ANALYSIS OF PSEUDO-SYMMETRY IN PROTEIN HOMO-OLIGOMERS

by

Catherine Jenifer Rajam Rajendran

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer Science

Indianapolis, Indiana

December 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Shiaofen Fang, Chair
Department of Computer & Information Science
Dr. Jing-Yuan Liu
Department of Computer & Information Science
Dr. Yao Liang
Department of Computer & Information Science

Approved by:

Dr. Shiaofen Fang
Head of the Graduate Program

TABLE OF CONTENTS

LIST OF TABLES 5
LIST OF FIGURES 6
ABSTRACT 7
CHAPTER 1. INTRODUCTION 8
1.1 Protein Structures 8
1.2 Finding Symmetry in Proteins 9
CHAPTER 2. SYMMETRY IN QUATERNARY STRUCTURES 11
CHAPTER 3. TOOLS USED TO FIND SYMMETRY 15
3.1 Importance of this project 15
3.2 Steps involved to find Symmetry 15
3.3 Tools Used 15
3.3.1 Chimera 16
3.3.2 Pdbcur 16
3.3.3 Defective Residue Fix 17
3.3.4 PDB File Format 17
3.3.5 Languages and Programs worked on for this project 18
CHAPTER 4. AXIS OF SYMMETRY 19
4.1 Off Symmetry 20
4.1.1 Rotation 21
4.2 Structural Index (SI) 24
4.3 Assembly Index (AI) 26
CHAPTER 5. B-FACTOR 29
CHAPTER 6. RESULTS 30
6.1 Off Symmetry Comparison between CA and CB atoms 30
6.2 Contribution of SI and AI to Off-Symmetry in CA and CB atoms 31
6.3 Correlation between B-Factor values and Off Symmetry among CA and CB atoms 32
CHAPTER 7. PROTEINS USED 34
7.1 Dimers 34
7.2 Trimers 35

7.3 Tetramers 36
CHAPTER 8. CONCLUSION & SUMMARY 38
REFERENCES 40

LIST OF TABLES

Table 6.1: Correlation of OS and BFactor for CA and CB Atoms 33
Table 6.2: Correlation of OS, SI and AI with BFactor 33
Table 7.1: Homo-Dimer Proteins 34
Table 7.2: Homo-Trimer Proteins 35
Table 7.3: Homo-Tetramer Proteins 36

LIST OF FIGURES

Figure 1.1: 180 rotation of water molecule 8
Figure 2.1: Two-fold and three-fold axis representation 11
Figure 2.2: Different types of Symmetries in Tetramers 12
Figure 2.3: Figure 2.3: Dihedral Symmetry and its axis illustration 13
Figure 2.4: Evolution of homo-tetramers 13
Figure 3.1: UCSF Chimera Tool used to rotate and align structures 16
Figure 3.2: PDB File format 17
Figure 4.1: Axis of Symmetry visualized in Homo-Dimers 20
Figure 4.2: Tetramer Rotations 21
Figure 4.3: Trimer Rotations 23
Figure 4.4: Dimer Rotations 24
Figure 4.5: Tetramer Alignments for SI 25
Figure 4.6: Tetramer Alignments for AI 27
Figure 6.1: Dimers Contribution of CB Atoms 30
Figure 6.2: Trimers Contribution of CB Atoms to OS 30
Figure 6.3: Tetramers Contribution of CB Atoms to OS 31
Figure 6.4: Dimers Contribution of SI and AI for OS 31
Figure 6.5: Trimers Contribution of SI and AI for OS 32
Figure 6.6: Tetramers Contribution of SI and AI for OS 32
Figure 7.1: 1nc7 protein PDB file sample 37

ABSTRACT

Author: Rajendran, Catherine Jenifer Rajam. MS
Institution: Purdue University
Degree Received: December 2018
Title: Analysis of Pseudo-Symmetry in Protein Homo-Oligomers
Committee Chair: Shiaofen Fang

Symmetry plays a significant role in protein structural assembly and function. This is especially true for large homo-oligomeric protein complexes due to stability and finite control of function. But, symmetry in proteins are not perfect due to unknown reasons and leads to pseudosymmetry. This study focuses on symmetry analysis of homo-oligomers, specifically homo-dimers, homo-trimers and homo-tetramers.

We defined Off Symmetry (OS) to measure the overall symmetry of the protein and Structural Index (SI) to quantify the structural difference and Assembly Index (AI) to quantify the assembly difference between the subunits. In most of the symmetrical homo-trimer and homo-tetramer proteins, Assembly Index contributes more to Off Symmetry and in the case of homo-dimer, Structural index contributes more than the Assembly Index. The main chain atom Carbon-Alpha (CA) is more symmetrical than the first side chain atom Carbon-Beta (CB), suggesting protein mobility may contribute to the pseudosymmetry. In addition, Pearson coefficient correlation between their Off-Symmetry and their respective atoms B-Factor (temperature factor) are calculated. We found that the individual residues of a protein in all the subunits are correlated to their average B-Factor of these residues. The correlation with BFactor is stronger in Structure Index than Assembly Index. All these results suggest that protein dynamics play an important role and therefore a larger off-symmetry may indicate a more mobile and flexible protein complex.

CHAPTER 1. INTRODUCTION

A symmetry operation is characterized as an operation that is performed on an object or molecule that is evidently unaltered in its structure. An object exhibits symmetry if it still looks the same after a transformation such as rotation. For instance (figure 1.1), if a water molecule is rotated by 180 around the axis that is perpendicular to the normal plane, which is passing through the center, i.e. origin, the resulting structure is indistinct from the original structure.

Figure 1.1: 180 rotation of water molecule

It is presumed that over half of homo-oligomeric proteins are homo-dimers or homo-tetramers and they are assumed to be symmetrical proteins (ref). Proteins are polymer chains formed by sequences of amino acids linked by peptide bonds. An amino acid in a protein chain can also be called a residue. Proteins can be composed by one such kind of chain and is called monomer. It can also be composed by multiple chains and in this case, they are called oligomer. If the chains are identical, then they are called as homo-oligomeric proteins and if they are not identical, they are called as hetero-oligomeric proteins.

1.1 Protein Structures

The quaternary level of protein structure includes the ways protein atoms meet up to make extensive assemblies or complexes. The interatomic cooperation that offer ascent to the folded structure of proteins additionally enable individual protein atoms to tie to each other to create

oligomers of different types and sizes. These bigger assemblies of protein atoms regularly have functional properties that the individual molecules don't. The arrangement of symmetric oligomers from identical structures is a proficient approach to create macromolecular assemblies with new control components and functions. Quaternary structure of protein complexes requires thought of how the subunits are symmetrically identified with created bigger assemblies or oligomers. This likewise includes examining the spatial connections between the identical protein protomers. Most people know about general thoughts concerning symmetry, particularly when it includes perfect symmetries. How ones right and left hands are connected is one of the most punctual exposures to ideas about symmetry. Those ideas should be expanded to some degree to see how proteins interface and pack into quaternary structures.

1.2 Finding Symmetry in Proteins

Homo-oligomers which are assembled in a cyclic fashion from multiple identical protein protomers symmetrically around a central axis of symmetry plays a significant role in many biological processes (Fallas, Ueda et al. 2017). Many methods and techniques were developed to investigate and to determine symmetry in proteins such as Nuclear magnetic resonance (NMR) spectroscopy, DISCO, SymD and CE-Symm Algorithms. NMR or (Magnetic Resonance Spectroscopy) MRS is a spectroscopic technique to observe the local magnetic fields around atomic nuclei, which is used as a tool to perform studies on protein (NMR spectroscopy Wikipedia). Restraints on distance between the pairs of nuclei in the protein is derived from Nuclear Overhauser Effect (NOE). NMR technique is also used to find symmetry in proteins, but they faced a lot of challenges to use it in protein homo-oligomers as it was difficult to determine which

subunits has the restrained protons. (Martin, Yan et al. 2011) proposed a new technique called DISCO to find distance restraints in oligomer structures.

