STRUCTURAL FLEXIBILITY OF AN ENZYME UPON BINDING OF LIGAND AT AN ACTIVE SITE: USING ESTIMATE OF ENTROPY OVER ENSEMBLE OF CONTACT MATRICES

by

Qazi Zaahirah

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2017

I start in the name of Allah most beneficent, and merciful. I dedicate this thesis to Allah Almighty for giving me courage, and blessing me with his support. Dear lord, I love you with all my heart. To my parents Qazi Aijaz Rasool, and Rubeena Buchh for supporting me throughout, especially to my dad for constant questions with regards to my thesis, to my grand father Qazi Gulam Rasool and his encouragement, and to my sister Mariyah for being so supportive and helpful.

Table of Contents

List of Tables

List of Figures

Abstract

Contact matrices have been widely used to represent proteins structures. They are transformed to protein contact networks to determine key functional residues, analyze effects of ligand binding, and protein protein interaction. We extend the approach of contact matrices to the aggregation of contact matrices across homologous samples of proteins by defining the frequency contact matrix (FCM). A FCM encodes the frequency of a contact between the side chains of structurally aligned sites. Using this approach, we analyzed the general sitewise contact entropy of a set of protein structures with and without a ligand-bound for residues close to the active site (≤ 10) and farther from the active site. Dataset comprised of enzymes from enolase, mandelate racemase, and muconate cycloisomerase subgroups within the enolase superfamily were constructed. The results show that the median of contact entropy of hydrophobic residues is higher than that for hydrophilic residues. A significant relationship was also observed between the mean contact entropy of residues and their respective properties such as size and hydrophobicity, which indicated that as the hydrophobicity or size of residues increases, the contact entropy of residues also increased. On comparing the contact entropy of residues in the ligand-bound and ligand-free structure, no change was observed for the enolase superfamily. It was also observed that the information obtained from the residue contact entropy values has significant relationship with sequence entropy values. Sequence entropy values indicate the uncertainty of residue type at structurally aligned positions. For some datasets, the residue contact entropy values were found to be sensitive to distance from the active site. The same entropy value distribution for ligand-bound and ligand-free datasets may be due to the dataset collection method and choice of sub-optimal parameters.

List of Abbreviations and Symbols Used

$C^{a'}$	simulated aggregate contact matrix
C^{a}	aggregate contact matrix
$C^{f'}$	simulated frequency contact matrix
C^{f}	Frequency contact matrix
$E^{c}(\mathcal{S}^{\mathcal{L}}, AS)$	contact entropy values at a distance ≤ 10 from active site for ligand-bound structures
$E^{c}(\mathcal{S}^{\mathcal{L}}, \neg AS)$	contact entropy values at a distance > 10 from active site for ligand-bound structures
$E^{c}(\mathcal{S}^{\mathcal{U}}, AS)$	contact entropy values at a distance ≤ 10 from active site for ligand-free structures
$E^{c}(\mathcal{S}^{\mathcal{U}}, \neg AS)$	contact entropy values at a distance > 10 from active site for ligand-free structures
E^{c}	contact entropy
E^{s}	sequence entropy
$\mathcal{D}^{\mathcal{A}}$	distance from active site
$\mathcal{S}^{\mathcal{L}}$	ligand-bound Structures
$\mathcal{S}^{\mathcal{R}}$	reference structure for each dataset
$\mathcal{S}^{\mathcal{U}}$	ligand-free structures
	ngstrom
BLAST	basic local alignment search tool
COMAR	contact map reconstruction
ECDF	empirical distribution function
FASTA	fast alignment search tool

FCM	frequency contact matrices
K-S	Kolmogorov-Smirnov test
KD	Kyte-Doolittle
MATT	multiple alignment with translations and twists
PDB	Protein Data Bank
PDB ID	Protein Data Bank identification code
RCSB	Research Collaboratory for Structural Bioinformatics
RNA	ribonucleic acid
SCOP	Structural Classification of Proteins
SFLD	Structure-Functional Linkage Database
VMD	visual molecular dynamics
XML	extensible markup language

