of secondary structures tend to re-use pre-existing folds to evolve new functions.

1.3 Residue contacts

Two residues are defined to be in contact, if any heavy atom of a residue is within atomic distance cut-off 4.5 Å to another residue heavy atom. All residues in contact with a residue constitute its residue contact list. The contact residue could be defined as short-range

(local), when it is within ± 6 neighboring residues in sequence; otherwise, it is referred to as long-range (distant) contacting residues. The long-range residue contacts (tertiary contacts) are essential for folding of proteins because this brings sequentially far residues in spatially close proximity. Therefore, both local and distant contacts in protein structures play essential roles in the stability and function of proteins. The knowledge of contacts can aid in tertiary structure of protein such as in Template Based Modelling (Yan & Moult, 2005). Usually, these predicted structures would have correct topology, whereas side-chain orientations need to be optimized. Recent studies have shown that knowledge of contacts is important to elucidate the allosteric mechanism. Moreover, network analysis of contacts can shed light into various aspects of protein structure and function such as mutation/s, difference in stabilities of mesophilic and extremophilic proteins.

1.4 Structural database SCOP

Protein domains are defined as independent folding units that form the basic 'building blocks' of proteins in evolution and architecture (Wetlaufer, 1973). It has also been proposed that protein domains are structurally self-sufficient in terms that if cleaved from protein backbone, the domain would retain their three-dimensional geometry and often their function. With the availability of multiple structures, the most accepted domain concept is based on the globularity or compactness of the proteins, which assumes that the atomic interactions within domains are stronger than between domains. Such domains are also defined as structural domains. Although, analysis of individual protein structure can reveal a great deal of information, a comprehensive view of proteins can be understood by comparison of multiple protein structures and studying evolutionary relationships between them. This necessitated the organization of protein structures into structural domains, which will be classified at various levels based on similarity of structures. Two databases are commonly used for structural domain definition and classification are: CATH and SCOP. In the present study, we have used SCOP domains (Murzin et al., 1995; Fox et al., 2014). The hierarchical classification of SCOP is mainly at four levels as detailed below-

- Class: Types of folds. This is the top-level or root of the SCOP hierarchical classification. These classes have structures with similar secondary structure composition but different tertiary structures.

- **Fold:** The different groups of domains within a class. This classification level indicates protein domains with similar tertiary structure.
- **Superfamily:** The domains within a fold are further grouped into superfamilies, which have at least a distant common ancestor. However, the different members of a superfamily have low sequence identities.
- **Family:** The domains in a superfamily are then grouped into families, which have a more recent common ancestor than superfamilies. Protein families are more closely related.

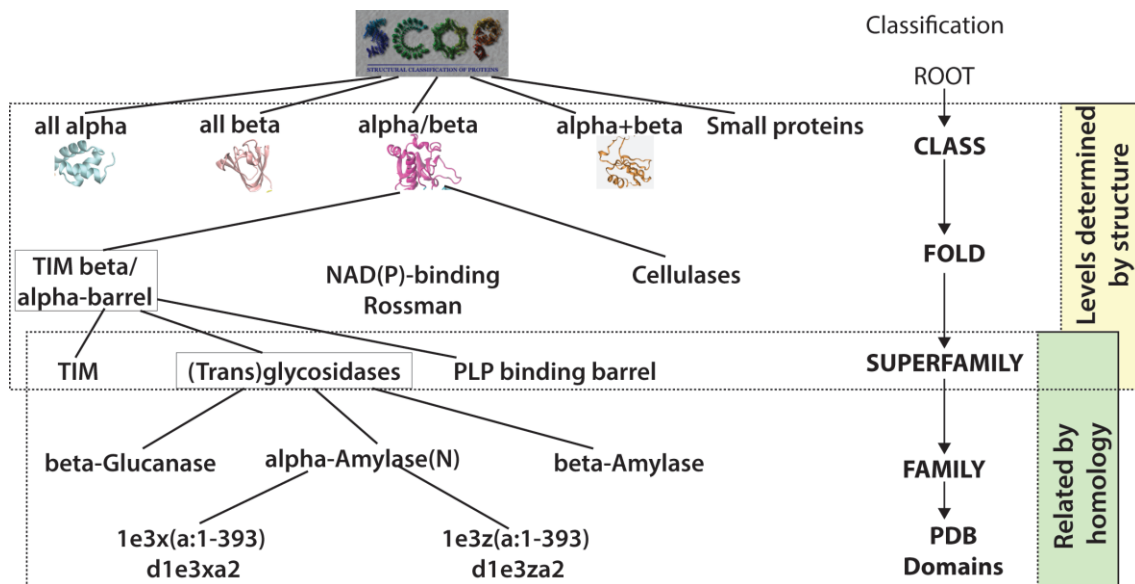


Figure 1.1. SCOP hierarchical classification of protein domains

In 2013, SCOPe (SCOP-extend) was developed, which is an extension to v1.75 (last released version) of SCOP database (Fox et al., 2013). It focuses on the classification of new PDB structures in SCOP-1.75 by utilizing automated methods.