SymD is a structural alignment of proteins residue wise by performing different transformations to evaluate the distance between the carbon-alpha atoms of the original structure and the aligned structure of the protein residues within a chain (Kim, Basner et al. 2010). CE-Symm is another algorithm used to find symmetry in proteins. They use Combinatorial Extension (CE) to perform alignment in protein structure to determine symmetry. Distance between the residues of the protein structure is measured from the aligned structures. Root Mean Square Deviation is calculated for the transitional and rotational alignment of the structure (Myers-Turnbull, Bliven et al. 2014). The method used in SymD algorithm is implemented to perform rotational alignment of the protein structure. These techniques and algorithm are used to determine the symmetry in protein structure, but they do not help us to quantify symmetry in proteins. Though it is a common notion that protein structures are symmetric, it does not contain perfect symmetry. (Bonjack-Shterengartz and Avnir 2015) introduces a special probability analysis on the protein structure as Continuous Symmetry Measure (CSM) to quantify symmetry in protein complexes and defines that protein structures are not perfectly symmetrical. In this project we quantify the pseudo-symmetry in protein structures by decomposing Off-Symmetry in protein complexes as Structural Index and Assembly Index between the subunits.

CHAPTER 2. SYMMETRY IN QUATERNARY STRUCTURES

Protein structures can be classified using the crystallographic point groups. All symmetry elements of a molecule pass through a central point within the molecule. The more symmetry operations a molecule has, the higher its symmetry is. If a molecule has rotational symmetry and rotating it by 360/n, it brings the object into an equivalent position, where the value of n is the order of an n-fold rotation. If the molecule has one or more rotational axes, the one with the highest value of n is the principal axis of rotation. The notation for n-fold symmetry is Cn. The actual symmetry group is specified by the point or axis of symmetry as discussed, together with the n. For each point or axis of symmetry, the abstract group type is cyclic group of order n. The n subunits are arranged in a cyclic arrangement in a head to tail form; a n-fold rotational axis is set perpendicular to the plane of the circular arrangement of atoms.

The Cyclic Groups, Cn is classified as C1 symmetry (monomeric protein) which has no symmetry. C2 Symmetry, which is homo-dimer and has a solitary two-fold axis. C3 symmetry, a homo-trimer must have a three-fold axis (120 revolution). Figure 2.1 gives us a clear illustration of two-fold and three-fold axis respectively for dimers and trimers.

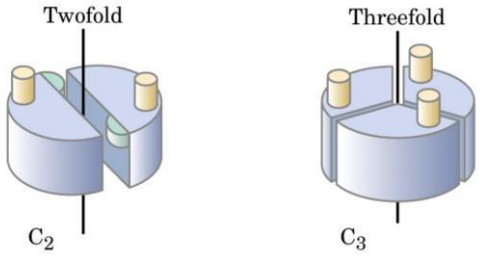

Figure 2.1: Two-fold and three-fold axis representation

(source:http://cbc.arizona.edu/classes/bioc462/462a/NOTES/Protein_Structure/quatern_structure.html)

A homo-tetramer can have two types of symmetries: either a four-fold axes in the cyclic point group C4 symmetry, or three orthogonal two-fold axes in the dihedral point group D2 symmetry (also noted as 222). Both point groups have n = 4, yet they yield altogether different quaternary structures (Figure 2.2).

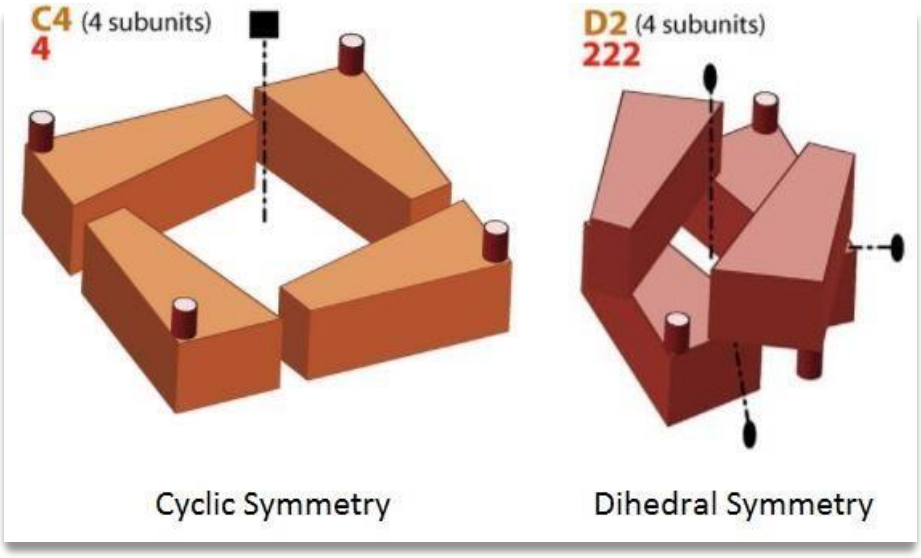

Figure 2.2: Different types of Symmetries in Tetramers

(source:http://cbc.arizona.edu/classes/bioc462/462a/NOTES/Protein_Structure/quatern_structure.html)

Dihedral Symmetry are represented by Dn, where D denotes that it is dihedral, and n is the number of times the element coincides with its original position in one complete turn around the center. It is a combination of a n-fold rotational axis and a two-fold axis perpendicular to it. A ”Dn” structure in this way comprises of two Cn structures stacked top-to-top or base to-base. Start to finish connections of Cn structures do happen yet those ordinarily lead to long linear polymers. The point group 222 (D2) has three-fold axes that are all perpendicular to each other. The minimum number of identical subunits needed to form a D2 symmetry is four subunits. For example, (figure 2.3) the protein streptavidin is made up of four subunits related in pairs by a two-fold axes. There

is a two-fold axis running on a horizontal plane that relates to the red and yellow subunits and the green and cyan ones. The two-fold axis arranged vertically relates the red and green subunits and the yellow and cyan. At last, the third two-fold axis guiding perpendicular toward the view in figure 2.3 relates the red and cyan and the green and yellow subunits.

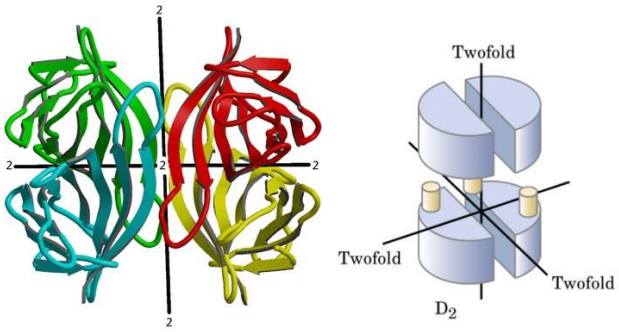

Figure 2.3: Figure 2.3: Dihedral Symmetry and its axis illustration

(source: http://www.els.net/WileyCDA/ElsArticle/refId-a0003121.html)

As mentioned earlier, a homo-tetramer can have either of the two symmetries. A tetramer is a protein with a quaternary structure of four subunits. Homo-tetramers have four identical subunits. For homo-tetrameric proteins, the structure is believed to have evolved going from a monomeric to a dimeric and finally a tetrameric structure in evolution.

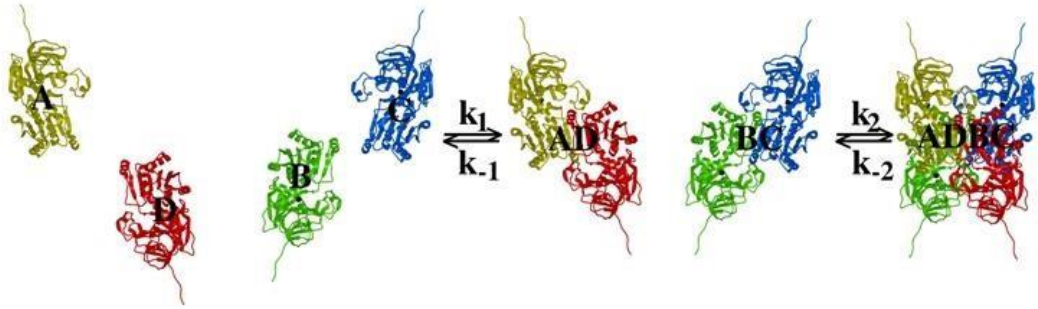

Figure 2.4: Evolution of homo-tetramers

(source: http://www.wikiwand.com/en/Tetrameric_protein)

D2 is more frequently seen than C4 symmetry in homo-tetramers (Sudha and Srinivasan 2016) which indicates the evolution path as illustrated in Figure 2.4 Dihedral symmetry, which requires the quantity of subunits to be even, is extremely normal in globular soluble proteins. For example, Hexamers generally have D3 symmetry and Octamers have D4 symmetry.