Acknowledgements

1. Christian Blouin

2. Stephen L. Bearne

3. Beiko Lab

4. Blouin Lab

5. NSERC

6. NSHRF Scotia Support Grant

Chapter 1

Introduction

1.1 Structure of Proteins

Proteins are the macromolecules which act as one of the building blocks of life. They are of great importance in almost all biological functions. Proteins are classified by their purpose and the results of their binding; for example structural proteins, enzymes, lectins, proteins of motility, receptors, repressors, immunoglobulins, hormones, and membrane-bound transfer proteins. They are produced in ribosomes of living cells with the help of RNA (ribonucleic acid) which acts as a blueprint for the type of protein that needs to be produced. These RNA blueprints are referred to as transcripts. Proteins are polypeptides i.e., they are long chains of individual units called residues of which there are 20 types with a general structure as shown in Figure 1.1 [14]. Amino acids contain amine ($-NH_2$) and carboxyl (-COOH) groups which form the backbone of the amino acid. The R group of each amino acid, also known as the side chain, may be hydrophilic (water attracting) or hydrophobic (water repelling). The former are generally found on the surface of protein and latter are generally buried inside the protein. The side chains give amino acids its properties, function and localization. For example, the side chain determines the relative hydrophobicity and molecular weight (size) of an amino acid (Table 1.1) [34].

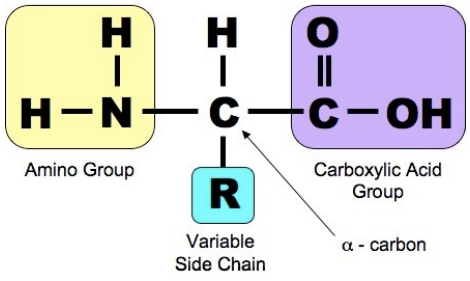

Figure 1.1: Structure of an amino acid. The backbone for all amino acids are the same but the R group varies for different amino acids [43].

Amino Acid	Name	Size	Hydrophobicity
G	glycine	75.0669	-0.4
A	alanine	89.0935	1.8
S	serine	105.093	-0.8
P	proline	115.131	-1.6
V	valine	117.1469	4.2
T	threonine	119.1197	-0.7
C	cysteine	121.159	2.5
I	isoleucine	131.1736	4.5
L	leucine	131.1736	3.8
N	asparagine	132.1184	-3.5
D	aspartate	133.1032	-3.5
E	glutamine	146.1451	-3.5
K	lysine	146.1882	-3.9
Q	glutamate	147.1299	-3.5
M	methionine	149.2124	1.9
H	histidine	155.1552	-3.2
F	phenylalanine	65.19	2.8
R	arginine	174.2017	-4.5
Y	tyrosine	181.1894	-1.3
W	tryptophan	204.2262	-0.9

Table 1.1: The molecular weight (size in Dalton) and hydrophobicity values (KD) of common amino acids [34].

1.1.1 Primary Structure of Proteins

The amino acids form a peptide bond between each other as shown in Figure 1.2 thereby forming a sequence which is known as the protein's primary structure (Figure 1.3). When two amino acids combine like this to form peptide bonds, the reaction releases a water molecule and the amino acids are now referred to as amino acid residues or just residues. Protein polypeptide chains vary in length, e.g., 50 to 25000. However, most protein polypeptide chains contain 200 to 500 residues [14].

1.1.2 Secondary Structure of Proteins

Polypeptides have specific local conformations or secondary structures which depend on hydrogen bonding between the residues. There are two main types of secondary structures: α-helices and β-sheets which are linked by loops as shown in Figure

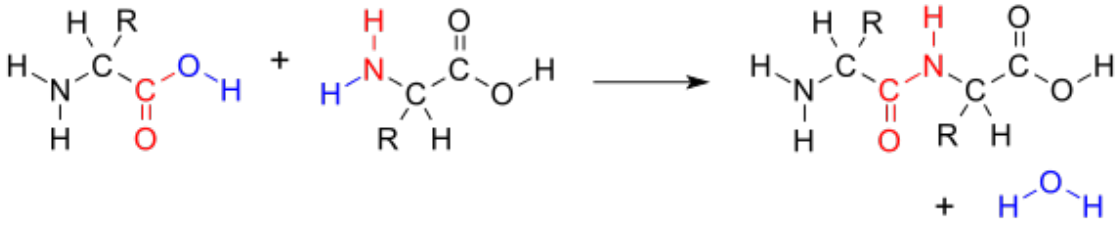

Figure 1.2: Peptide bond between the two residues [51]. During the formation of peptide bond a water molecule is released (blue).