1.5 Structural alignment of proteins

As described before, protein tertiary structures can show remarkable similarity in structure despite having low sequence identity (Chothia, 1992). During evolution, proteins

evolve mostly by mutation and insertion/deletion events leading to an extensive change in sequences, however, conserving the topology of their constituent secondary structures. In fact, analyses of structural families have shown that homologs frequently share fewer than 15% identical residues. Comparison of structures across a protein family gives insights into the tolerance to structural change for a given family and also the impacts any changes have on the functions of the proteins. The degree to which structures diverge as their sequences change during evolution is complex and varies with the structural class and architecture of the protein and also whether there are functional constraints. In some families, structures are highly constrained and sequences can change significantly before any large structural change is observed. Recent analyses have revealed that the degree of structural variability across a protein family, also described as the structural plasticity, varies considerably between protein families. That is, some protein families appear to tolerate much more structural change than others.

The structural alignment is referred to as sequence independent when structurally equivalent residues are not known a priori. In this case, the first stage is to make an estimated guess of structurally equivalent residues followed by a process of heuristics and refinement steps to obtain structural alignment (Pandit and Skolnick, 2008). The most critical step in later stages is maximization of score, which mostly is Root Mean Square Deviation (RMSD). An alternate score (TM-score) for maximization has shown great improvement in structural alignment (Pandit and Skolnick, 2008). Most structural alignment methods perform pairwise alignment because obtaining reliable structural equivalent residues across multiple structures is rather difficult. Recently, a structural alignment method mTMalign has been developed that uses TM-score for maximization and results in core structural equivalent residues (Dong et al, 2018).

RMSD is the most commonly used metric while quantifying the similarities between superimposed atomic coordinates, which is performed using Kabsch algorithm (Kabsch, 1976). Usually, RMSD is computed from Euclidean distance of equivalent C- α atoms of superposed coordinates C- α atoms given by the equation below:

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (x_i^b - x_i^a)^2 + (y_i^b - y_i^a)^2 + (z_i^b - z_i^a)^2}{N}}$$

where, x, y, z are the coordinates of atoms with a and b denoting coordinates from 2 structures; N is the total number of superposed C-atoms

The structural alignments of proteins are essential to detect distantly related sequences (remote homologs), hence, tracing their molecular evolution. Apart from this, alignments can be used for function prediction of a new protein by detecting regions of local and global similarity to a protein with known function. Last, but not least, finding similar protein structures during database scanning to fetch potential templates used for homology modeling (Carugo, 2006).

1.6 Objective

In order to understand the role of contacts, we have analysed the extent of contact conservation among homologous proteins by structural alignments of multiple structures to define structural equivalent positions. The contact of a residue is defined to be conserved, if its contacting residues are structural equivalent to each other. These will be explored for their association with known function of proteins and their role in stability of the superfamily/family. In the present work, we have addressed following questions:

- How much protein atomic contacts are conserved within homologous protein structures, which are closely and distantly related as described by the SCOP level of family and superfamily respectively?
- Do residues having contact conservation are mostly buried/exposed or belong to a certain secondary structure class?
- What is the extent of sequence conservations at structurally conserved residues having conserved contacts?

Chapter 2

Methodology

2.1 Dataset generation

The present work required a dataset of structures belonging to SCOP family/superfamily. We used previously constructed non-redundant datasets of structures instead of generating our own dataset. We have commonly used database Phylogeny and ALIgnment of homologous protein structures (PALI) for structure representative of family (Sujatha et al., 2001). We selected representatives for structures belonging to four major classes: α , β , α/β , $\alpha+\beta$. We removed any structure determined using NMR. Since we were interested in performing alignment of multiple structures, we eliminated families having 2 members. Thus, we have a dataset of SCOP families, which have at least 3 members each. This dataset consisted of following number of families in each class:

Class α	-	147
Class β	-	198
Class α/β	-	311
Class $\alpha+\beta$	-	262

The dataset of non-redundant sequences for superfamily was obtained from CAMbridge database of Protein Alignments organised as Structural Superfamilies (Compass) database (Sowdhamini et al., 1998). Following the same procedure as used above, we obtained structure representatives for superfamilies. The number of superfamilies selected from the database is shown below:

Class α	-	129
Class β	-	141
Class α/β	-	116
Class $\alpha+\beta$	-	125

2.2 mTMalign structural alignment

In our study, we used the mTMalign program (Dong et al., 2018) to create multiple structure alignments (MSAs). The MSAs were generated for structures belonging within family and representatives between family, *i.e.*, at the superfamily level. These MSAs were then used to identify structurally equivalent residues for which we calculated contact conservation and performed subsequent analyses.