In this study, we will investigate imperfect symmetry of dimeric proteins with C2 symmetry, trimeric proteins with C3 symmetry and tetrameric proteins with C4 symmetry. All this Off-Symmetry data for all these symmetry molecules will provide us with more information about the formation of these proteins and why there is an off symmetry in homo-oligomers.

CHAPTER 3. TOOLS USED TO FIND SYMMETRY

3.1 Importance of this project

Symmetrical oligomeric complexes are more favored in large proteins because of stability and finite control of assembly. But this symmetry is not perfect. The main purpose of this study is to learn why there is just an approximate symmetry and what contributes to this disrupted symmetry by quantifying pseudo-symmetry in homo-oligomers.

3.2 Steps involved to find Symmetry

There are several strategies involved to find symmetry in proteins. Below list gives us a brief information involved.

- Download Symmetrical Homo-Oligomers
- Delete duplicate atoms in residues and ANISOU (anisotropic temperature factors) records
- Atoms count is matched as per residues with all 4 chains
- Axis of Symmetry
- Rotation Symmetry
- Protein Alignment

3.3 Tools Used

Few tools have been used to calculate, convert structures and get more assembled data to find symmetry in proteins.

3.3.1 Chimera

UCSF Chimera (figure 3.1) is a highly extensible program for interactive visualization and analysis of molecular structures and related data for further study of structures. High-quality images and animations can be generated. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics and it is funded by the National Institutes of Health. Structure optimization is done using Chimera.

Figure 3.1: UCSF Chimera Tool used to rotate and align structures

3.3.2 Pdbcur

PDBCUR is a curation tool providing various analyses and manipulations of PDB files.

- To delete anisou records from PDB file noanisou
- To delete duplicate CA and CB atoms in a residue

3.3.3 Defective Residue Fix

To work on these PDB files, firstly, we must make sure that the residues in all chains are equal and same atoms count is determined in all chains. This is done by a shell script, which finds the Minimum and Maximum Residue in a PDB file. Residue number is checked in all chains in a PDB file and if any residue is it not available in any of the chain, it is deleted from the PDB file. Similarly, CA, CB atoms count are verified.

3.3.4 PDB File Format

PDB (Protein Data Bank) file format is a textual file format describing the 3D structures of molecules held in Protein Data Bank. PDB file format provides a standard representation for macromolecular structure data derived from X-ray diffraction.

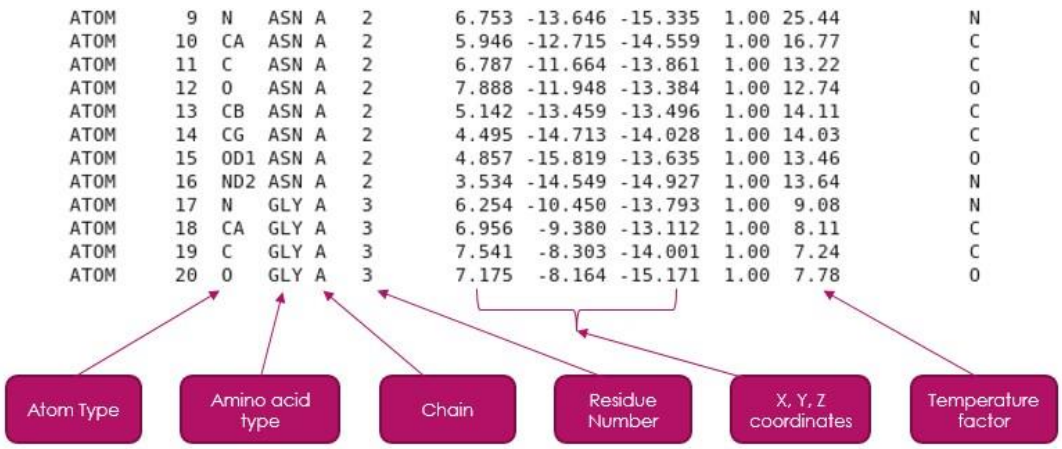

Figure 3.2: PDB File format

In a PDB file, as mentioned in the above figure 3.2, the Atom type is used to capture the carbon alpha atoms and carbon beta atoms. Chain represents the name of the chain. Residue number to compare between the chains. X, Y, Z are the 3D coordinates of each atom. Temperature factor is used to find the B-Factor and the correlation among the Off-Symmetry.

3.3.5 Languages and Programs worked on for this project

1) Shell Script

2) Python

3) JAVA

4) PDBCUR

R Language.

CHAPTER 4. AXIS OF SYMMETRY

Perfect rotational symmetry of order n, also called n-fold rotational symmetry, with respect to a particular axis (in 3D) means that rotation by an angle of 360/n does not change the object. The actual symmetry group is specified by the axis of symmetry, together with the n. Axis of symmetry plays an important role in finding the pseudo symmetry in proteins. As mentioned earlier, Axis of symmetry can be generated by finding the perpendicular axis to plane. In Chimera, there is a built-in feature which was used to find the rotational axis. First, it generates a plane to the protein and then Normal axis to plane (which is perpendicular axis) is determined. But since we could not get the accurate axis using this feature for our research, we generated the axis of symmetry by using Singular Value Decomposition (SVD).

In SVD, we decompose an input source data matrix of size (m x n) and represent it as a product of three matrices. A = UDVT. where, U is (m x m) and V (n x n) and they are defined to be orthogonal. D is a (m x n) diagonal matrix, and the values along this diagonal are the singular values which contains the square of eigen values of the symmetry matrix formed by U and V. This factorization of matrices means that we can multiply the three matrices to get back the original matrix. This articulates the idea that each component matrix can be thought of as applying a different part of the total transformation, which is a rotation and a stretch. Breaking this up, VT does a rotation, D does a stretch or scaling, and U does another rotation.

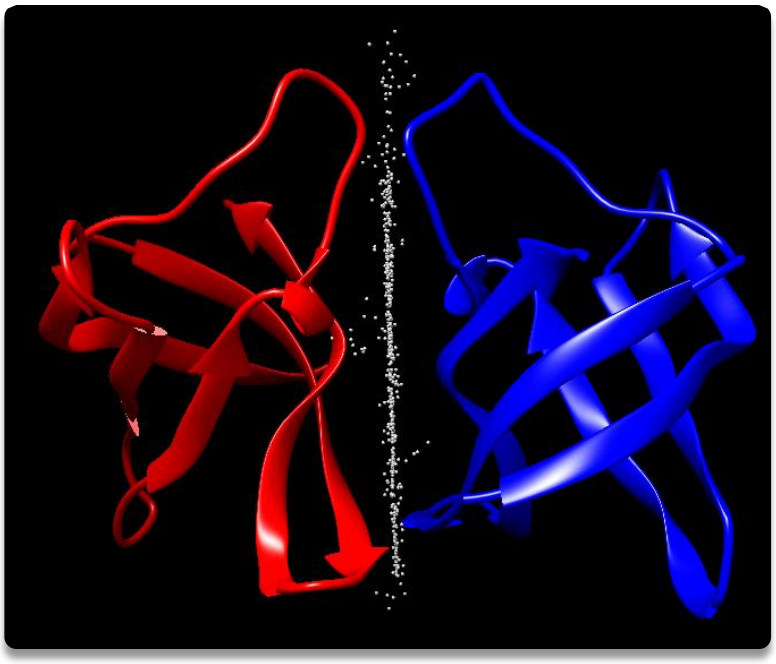

Figure 4.1: Axis of Symmetry visualized in Homo-Dimers

The centroid of the two atom pairs of the protein chains are given as the input source matrix to decompose it using SVD. Eigenvectors or values are calculated from the atomic coordinates. The largest eigenvector is taken as the Axis of Symmetry

4.1 Off Symmetry

A tetramer with C4 symmetry has a single 4-fold rotational axis. This means that rotations of 360/4 = 90 degrees will move each protomer to the equivalent position of another protomer. Only by marking one subunit, we can see the results of such rotations. Euclidean distance between the equivalent atoms in the original tetramer and the rotated structures gives us the OS.