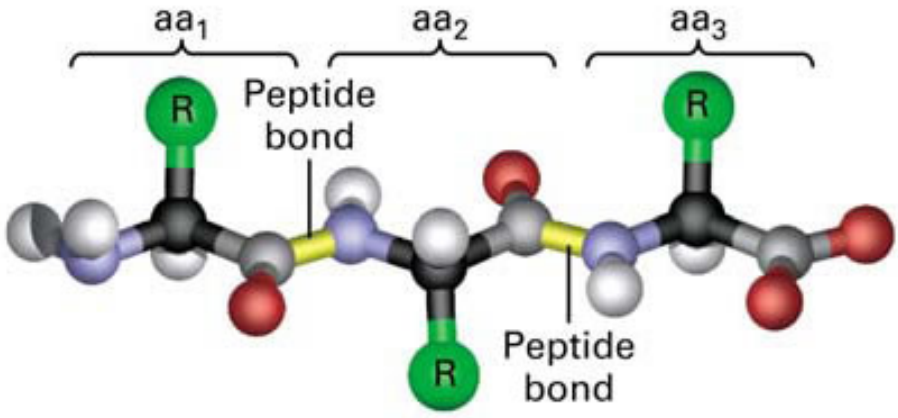

Figure 1.3: Primary structure of proteins [40]. This primary structure is a three residue polypeptide chain.

1.4. α-helices are a righthand-coiled conformation which is formed by donation of a hydrogen bond. β-sheets are pleated sheets connected laterally by two or three backbone hydrogen bonds. Helices are often represented by cylinders and coiled ribbons and β-sheets are represented by arrows.

Figure 1.4: Example of yeast enolase secondary structure (PDB ID: 1L8P chain: A). α-helices (red), β-sheets (yellow) and loops (green).

1.1.3 Tertiary Structure of Proteins

The residues and secondary structural features interact with each other forming the three-dimensional shape of proteins i.e., the tertiary structure. The folding of proteins allow for the interaction of residues that may be distant from each other in the primary sequence of the protein. This 3D structure is roughly spherical or partially compact for globular proteins [14]. The residues buried inside the protein core are primarily hydrophobic, so that they avoid contact with aqueous medium that most proteins generally exist within [55]. Acidic or basic residue sidechains will generally be exposed on the surface of the protein as they are hydrophilic. The structures of most of the proteins that have more than 200 residues have two, three, or more structural units called domains. Quite often, a single chain of polypeptides connects different domains in the protein. A domain is a conserved part of the protein structure that can evolve, function and exist independently. Small proteins tend to have only one domain and

Figure 1.5: Example of the tertiary structure of yeast enolase (PDB ID: 1L8P chain: A).

are grouped in the same group; large proteins, on the other hand, often have more than one domain and can be classified individually [44]. The definition of domain and the division of proteins into domains is very subjective and lacks clear rules. Figure 1.5 shows the typical tertiary structure of a protein.

1.1.4 Quaternary Structure in Proteins

Proteins can consist of multiple polypeptide chains, which can be either identical or different depending on the type and functionality of the protein. Different polypeptide chains are referred to as subunits, monomers, chains, or protomers. These subunits may be the same (homo) or different (hetero). The number of chains can vary, which makes a protein homodimer (two identical chains) or heterodimer (two different chains), homotrimer (three identical chains) or heterotrimer (three different chains) or even higher order combinations of identical or different chains. Such combinations of the chains make the quaternary structure of the protein [14]. The quaternary structures refers to how these chains interact with each other and arrange themselves to form protein complexes. Figure 1.6 shows a typical quaternary structure of a protein.