2.3 Contact conservation

We have used atomic contacts, which are defined as, two residues are said to be in contact if any heavy atoms of these residues are within a distance of 4.5 Å of each other. In order to avoid local contacts, we considered contacting residues to be at least three residues apart from each other in sequences. Therefore, residues i and j are in contact, if $j > i+2$ and any heavy atom of i and j are within a distance of 4.5 Å of each other.

As we are interested in analyzing contact conservation, we defined the extent of contact conservation as when residues in contact to a residue are also structurally equivalent (see figure 2.1). Therefore, a residue i having a contact conservation score of 1 means that all interacting

residues of i are also structurally equivalent. On the contrary, a contact conservation score of 0 would mean none of contacting residues is structurally equivalent.

Since equivalent residues can have different numbers of contacting residues, we normalize it to a minimum contact residue and this also serves as a reference to compute equivalent residues for all other structures. Such normalization can provide slight conservation on the higher side. The equation to calculate contact conservation is given below:

$$Fc_{i,l} = \frac{\sum_{k=1}^{N_i} \left(\sum_{j=1}^N d_{i,k,j} / N \right)}{\min(Q_{ij}) \forall j = 1, N_j}$$

where, $Fc_{i,l}$ is the fraction of contact conservation of residue i of structure l ;

N_i is the number of contacting residue of residue i ;

N is the total number of aligned structures;

$d_{i,k,j}$ is knocker delta function,

$$d_{i,k,j} = \begin{cases} 1 & \text{if } p_{i,k,l} \text{ has structural equivalent residue in structure } j \\ 0 & \text{otherwise} \end{cases}$$

$p_{i,k,l}$ is the k^{th} residue in contact with residue i in structure l ;

$Q_{i,j}$ is the number of residues in contact of residue i in structure j

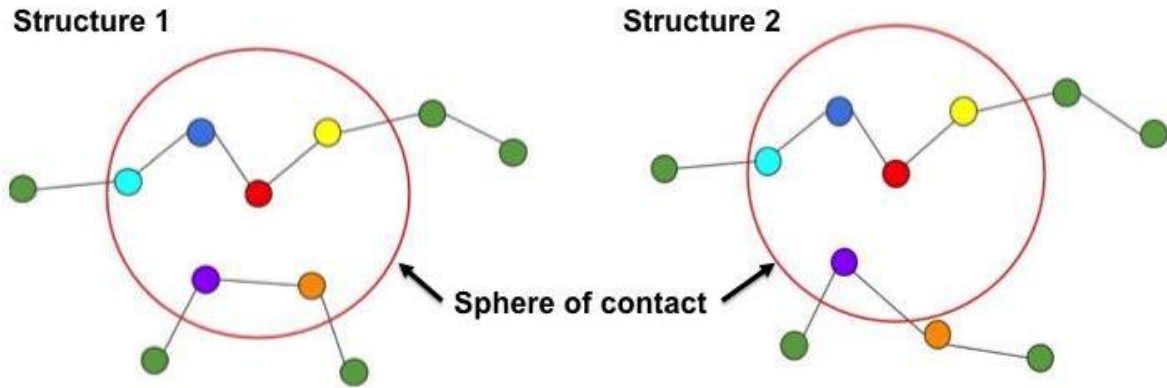


Figure 2.1: Schematic representation to show contact conservation of structurally equivalent residues.

The circle represents a sphere of contact showing all residues in contact to the residue as shown in 'red'. The residues are shown as small colored balls with similar colors showing structurally equivalent residues in two structures (structure 1 and 2).

2.4 Calculation of sequence conservation

The sequence conservation of position of structurally equivalent residues were calculated using Gerstein substitution matrix and method using the AACon program (Agnieszka Golicz et al., 2018). The program can calculate sequence conservation using 18 different scoring schemes or methods. The calculation of V_{Gerstein} , which measures the entropy of a position relative to that if the sequences were aligned randomly, is defined as below:

$$V_{\text{Gerstein}} = \sum_i^K \bar{p}_i \ln \bar{p}_i - \sum_i^K p_i \ln p_i,$$

where \bar{p}_i is the average frequency of amino acid i in the alignment and $K = 20$.

2.5 Secondary structure and residue accessibility

Protein secondary structure assignment was performed using the STRIDE program (Henig and Frishman, 2004). Using the assignment, we performed 3 state classification of secondary structure. All types of helices are classified as helix, all types of strand are classified as β -strand, and rest other residues are assigned as Coil.