The OS value quantifies how much a homo-tetramer is off the perfect symmetry. OS value is different from structural difference and assembly difference. We calculated OS separately for

the side chain and the main chains CA and CB atoms to find out how much the chains and the atoms contribute to the Off Symmetry.

4.1.1 Rotation

When we tried to rotate the protein structure using our Axis of Symmetry. Our Python code, which uses the chimera, rotated the structure by the vector or anti-vector at times. To fix it, we found the largest magnitude of the vector, if it is positive then the direction of the vector is changed, so all the protein structures are rotated in same direction.

Homo-Tetramers

Tetramers structure is little complicated. Direction of tetramer rotation is not standard, for few, it rotates clock wise and few are rotated anti-clockwise. Firstly, the protein chains are renamed to a common structure order as ABCD and then these proteins are separated into two groups as per the rotation. Using the axis of symmetry, proteins are rotated and respective A, B, C, D are configured (figure 4.2).

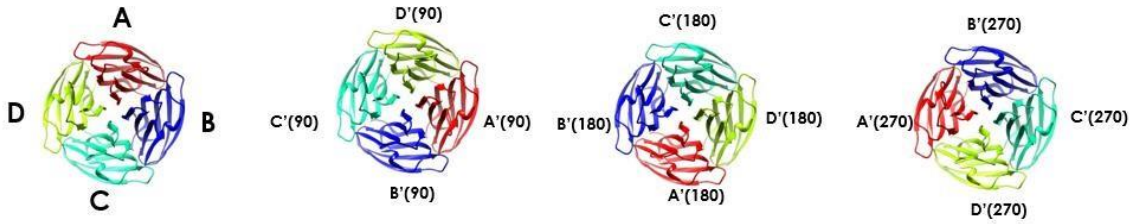

Figure 4.2: Tetramer Rotations

Rotation of homo-tetrameric protein is done using chimera tool, using the command

turn x, y, z 90

where x, y and z represent the 3D coordinates of the axis of symmetry of that specific protein.

The equations used to calculate Off-Symmetry is mentioned below. The distance between the original structure and the rotated structure is calculated for each residue and the value is summed up for each rotation, which are 90, 180 and 270, where N is the number of comparing atom pairs.

$$OS(90) = \frac{\sum_{n=1}^{N}(|A_n'(90)B_n| + |B_n'(90)C_n| + |C_n'(90)D_n| + |D_n'(90)A_n|)}{4N}$$

$$OS(180) = \frac{\sum_{n=1}^{N}(|A_n'(180)C_n| + |B_n'(180)D_n| + |C_n'(180)A_n| + |D_n'(180)B_n|)}{4N}$$

$$OS(270) = \frac{\sum_{n=1}^{N}(|A_n'(270)D_n| + |B_n'(270)A_n| + |C_n'(270)B_n| + |D_n'(270)C_n|)}{4N}$$

Where, A, B, C and D are the rotated structure values and A, B, C, D are the original structure. And then the mean value of all these rotated structures are calculated to get the overall Off Symmetry of the protein structure.

$$OS = \frac{OS(90) + OS(180) + OS(270)}{3}$$

Homo-Trimers

A trimer is a protein with a structure of three identical subunits.

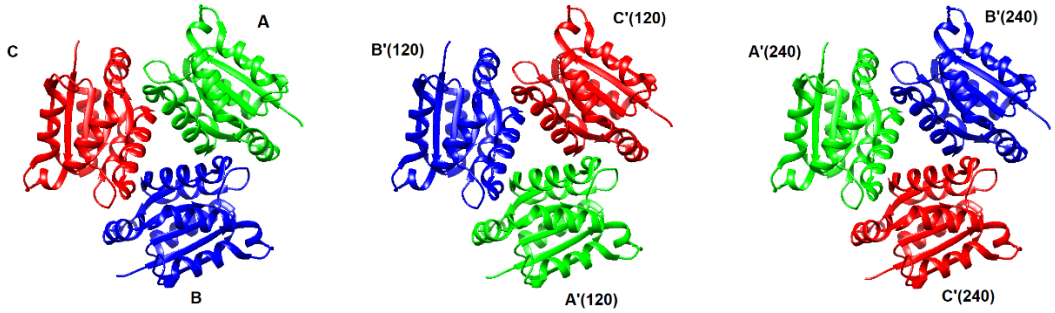

Figure 4.3: Trimer Rotations

The equations used to calculate Off-Symmetry in Homo-Trimer is mentioned below. The distance between the original structure and the rotated structure is calculated for each residue and the value is summed up for each rotation, which are 120 and 240, where N is the number of comparing atom pairs.

$$OS(120) = \frac{\sum_{n=1}^{N}(|A_nC'_n(120)| + |C_nB'_n(120)| + |B_nA'_n(120)|)}{3N}$$

$$OS(240) = \frac{\sum_{n=1}^{N}(|A_nB'_n(240)| + |C_nA'_n(240)| + |B_nC'_n(240)|)}{3N}$$

$$OS = \frac{OS(120) + OS(240)}{2}$$

Where, A, B and C are the rotated structure values and A, B, C are the original structure. And then the mean value of all these rotated structures are calculated to get the overall Off Symmetry of the protein structure.

Homo-Dimers

A dimer is a protein with a structure of two identical subunits.

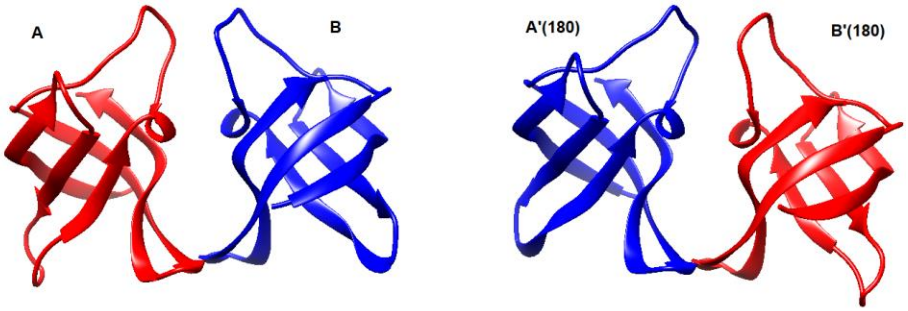

Figure 4.4: Dimer Rotations

The equations used to calculate Off-Symmetry in Homo-Dimer is mentioned below. The distance between the original structure and the 180 rotated structure is calculated for each residue, where N is the number of comparing atom pairs.

$$OS = \frac{\sum_{n=1}^{N}(|A'_nB_n| + |B'_nA_n|)}{2N}$$

Where, A and B are the rotated structure values and A, B are the original structure. And then the mean value of all these rotated structures are calculated to get the overall Off Symmetry of the protein structure.

4.2 Structural Index (SI)

Structural Index in Homo-Oligomers are the structural difference between the original protein structure and the Structurally aligned chains of the proteins. Euclidean distance between the equivalent atom pairs in the aligned structure and the original structure gives the SI. It aligns one chain over the other and calculates the distance.

Homo-Tetramers

Below shows the alignment for Tetramers.