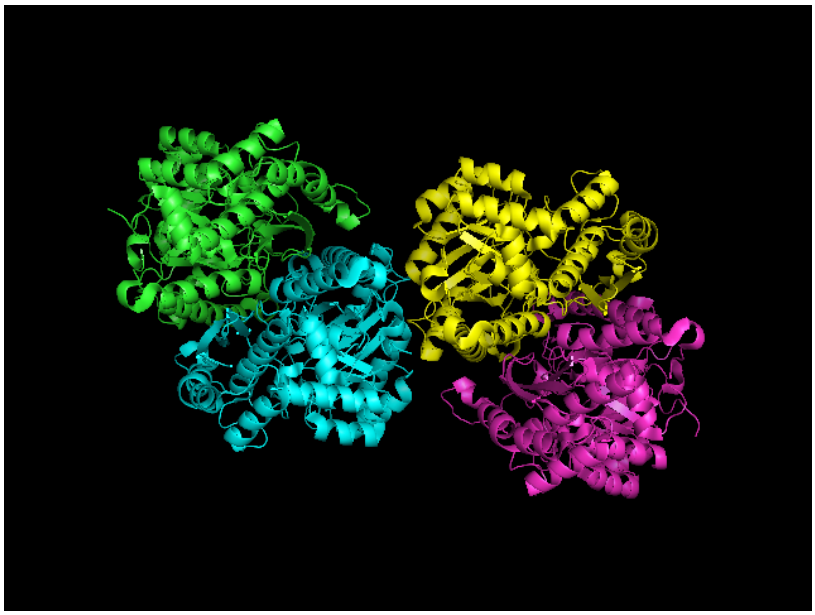

Figure 1.6: Quaternary structure of enolase1 (PDB ID: 2PSN). This structure comprises of 4 chains (A,B,C, and D) coloured green, blue, yellow, and pink respectively.

1.1.5 Protein Structure and its Relation to Functionality

The 3D structure of proteins plays an important role in determining their function, and it is the amino acid sequence of the protein that determines the function and the structure of a given protein. One characteristic that affects the function of a protein is its hydrophobicity; determined largely by primary and secondary structure, for example, the regions of membrane proteins that interact with lipids (hydrophobic in nature) mostly comprises of hydrophobic residues (water repelling) or the mutated haemoglobin in red blood cells found in sickle cell disease have high hydrophobicity which causes the protein molecules to stick to each other. Most proteins perform their specific function when they are folded into an ordered and stable structure called its native state. The sequence of changes that the protein undergoes to reach its native structure is known as protein folding. The native state can be perturbed by a number of external factors such as temperature, pH, absence of water as solvent, presence of a hydrophobic surface such as a membrane, and presence of metal ions [6].

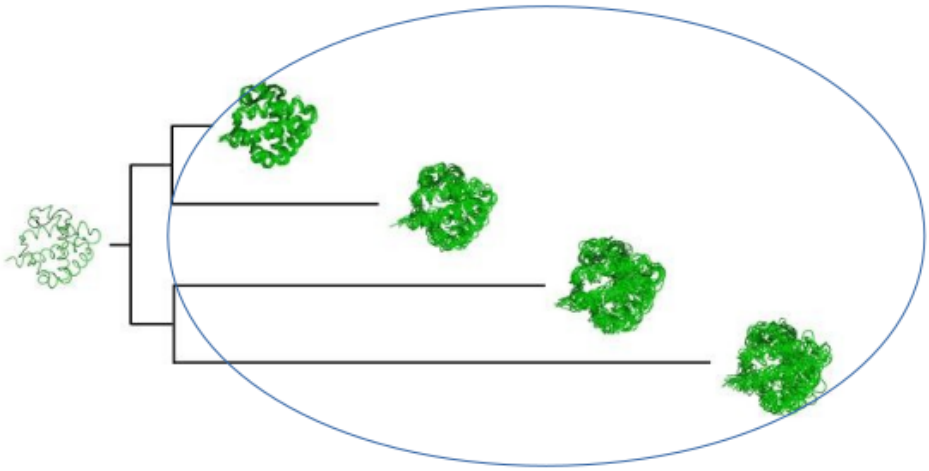

Figure 1.7: The blue circle shows homologous structures that have evolved from the same ancestor (root).