The solvent accessibility of residue is calculated using the NACCESS program (Hubbard and Thornton, 1993). We defined a residue as exposed or buried based on its relative solvent accessible surface area (RSA). If a residue has RSA $<20\%$, it is defined to be buried, else it is classified as solvent-exposed residue. Since we are considering multiple structures, an aligned position is defined as buried or exposed if $>60\%$ of aligned residues are buried or exposed.

Chapter 3

Results and Discussions

3.1 Extent of contact conservation in homologous sequences

In the present work, we have investigated the extent of contact conservation in homologous sequences as identified by structures related at the family/superfamily of SCOP. The structures related at the family level are suggested to be closely related, typically, having 30% sequence identity or higher. However, superfamily-related structures have low sequence identities. As described in methods, contact conservation of a structurally equivalent residue is defined as a fraction of all contacting residues, which are also structurally equivalent normalized by the number of residues in contact. A high contact conservation means its contacting residues are also structurally equivalent. Alternatively, it shows how many neighboring residues of a residue of a protein are also structurally similar. Such information can be useful for modeling the tertiary structure of the protein as well as to identify residues, which have a functional role in proteins.

First, we performed multiple structure alignment of structures within a family using mTMalign and calculated the extent of contact conservation of structurally equivalent residues. In general, ~70% residues constitute the equivalent residues in multiple structural alignment.

The Figure 3.1 summarizes the distribution of contact conservation for families belonging to four classes. As can be seen, in general, more than 90% of structurally equivalent residues have 0.8 or higher contact conservation in various classes of proteins. There is a small fraction of residues (<10%), which have almost no contact conservation. At family level, most of the residue's contacts are conserved, as expected, because within family the protein structures are also typically closely related to each other.

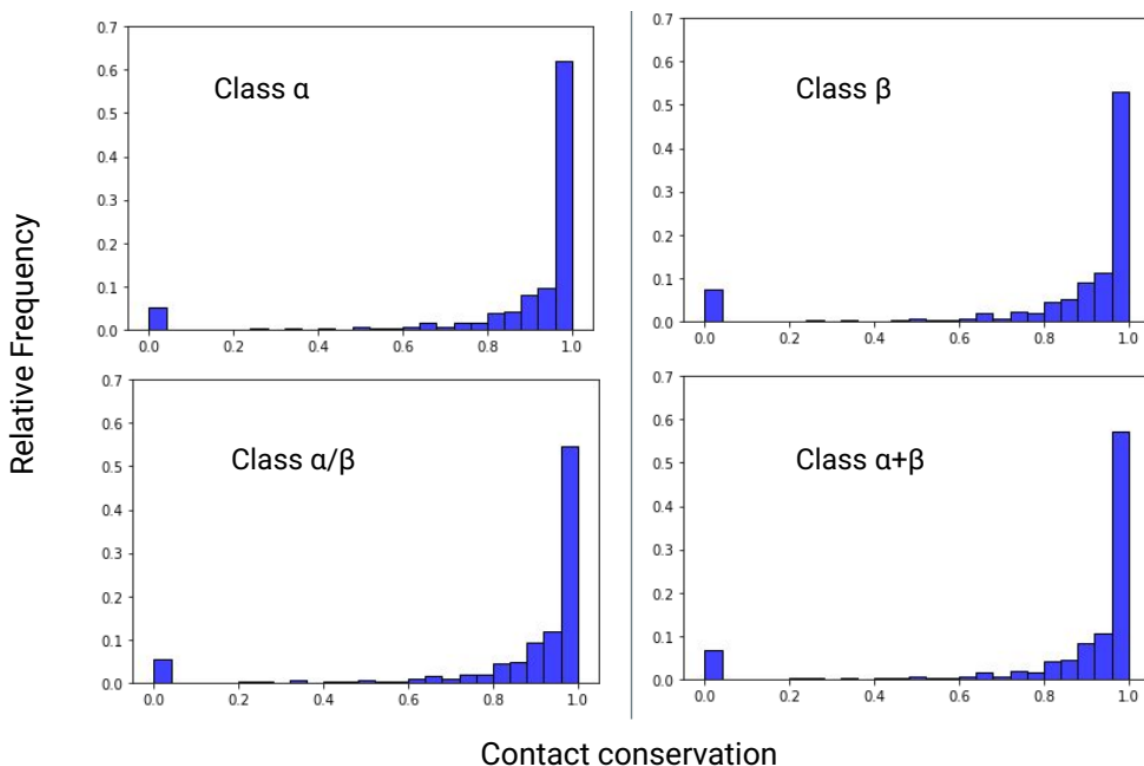


Figure 3.1: Histogram showing distribution of contact conservation of structures related at family level in different classes

Having observed high contact conservation for residues in proteins related at family levels, we examined whether spatially neighboring residues are also conserved in structures related at superfamily level. As described in methods, we selected representative structures from CAMPASS database and aligned them using mTMalign. We found that structurally equivalent residues show average coverage of ~40% in various multiple structural alignments that is small in comparison to family level. Following the approach described before, we calculated the extent of contact conservation, whose distribution is shown in Figure 3.2. On

average, we observe that α -class superfamily shows conservation similar to family level. However, superfamilies in other classes have distribution shifted to a lower side with a decrease in the fraction of superfamilies showing contact conservation close to 1. Despite a slight shift in the distribution, the contact conservation is still high among proteins related at superfamily.