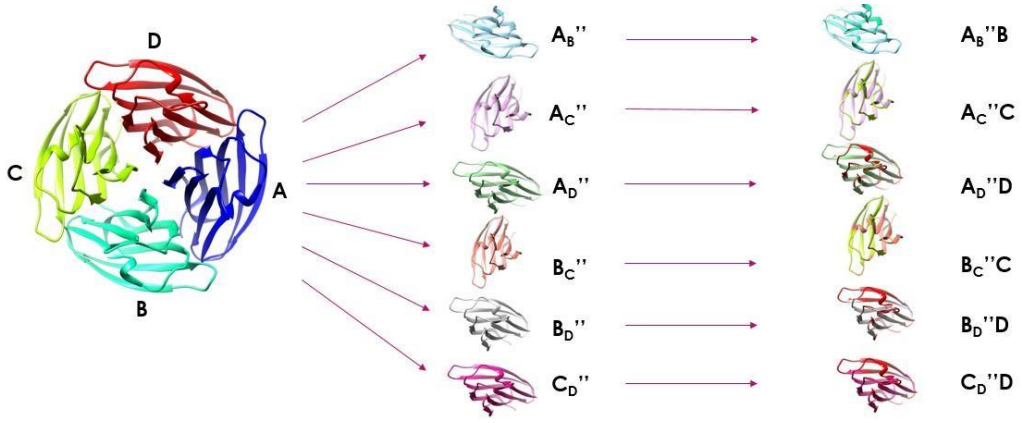

Figure 4.5: Tetramer Alignments for SI

Structure Index for Tetramers can be calculated by measuring the distance between the original structure and the aligned structure for the above mentioned 6 alignments and finding the Euclidean distance between the equivalent atom pairs in the aligned structure and the original structure (figure 4.4). In the below formula, AB, AC, AD, BC, BD and CD are the aligned structure values and A, B, C, D are the original structure. And then the mean value of all these aligned structures are calculated to get the overall structural index

$$SI = \frac{\sum_{n=1}^{N}(|(A_B)_n''B_n|+|(A_C)_n''C_n|+|(A_D)_n''D_n|+|(B_C)_n''C_n|+|(B_D)_n''D_n|+|(C_D)_n''D_n|)}{6N}$$

Homo-Trimers

Structure Index for Trimers can be calculated by measuring the distance between the original structure and the aligned structure for 3 alignments and finding the Euclidean distance between the equivalent atom pairs in the aligned structure and the original structure. In the below formula, AB,

AC and BC are the aligned structure values and A, B, C are the original structure. And then the mean value of all these aligned structures are calculated to get the overall structural index.

$$SI = \frac{\sum_{n=1}^{N}\left(\left|(A_B)_n^{\text{“}}B_n\right| + \left|(A_C)_n^{\text{“}}C_n\right| + \left|(B_C)_n^{\text{“}}C_n\right|\right)}{3N}$$

Homo-Dimers

Dimers Structure Index can be calculated by measuring the distance between the original structure and the aligned structure. Euclidean distance between the equivalent atom pairs in the aligned structure and the original structure. Aligned structure, AB values are measured from original structure B. And then the mean value of all the aligned structures are calculated to get the overall structural index.

$$SI = \frac{\sum_{n=1}^{N}\left(\left|(A_B)_n{}^{\text{“}}B_n\right|\right)}{N}$$

4.3 Assembly Index (AI)

Assembly Index (AI) is the structural difference between the Structurally aligned chains of the proteins and the rotated structure. Euclidean distance between the equivalent atom pairs in the aligned structure and the rotated structure gives the Assembly Index (AI). Assembly Index can be calculated by finding the real distance of differences between the rotated and the aligned structure.

Homo-Tetramers

For Tetramers, the below equation, A, B, C, D are rotated structures and AB, BC, CD and DA are aligned structures for 90 alignment (figure 4.5).

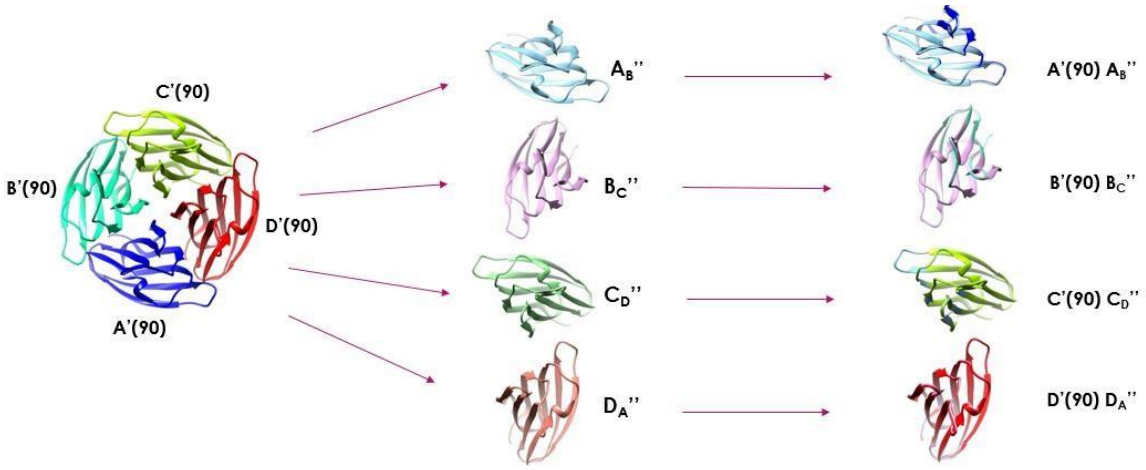

Figure 4.6: Tetramer Alignments for AI

Structure Alignment and Rotated structures are mapped to find distance using RMSD and same procedure is carried on for 90, 180 and 270 separately.

$$AI(90) = \frac{\sum_{n=1}^{N}(|(A_B)_n''A_n'(90)| + |(B_C)_n''B_n'(90)| + |(C_D)_n''C_n'(90)| + |(D_A)_n''D_n'(90)|)}{4N}$$

$$AI(180) = \frac{\sum_{n=1}^{N}(|(A_C)_n''A_n'(180)| + |(B_D)_n''B_n'(180)| + |(C_A)_n''C_n'(180)| + |(D_B)_n''D_n'(180)|)}{4N}$$

$$AI(270) = \frac{\sum_{n=1}^{N}(|(A_D)_n''A_n'(270)| + |(B_A)_n''B_n'(270)| + |(C_B)_n''C_n'(270)| + |(D_C)_n''D_n'(270)|)}{4N}$$

$$AI = \frac{AI(90) + AI(180) + AI(270)}{3}$$

Once the AI (90), AI (120) and AI (270) is found, the mean value of all the three values gives us the final AI value.

Homo-Trimers

Assembly Index for Trimers is the structural difference between the Structurally aligned chains of the proteins and the rotated structure. Euclidean distance between the equivalent atom pairs in the aligned structure and the rotated structure gives the Assembly Index (AI). Assembly Index can be calculated by finding the real distance of differences between the rotated and the aligned structure. For Trimers, the below equation, A, B, C are rotated structures and AB, BC and CA are aligned structures for 120 alignment

$$\mathrm{A}I(120) = \frac{\sum_{n=1}^{N}\left(\left|(A_B)_n^{\text{“}}A_{n\prime}(120)\right|+\left|(B_C)_n^{\text{“}}B_{n\prime}(120)\right|+\left|(C_A)_n^{\text{”}}C'_n(120)\right|\right)}{3N}$$

$$\mathrm{A}I(240) = \frac{\sum_{n=1}^{N}\left(\left|(A_C)_n^{\text{“}}A_{n\prime}(240)\right|+\left|(B_A)_n^{\text{“}}B_{n\prime}(240)\right|+\left|(C_B)_n^{\text{”}}C'_n(240)\right|\right)}{3N}$$

$$\mathrm{AI} = \frac{AI(120)+AI(240)}{2}$$

Homo-Dimers

Dimers Assembly Index is calculated by finding the structural difference between the Structurally aligned chains of the proteins and the rotated structure. For Dimers, the below equation, A is rotated structure and AB is aligned structures for 180 alignment structure.

$$AI(180) = \frac{\sum_{n=1}^{N}\left(\left|(A_B)_n^{\text{“}}A_{n\prime}(180)\right|\right)}{N}$$

CHAPTER 5. B-FACTOR

The B-factor describes the displacement of the atomic positions from an average (mean) value. The more flexible an atom is the larger the displacement from the mean position. The average BFactor, OS, SI and AI of every residue in the protein is calculated. Moreover, The B-Factor for CA and CB atoms for each residue is compared with CA and CB atoms of OS. The B-Factor is normalized with Min-Max Normalization

$$B_{norm} = \frac{B - B_{\min}}{B_{\max} - B_{\min}}$$

To find out whether OS of each residue has correlation with their mobility or flexibility, we investigate B-Factors of these residues. However, it is determined that B-Factor can absorb errors and can be influenced by the quality of X-ray diffraction data.