1.2 Homologous Proteins

The protein structures that have common ancestry across the evolutionary timeline [45] are said to be homologous proteins. They may have statistically significant similarity. For example, the three-dimensional structure of horse and human haemoglobin is homologous even though 43 of their residues are different [14]. Figure 1.7 shows the evolution of protein sequences. The similarity of the sequence is determined by a sequence alignment. There are numerous software programs that are available for aligning the sequences, e.g., BLAST [4] and HMMER [22].

1.3 Families, Subfamilies, and Subgroups in Proteins

Evolutionarily related proteins may be grouped together in a family. Some studies suggest that if two protein sequences share 30% sequence similarity or have similar structure or similar function, they are usually grouped in a family [44]. A superfamily is comprised of families that have different sequences, but the similar structure and

function of these families suggest the possibility of common evolutionary origin. If the secondary structures are in the same arrangement for families and superfamilies, then these are grouped in common folds. The folds are grouped into classes for convenience [44]. For example, in this research, we examine the enolase superfamily of enzymes which consists of a common fold called a TIM barrel. This superfamily is composed of members with high sequence and structural similarity (see Section 1.6).

1.4 Types of Databases

Known protein structures are generally referred to by their PDB ID. The Protein Data Bank (PDB) is a repository of information about the structures of large biological molecules. Currently the Research Collaboratory for Structural Bioinformatics (RSCB) is responsible for maintaining the repository [18]. The data is available publicly as a free open-access resource.

The Structural Classification of Proteins (SCOP) database for the investigation of sequences and structure provides a detailed and comprehensive description of structural and evolutionary relationships of proteins whose three-dimensional structures have been determined. It includes all the proteins in the current version of PDB. The proteins in SCOP are classified by visual inspection and comparison. The unit of classification in SCOP is the protein domain, that is the conserved part of the protein which can fold independently from the rest of the protein.

For this research, the Structure-Function Linkage Database (SFLD) was used to identify and classify protein structures. The SFLD is a manually curated database that classifies functionally diverse enzymes on the basis of structure-function relationships. In other databases, some of the members of a superfamily may seem to be homologous or evolutionarily related due to fact that they share all or part of their functions. This sometimes leads to misannotation of these members. SFLD tackles such misannotation by manual curation of the superfamilies of enzymes with diverse functionality [3]. In the SFLD, the enzymes are divided into superfamilies that may or may not have functional and evolutionary relationships. On the basis of sequence, these superfamilies are divided into subgroups. Enzymes belonging to one subgroup

are evolutionarily related enzymes and have more shared features than the superfamily as a whole, however the enzyme may still catalyse different reactions. Lastly, the database is divided into families, if two enzymes perform same function using the same mechanism they are categorized into the same family [3].

1.5 Structural Alignment of Proteins

Structural alignment of protein structures is important to understand the underlying similarity of the structures. There are multiple software programs available for preparing structural alignments of proteins. Multiple Alignment with Translations and Twists (MATT) is among the most commonly used [41]. MATT is an aligned fragment pair chaining algorithm that allows for local flexibility between the fragments. Local flexibility implies that small translations and rotations are allowed to bring the set of aligned fragments closer using dynamic program assembles. It superimposes mostly the backbone of a protein structure in close spatial alignment to every other structure. In a rigid body transformation, such flexibility is not possible [41]. MATT considers fragments of five to nine amino acid residues for each structure. For each fragment pair from two different structures, an alignment score (based on the p-value of the minimum RMSD (root mean square deviation) achievable by the rigid body transformation) is calculated. MATT builds up sets of aligned fragments of increasing length using dynamic programming. Structural alignment can reveal areas that have a high degree of superimposition: such areas can indicate functionally important parts of proteins such as catalytic sites or highly conserved structural folds. Figure 1.8 shows the structural alignment of mandelate racemases using MATT and visualized using PyMOL [17].