The high contact conservation in protein structures related at family/superfamily suggests that structurally equivalent residues tend to conserve spatially their neighboring residues. It will be interesting to see whether a similar trend is observed in proteins related at the level of fold.

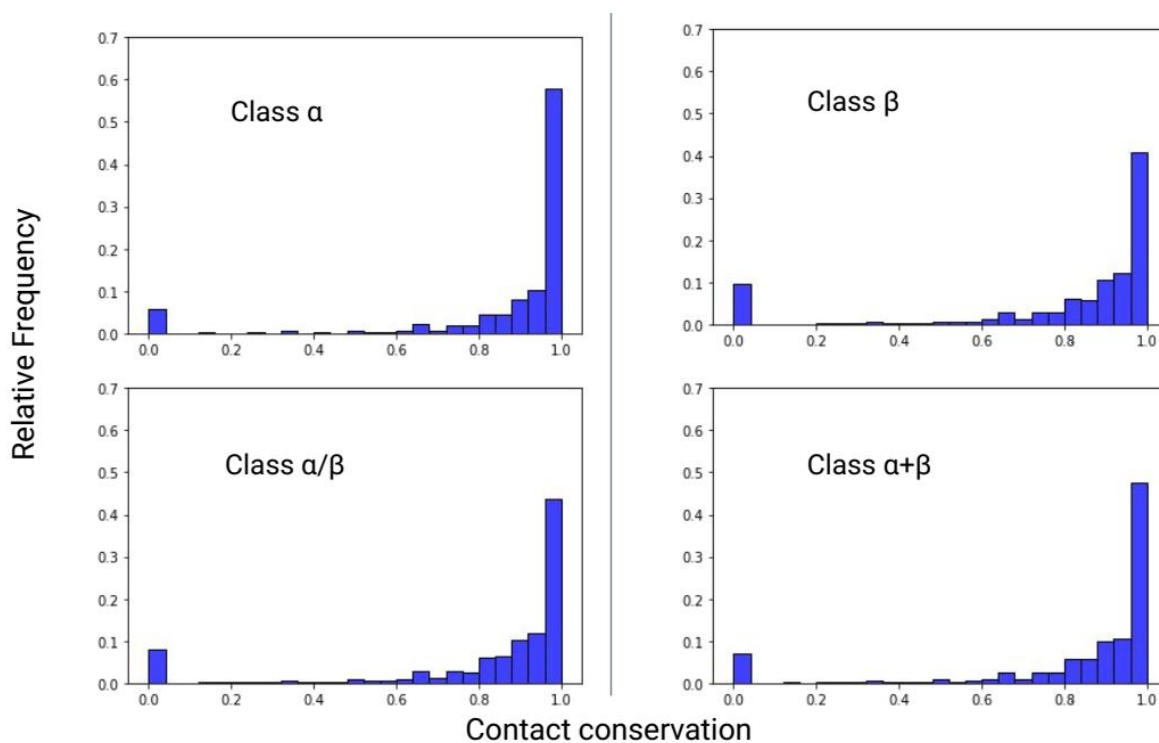


Figure 3.2 Histogram showing the distribution of contact conservation of structurally equivalent residues for structures in 4 classes.

3.2 Contact conservation of buried/exposed and type of secondary structure

Next, we analyzed whether residues having high/low contact conservation belong to a certain secondary structure type or are buried/exposed residues. It is known that buried residues are structurally conserved part of the protein core, so we analyzed whether they also have a

greater extent of contact conservation. Figure 3.3 shows the distribution of contact conservation as a box plot for three secondary structure types and whether a residue is buried or exposed for structures related at the family level. As can be seen, contact conservation for both regular secondary structures is higher (>0.9) for either buried/exposed residues. As expected, coil region shows on average poor contact conservation. Interestingly, we observed that buried residues show a greater contact conservation than exposed residues except for α -helices. It could be because of less number of exposed helical residues.