CHAPTER 6. RESULTS

6.1 Off Symmetry Comparison between CA and CB atoms

CB atoms contributes to the Off Symmetry more than CA Atoms in Dimers, Trimers and Tetramers.

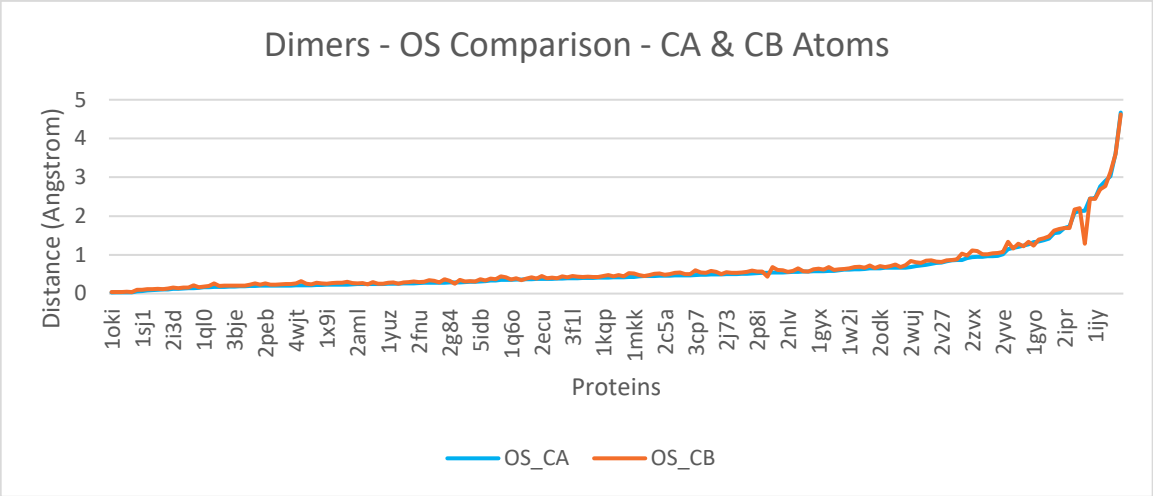

Figure 6.1: Dimers Contribution of CB Atoms

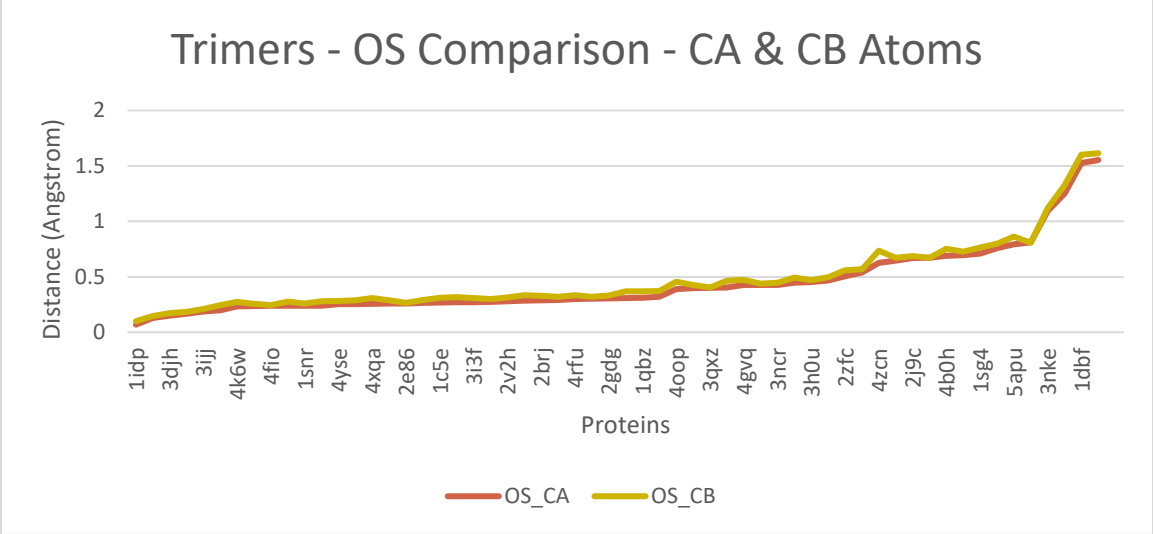

Figure 6.2: Trimers Contribution of CB Atoms to OS

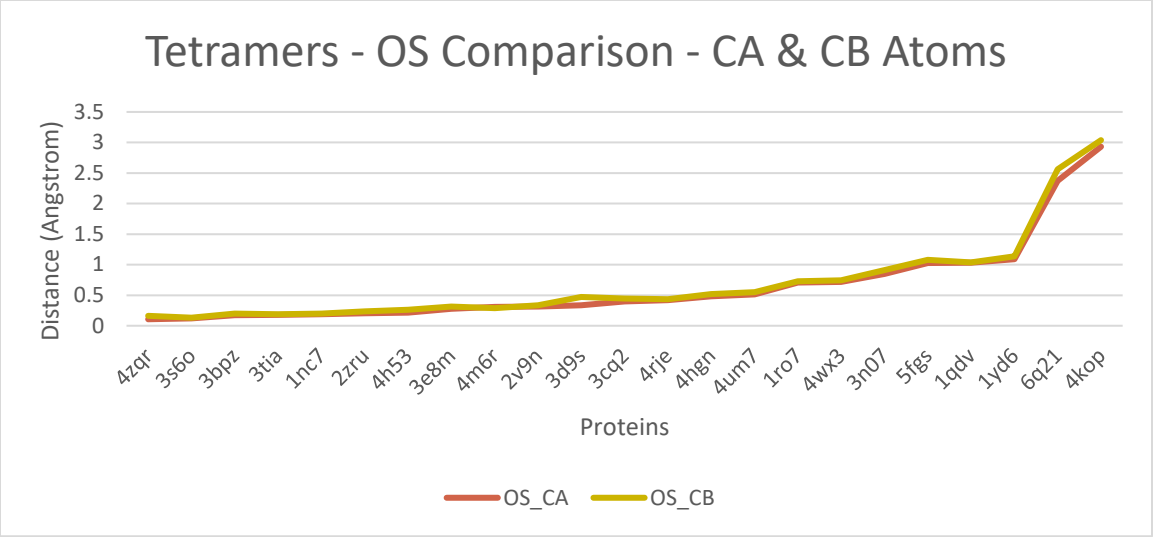

Figure 6.3: Tetramers Contribution of CB Atoms to OS

6.2 Contribution of SI and AI to Off-Symmetry in CA and CB atoms

In most cases, SI contributes to OS in Dimers, but when the OS increases, AI contributes to OS than SI. In Trimers, AI and SI contribution is distributed. In case of Tetramers, AI contributes to Off-Symmetry than SI.