1.6 Enolase Superfamily

The enolase superfamily comprises a group of enzymes which catalyse a range of reactions, yet are related by their common partial chemical mechanism: abstraction of the α-proton of a substrate to form an enol(ate) intermediate. The catalysis by enzymes of this superfamily is performed by a conserved set of residues located

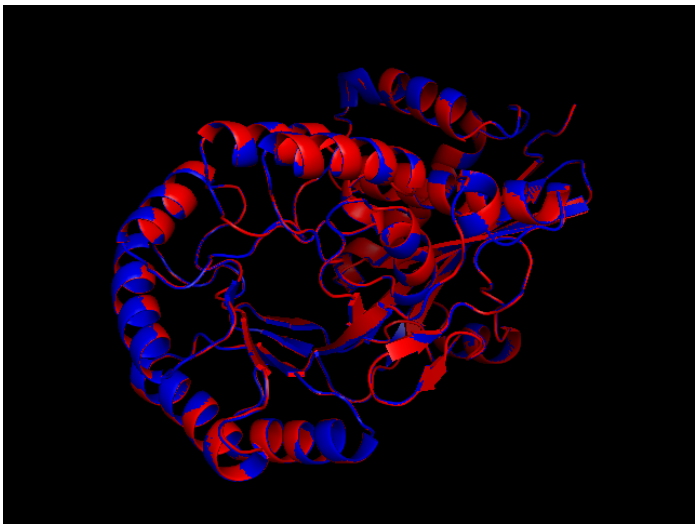

Figure 1.8: Structure alignment of proteins using MATT. The image of alignment was generated using the visualization software PyMOL and the proteins aligned are mandelate racemases (1MDR chain A (red) and 2MNR chain A (blue))

in the active site. Enzymes in this superfamily have an important structural domain called the triosephosphate isomerase (TIM) barrel - a conserved protein fold consisting of eight α-helices and eight β-sheets [3, 57]. The enolase superfamily includes enolase as well as other metabolically specialized enzymes which are categorized into other subgroups: mandelate racemase, galactarate dehydratase, glucarate dehydratase, muconate-lactonizing enzyme, β-methylaspartate ammonia-lyase, and D-mannoate dehydratase [8]. One important member of the superfamily is enolase, also known as phosphopyruvate hydratase. It acts as a catalyst in the penultimate step of glycolysis (breakdown of glucose into pyruvate molecules). Like other members of the superfamily, it also contains a TIM barrel. Although the TIM barrel fold is common to many other superfamilies, none of these superfamilies have as significant level of sequential and functional similarity as the enolase superfamily [8]. Since the enzymes belonging to this superfamily catalyse a range of imperative reactions, numerous studies have been conducted on this superfamily [48, 58]. The enolase superfamily is known to be mechanistically diverse and studies have been conducted to assign functions to its members which can be used as a template to assign functions to other superfamilies [26]. The diversity and number of enzymes included in the enolase superfamily were one of the prime reasons of choosing it for this study. The number of structures in enolase superfamily increased the breadth of our sampling.

Figure 1.5 shows the structure of enolase from yeast (PDB ID: 1L8P chain: A).

1.7 File Format of Protein Structures

The secondary and tertiary structures of proteins are represented in the form of Cartesian coordinates with each atom of an individual amino acid having its respective x, y, and z coordinates. These coordinates, along with variables such as the occupancy (conformation of side chain or main chain atoms) and the temperature factor (displacement of atomic position from the mean value) are stored in a file format called PDB (http://www.rcsb.org/pdb/home/home.do). The PDB file format contains metadata of the protein structure. It also gives the position of ligands that are bound to the protein, which include ions, water molecules, and substrates. The primary structure of the protein is a sequence of amino acids or the residues that can be computationally represented in the form of a string. The FASTA file format of proteins contains the sequence of residues that make up the protein. FASTA files can also represent alignments both sequential and structural, where '-' is used to represent gaps in the alignment. A gap simply means that no matching residue was found in a given position of the query sequence when aligned to a reference sequence.