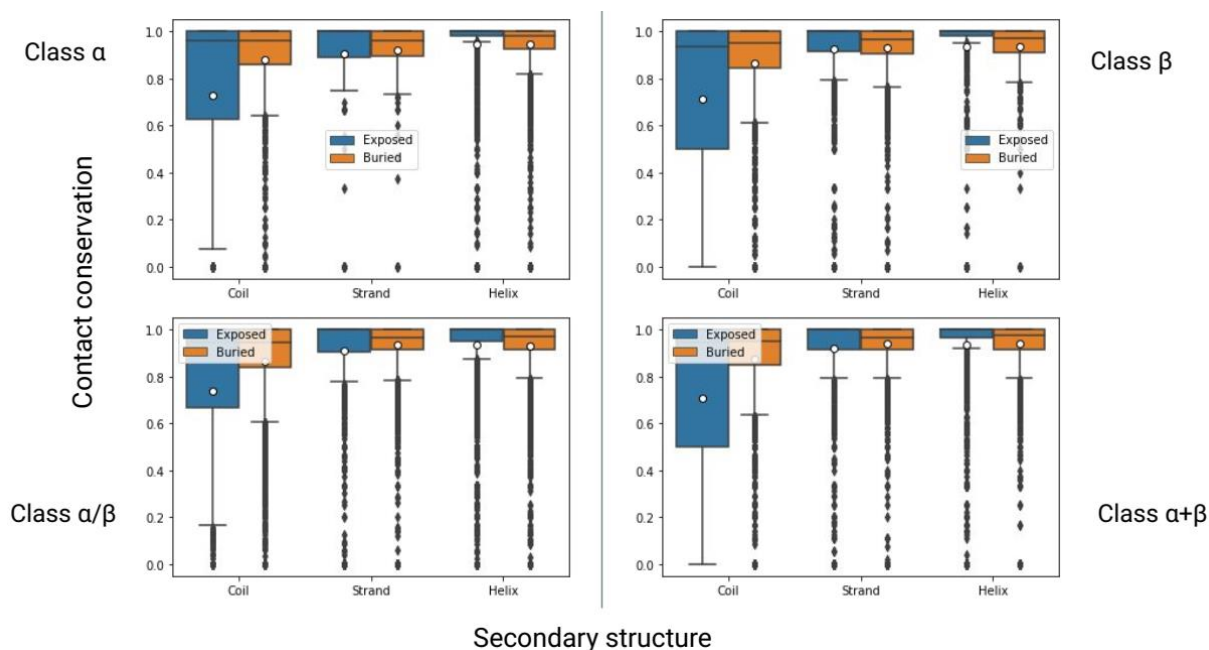


Figure 3.3: Box plot showing the distribution of contact conservation of residues of secondary structure type and buried/exposed state for structures related at family level for all four classes

We performed the same analyses on proteins related at the superfamily level. The results are summarized in Figure 3.4. As has been observed in the family, the contact conservation relation to secondary type and buried/exposed residues are seen similar for proteins related at the superfamily level.

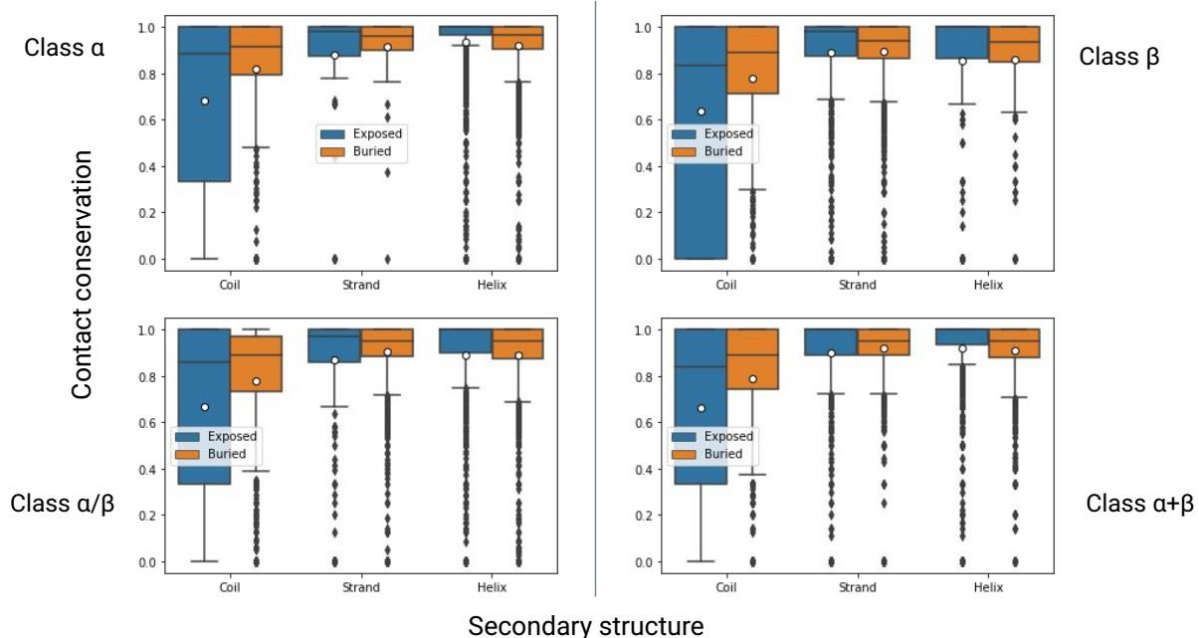


Figure 3.4: Box plot showing distribution of contact conservation of residues of secondary structure type and buried/exposed state for structures related at superfamily level for all four classes