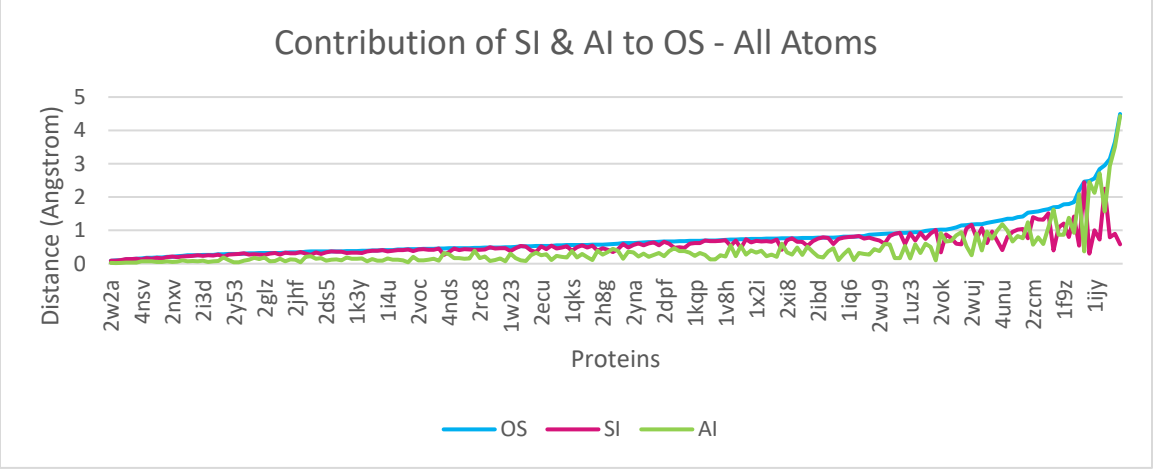

Figure 6.4: Dimers Contribution of SI and AI for OS

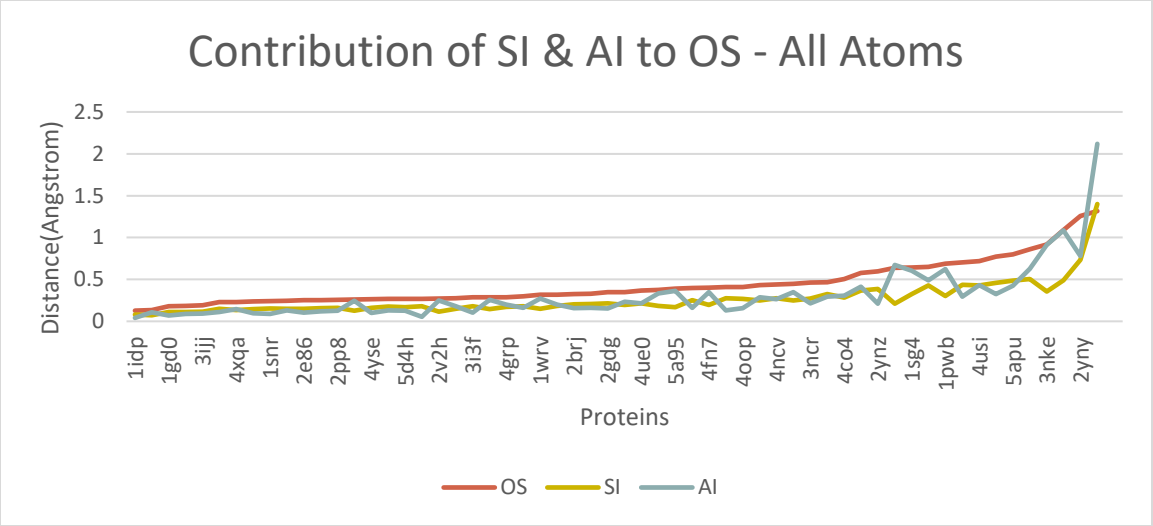

Figure 6.5: Trimers Contribution of SI and AI for OS

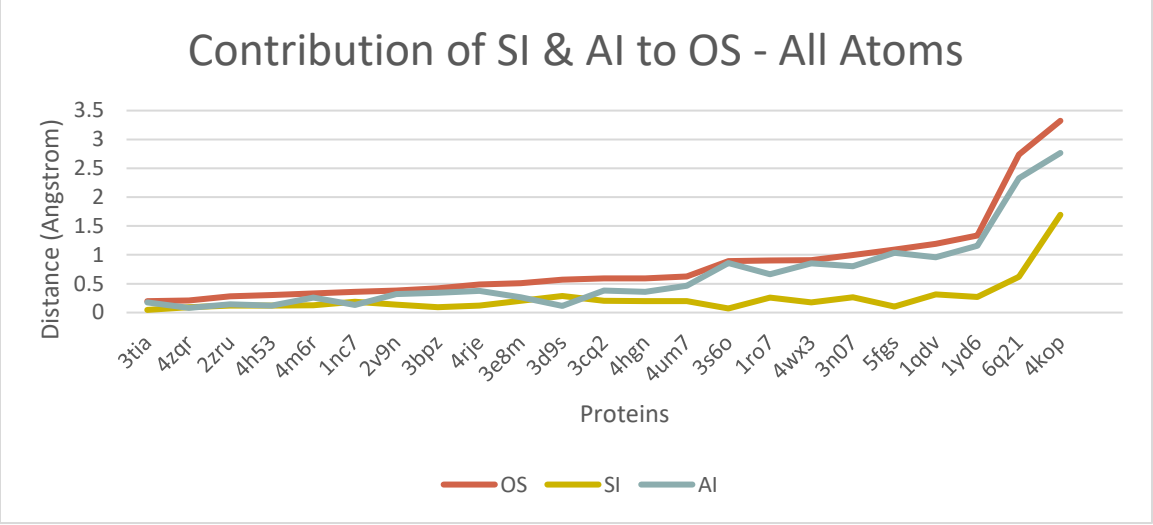

Figure 6.6: Tetramers Contribution of SI and AI for OS

6.3 Correlation between B-Factor values and Off Symmetry among CA and CB atoms

The Correlation between the Off Symmetry and the B-Factor for the CA and CB atoms are captured, and we could see that CA atoms has higher correlation to BFactor than CB Atoms in Dimers and Trimers, which contributes to the displacement.

Table 6.1: Correlation of OS and BFactor for CA and CB Atoms

	Dimers			Trimers			Tetramers		
	CA	CB	All	CA	CB	All	CA	CB	All
Weak 0 – 0.39	34.84%	37.37%	28.78%	24.13%	18.96%	15.52%	30.43%	34.78%	26.09%
	69	74	57	14	11	9	7	8	6
Moderate 0.4 - 0.59	44.45%	48.48%	55.05%	46.55%	55.17%	50.00%	52.17%	39.13%	39.13%
	88	96	109	27	32	29	12	9	9
Strong 0.6 – 1	17.17%	12.12%	15.65%	29.31%	25.86%	34.48%	17.39%	26.09%	34.78%
	34	24	31	17	15	20	4	6	8
< 0	3.53%	2.02%	0.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	7	4	1	0	0	0	0	0	0

The Correlation between the B-Factor, and Structure Index and Assembly Index shows that SI is positively Strongly Correlated than AI.

Table 6.2: Correlation of OS, SI and AI with B-Factor.

	Dimers			Trimers			Tetramers		
	OS	SI	AI	OS	SI	AI	OS	SI	AI
Weak 0 – 0.39	23.23%	22.22%	54.04%	10.34%	8.62%	24.14%	17.39%	26.09%	21.74%
	46	44	107	6	5	14	4	6	5
Moderate 0.4 - 0.59	45.96%	45.96%	29.29%	22.41%	31.03%	32.76%	17.39%	30.43%	34.78%
	91	91	58	13	18	19	4	7	8
Strong 0.6 – 1	30.30%	31.82%	9.56%	67.24%	60.34%	36.21%	65.22%	43.48%	43.48%
	60	63	19	39	35	21	15	10	10
< 0	0.50%	0%	7.07%	0.00%	0%	6.90%	0.00%	0%	0.00%
	1	0	14	0	0	4	0	0	0

CHAPTER 7. PROTEINS USED

7.1 Dimers

198 Homo-Dimer Proteins

Table 7.1: Homo-Dimer Proteins

1c9o	1mkk	1uww	2dsk	2ob3	2w6a	3ct6
1dj0	1mxr	1uz3	2dxu	2odk	2wtp	3ctp
1djt	1nki	1v8h	2dy0	2ofc	2wu9	3cwr
1e7l	1nww	1v9y	2e5f	2p8i	2wuj	3f1l
1e9g	1nxm	1vh5	2e6f	2pa7	2x02	3g46
1eaj	1nzi	1vl7	2ecu	2peb	2xhf	3vrc
1ezg	1o1h	1vzi	2egv	2phn	2xi8	4axo
1f9z	1ofz	1w23	2ehp	2pl7	2xmj	4nds
1g6u	1oi6	1w2i	2f22	2prv	2y53	4nsv
1gve	1oki	1wkq	2fnu	2prx	2yna	4qiu
1gyo	1psr	1wpn	2ftr	2q20	2yve	4rt5
1gyx	1pvm	1x2i	2g84	2q9o	2z6r	4unu
1h41	1pyz	1x9i	2glz	2qe8	2zcm	4wjt
1h4w	1q6o	1xrk	2gom	2qif	2zdp	4yag
1i0r	1qks	1xy1	2gty	2qjw	2zew	4ypo
1i4u	1ql0	1y5h	2gu9	2ql8	2zvx	4ysl
1ijy	1qlw	1yuz	2gyq	2r5o	3a6r	5i5m
1iq6	1rku	1zrs	2h8g	2r8q	3aia	5idb
1isu	1s0p	1zuy	2hin	2rc8	3ayj	
1ix9	1sby	2aib	2i3d	2rl8	3b0f	
1iyb	1sh8	2aml	2i51	2v27	3b4u	

Continued to next page

Table 7.1, Continued

1jr8	1sj1	2arc	2i8t	2vha	3bje	
1k20	1sqs	2axw	2ibd	2voc	3bmz	
1k3y	1sr7	2c5a	2ipr	2vok	3bxu	
1kdg	1t6f	2car	2it2	2vv6	3c3y	
1kqp	1u07	2d8d	2j73	2w1v	3c8e	
1l6r	1u0k	2dkj	2jae	2w2a	3c9u	
1lq9	1ucr	2dpf	2jhf	2w31	3ccd	
1m2d	1usc	2dpl	2nlv	2w3g	3cov	
1m4i	1uwk	2ds5	2nxv	2w3p	3cp7	

7.2 Trimers

58 Homo-Trimer Proteins

Table 7.2: Homo-Trimer Proteins

1c5e	2wpz	4grn
1dbf	2wq4	4grp
1gd0	2yny	4gvq
1idp	2ynz	4k6v
1jlj	2zfc	4k6w
1pwb	3djh	4ncv
1qbz	3fuc	4oop
1sg4	3h0u	4rfu
1sjm	3i3f	4ue0
1snr	3ijj	4usi
1uxa	3mhy	4xqa
1vmf	3ncq	4yse
1wrv	3ncr	4zcn

Continued to next page

Table 7.2, Continued

2bcm	3nke	5a95
2brj	3qxz	5apu
2e86	3rwn	5b4o
2gdg	4b0h	5d4h
2j9c	4co4	5jbx
2pp8	4fio	
2v2h	4fn7	

7.3 Tetramers

23 Homo-Tetramer Proteins

Table 7.3: Homo-Tetramer Proteins

1nc7	3tia
1qdv	4h53
1ro7	4hgn
1yd6	4kop
2v9n	4m6r
2zru	4rje
3bpz	4um7
3cq2	4wx3
3d9s	4zqr
3e8m	5fgs
3n07	6q21
3s6o	

Figure 7.1 contains a sample PDB file of 1nc7 protein. It contains the protein coordinates in residual level for each chain and much more

```
ATOM      9  N   ASN A   2       6.753 -13.646 -15.335  1.00 25.44           N
ATOM     10  CA  ASN A   2       5.946 -12.715 -14.559  1.00 16.77           C
ATOM     11  C   ASN A   2       6.787 -11.664 -13.861  1.00 13.22           C
ATOM     12  O   ASN A   2       7.888 -11.948 -13.384  1.00 12.74           O
ATOM     13  CB  ASN A   2       5.142 -13.459 -13.496  1.00 14.11           C
ATOM     14  CG  ASN A   2       4.495 -14.713 -14.028  1.00 14.03           C
ATOM     15  OD1 ASN A   2       4.857 -15.819 -13.635  1.00 13.46           O
ATOM     16  ND2 ASN A   2       3.534 -14.549 -14.927  1.00 13.64           N
ATOM     17  N   GLY A   3       6.254 -10.450 -13.793  1.00  9.08           N
ATOM     18  CA  GLY A   3       6.956  -9.380 -13.112  1.00  8.11           C
ATOM     19  C   GLY A   3       7.541  -8.303 -14.001  1.00  7.24           C
ATOM     20  O   GLY A   3       7.175  -8.164 -15.171  1.00  7.78           O
```

Figure 7.1: 1nc7 protein PDB file sample

CHAPTER 8. CONCLUSION & SUMMARY

Symmetry is a property of protein homo-oligomers and may play an important role in protein assembly, stability and function. However, proteins are not perfectly symmetrical due to unknown reasons and this leads to pseudosymmetry. We defined Off Symmetry (OS) to measure the overall symmetry of the protein and separately examined two factors, Structure Index (SI) and Assembly Index (AI), that contribute to overall OS. By computing the OS, SI and AI, we found that AI shows major contribution to Off-Symmetry when compared with Structural Index in Tetramers. It is vice versa for Dimers, SI shows major contribution than AI to OS, but as the OS increases, AI starts to take over SI contribution. In Trimers, it is almost equally distributed between SI and AI. When comparing Off-Symmetry, we were able to show that CA atoms has lower Off-Symmetry values than CB atoms in homo-oligomers. As it is known that CB atoms are present in the side chain of the protein, it is more flexible than CA atoms, we hypothesized that protein flexibility may be one of the causes of protein Off-Symmetry.

To understand pseudo-symmetry in protein atoms, we performed Pearson coefficients correlation on OS values along with B-Factor for each residue on CA or CB atoms. We found that in most structures, OS in both CA atoms and CB atoms are moderately or strongly correlated with their B-Factors. Even though B-Factors can generate more mobility in atoms, it also absorbs errors, which is why OS of both CA and CB atoms are strongly correlated to B-Factor. We also captured correlation between OS, SI and AI in Homo-oligomer proteins, we found strong positive correlation in Structure Index than Assembly Index for Off-Symmetry with B-Factor. These results suggest that molecular dynamics is likely involved in developing pseudo-symmetry in proteins.

In the future, we can quantify Off-Symmetry by calculating the distance of residue atom centroid from the axis of symmetry. We can identify a pattern or trends in protein off-symmetry and diversity of protein functions.

REFERENCES

[1] Fallas JA, Ueda G, Sheffler W, Nguyen V, McNamara DE, Sankaran B, Pereira JH, Parmeggiani F, Brunette TJ, Cascio D, Yeates TR, Zwart P, Baker D. *Computational design of self-assembling cyclic protein homo-oligomers*. Nat Chem. 2017 Apr;9(4):353-360. doi: 10.1038/nchem.2673. Epub 2016 Dec 5.

[2] Nuclear Magnetic Resonance Spectroscopy or Magnetic Resonance Spectroscopy, *https://en.wikipedia.org/wiki/Nuclear_magnetic_resonance_spectroscopy*

[3] Jeffrey W. Martin, Anthony K. Yan, Chris Bailey-Kellogg, Pei Zhou and Bruce R. Donald, *A Geometric Arrangement Algorithm for Structure Determination of Symmetric Protein Homo-Oligomers from NOEs and RDCs*. J Comput Biol 2011 Nov; 18(11): 1507–1523. doi:10.1089/cmb.2011.0173.

[4] C. Kim, J. Basner, and B. Lee, *Detecting internally symmetric protein structures*, BMC bioinformatics, vol. 11, no. 1, p. 303, 2010.

[5] D. Myers-Turnbull, S. E. Bliven, P. W. Rose, Z. K. Aziz, P. Youkharibache, P. E. Bourne, and A. Prli´c, *Systematic detection of internal symmetry in proteins using ce-symm*, Journal of molecular biology, vol. 426, no. 11, pp. 2255–2268, 2014.

[6] M. Bonjack-Shterengartz and D. Avnir, *The near-symmetry of proteins*, Proteins: Structure, Function, and Bioinformatics, vol. 83, no. 4, pp. 722–734, 2015.

[7] M Govindarajan Sudha, Narayanaswamy Srinivasan, *Comparative analyses of quaternary arrangements in homo-oligomeric proteins in superfamilies: Functional implications* 14 May 2016 https://doi.org/10.1002/prot.25065