3.3 Relation between sequence conservation and contact conservation

We explored whether extent of contact and sequence conservation are correlated with an assumption that structurally equivalent residues with high contact conservation will probably have a greater sequence conservation as well. The sequence conservation was computed using aaCons and the Gerstein scoring scheme, which is an entropy based score. As can be seen Figure 3.5 for family level and Figure 3.6 for superfamily related proteins, we could not find any correlation of sequence to their contact conservation. Therefore, suggesting that it is possible that structurally equivalent residues are able to conserve their neighboring interacting residue environment by co-evolving residues.

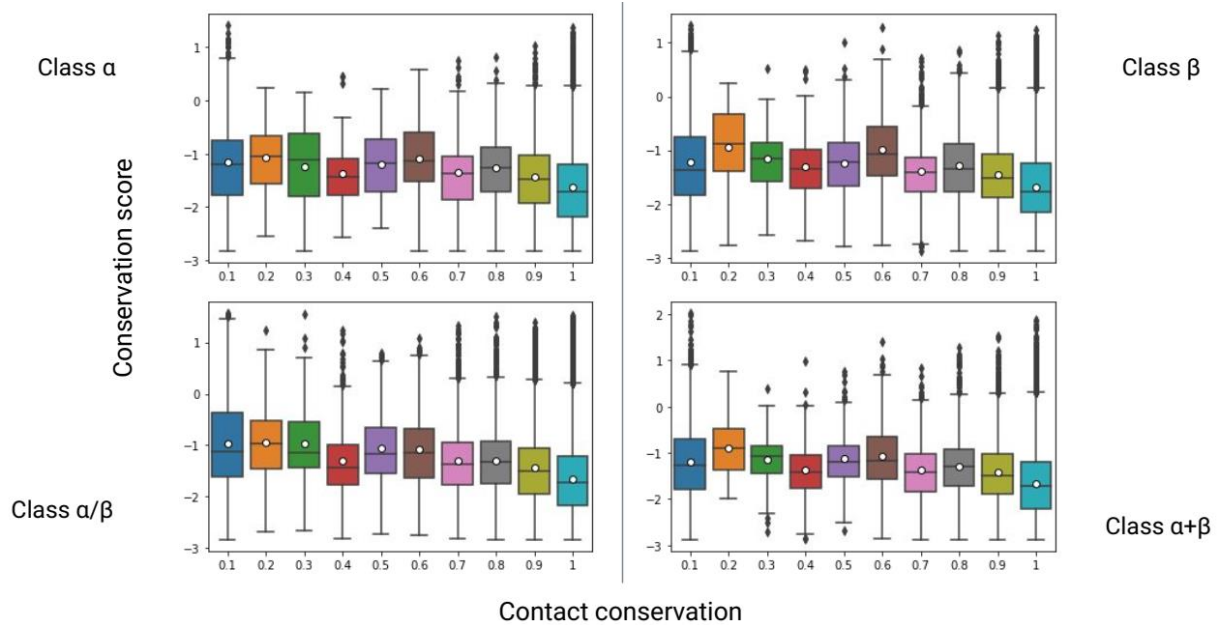


Figure 3.5: Box plot showing relation between sequence and contact conservations of family related proteins for all classes.

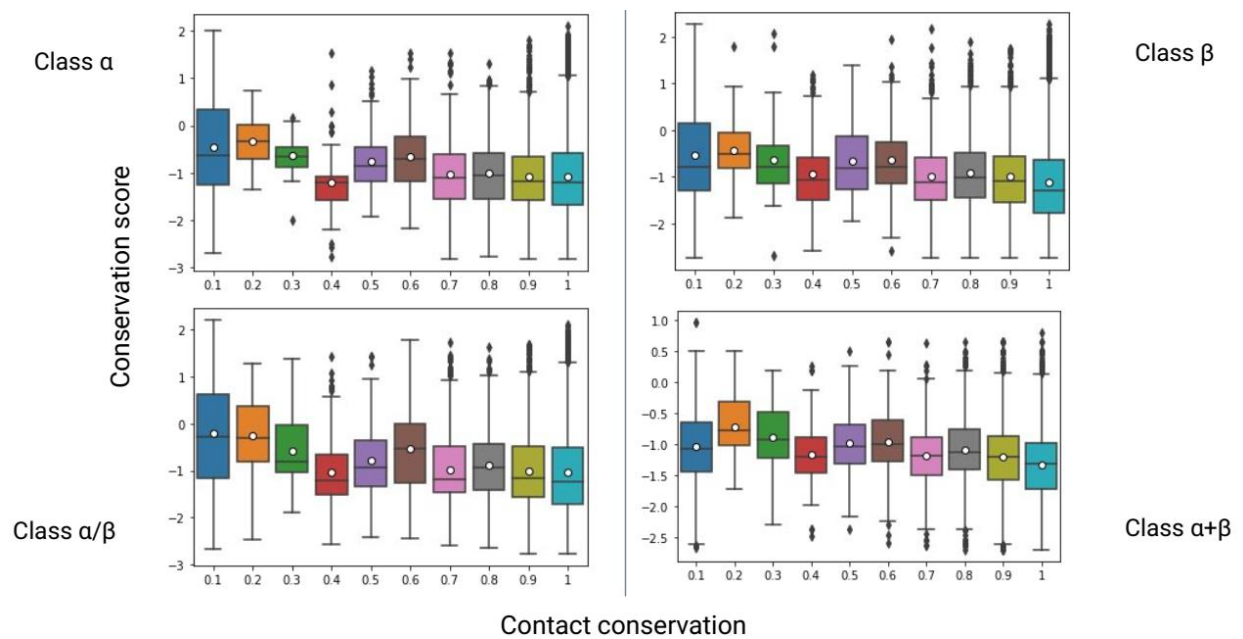


Figure 3.6: Box plot showing relation between sequence and contact conservations of superfamily related proteins for all classes.

Chapter 4

Conclusions

The present study is a preliminary work to understand conservation of spatial neighbors of structurally equivalent residues when proteins are related at family/superfamily level. For this, we analyzed the contact conservation of closely/remotely related proteins. We found that contact conservation of equivalent residues is conserved despite proteins being distantly related to each other. The examination of dependence of contact conservation on secondary structure type or whether a residue is buried/exposed showed that residues in regular secondary structure, as expected, show a greater level of conservation. Interestingly, we could not find correlation between extent of contact and sequence conservation. This raises an interesting question for further investigation for this observation. It can be speculated that structurally equivalent residues are able to conserve their neighboring interacting residue environment by co-evolving residues.

Bibliography

Agnieszka Golicz, Peter V. Troshin, Fábio Madeira, David M. A. Martin, James B. Procter and Geoffrey J. (2018) AACon: A Fast Amino Acid Conservation Calculation Service

Branden, C., & Tooze, J. (1999). *Introduction to Protein Structure* (2nd ed.). Garland Publishing.

Burley, S. K., & Bonanno, J. B. (2002). Structuring the universe of proteins. *Annual Review of Genomics and Human Genetics*, 3, 243-262.

Carugo, Oliviero. (2006). *Rapid Methods for Comparing Protein Structures and Scanning Structure Databases* (Vol. 1). Bentham Science Publishers.

Chothia, Cyrus. (1992). One thousand families for the molecular biologist. *Nature*, 357(6379), 543–544.

Dong, R., Pan, S., Peng, Z., Zhang, Y., & Yang, J. (2018). mTM-align: a server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky430>

Fox NK, Brenner SE, Chandonia JM. 2014. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* 42:D304-309. doi: 10.1093/nar/gkt1240.

Heinig, M., Frishman, D. (2004). STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins. *Nucl. Acids Res.* , 32, W500-2

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5), 922–923.

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences*.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4), 536–540.

Pandit, S.B., Skolnick, J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* 9, 531 (2008).

Sowdhamini et al. (1998). CAMPASS: a database of structurally aligned protein superfamilies. *Structure*, 6(9), 1087–1094. [https://doi.org/10.1016/s0969-2126\(98\)00110-5](https://doi.org/10.1016/s0969-2126(98)00110-5)

S. Hubbard and J. Thornton. 1993. NACCESS, Computer Program. Department of Biochemistry Molecular Biology, University College London.

S. Sujatha, S. Balaji, N. Srinivasan, PALI: a database of alignments and phylogeny of homologous protein structures , *Bioinformatics*, Volume 17, Issue 4, April 2001, Pages 375–376, <https://doi.org/10.1093/bioinformatics/17.4.375>

Yan, Y., & Moult, J. (2005). Protein Family Clustering for Structural Genomics. *Journal of Molecular Biology*, 16.

Wetlaufer, D. B. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences*, 70(3), 697–701.