Using Python scripts, the FASTA file format can be parsed into a string to perform an analysis. PDB files, however, require more elaborate parsing. Some PDB files do not have a standard format hence retrieval of the coordinates of residues or ligands requires extensive parsing. The presence of blank spaces and the non-conventional format of the meta data can also make parsing challenging. In addition to the FASTA and PDB formats, the protein can also be represented in XML format. The description of the XML format is provided in the XML schema of PDB Exchange Data Dictionary [56].

1.8 Contact Matrices and Contact Entropy

1.8.1 Contacts

The structure of proteins can be represented as a three-dimensional model. As described in Section 1.1.3, residues interact with each other by forming contacts, which

gives shape to the protein structure. In the present study the word contacts is used to refer to the distance between two residues due to London-van der Waals forces [7] which can be used to investigate the protein structures. The definitions of residue contacts used in literature are very diverse and various researchers have used multiple definitions. In some cases, the residues are said to be in contact if one of the atoms of two residues are at a distance of ≤ 5 [7]. This makes the contact calculation quite sensitive because all of the atoms of a residue are involved in the determination of contact. The cut-off threshold distance to determine a contact depends on the method of inter-residue distance calculation. In some studies, the α-carbon of the residues (first carbon that attaches to functional group) is used to calculate the distance, with a cut-off threshold of 7 [9]. In other studies contacts are determined by the β-carbon of the residue (second carbon atom that attaches to functional group), with a cut-off distance of 8.5 [7]. Measurement of distances using α and β-carbons are fast to compute due to the fact that only one distance comparison is made. The centroid of a residue can also be used; however, as stated earlier, to increase the sensitivity of the contacts, the minimum distance between any of the atoms from the amino acid is used. The cut-off in such cases is ≤ 5 [42] [2] or 4.5 [37]. Using all of the atoms of the residue or the centroid requires more comparisons and calculations.

1.8.2 Contact Matrices

The contacts between the residues may be used to construct a contact map or contact matrix. A contact matrix is a binary matrix M where $M_{ij} = 1$ if there is a contact between residue i and residue j, or else $M_{ij} = 0$. Contact matrices may also be used as an adjacency matrix to generate amino acid networks or protein contact networks. Applications of contact matrices are diverse. For example, they may be used for prediction of 3D structures of proteins which is difficult using only the residue sequence [53]. A heuristic algorithm called contact map reconstruction (COMAR) has been used for this purpose [53]. Some studies have also employed contact matrices to cluster the contacts in order to identify secondary structures. The folding of proteins allows for the interaction of residues that may be distant from each other in the primary sequence of the protein. These clusters of contacts are able to capture non-local interactions to aid prediction of tertiary structures [29]. Contact matrices

can also be used to analyze protein folding pathways (the physical process by which a protein acquires its native 3D structure) [29], interactions at protein-protein interfaces (when two proteins interact as a result of biochemical event caused by electrostatic forces such as hydrophobic effects) [16], protein dynamics by identifying core residues [7], and key functional residues in protein structures, i.e., the residues that act at the active site of enzymes [5].

Some studies convert contact matrices into contact networks, residue interaction graphs, or protein contact networks [5]. In contact networks, residues act as nodes and contacts between them are edges. These networks are usually undirected and possess small-world properties, i.e., each node (residue) can be reached from other nodes by passing along a small number of edges (contacts) [20]. Graph properties such as degree distribution (probabilities of number of contacts associated with each residue over the whole protein network), shortest path length and average shortest path (average number of steps along the shortest path for all possible pairs of residues in the network), clustering coefficient (measure of degree to which residues in the protein network tend to cluster together), and closeness centrality of residues have been used to identify functionally important residues. Contact matrices have also been used to create a hierarchical classification of amino acids in a protein network into successive layers from the core (having high density and contacts) to the periphery (having low density and contacts) of the protein [33].

The diverse implementation of contact matrices makes them essential in the field of structural bioinformatics. However, the information provided by a contact matrix encompasses the structural property of a single protein. In this study, we explored the use of contact matrices to characterize sets of homologous protein structures which have common ancestry. This new contact matrix calculated over a set of homologous structures is called a frequency contact matrix (FCM). The calculation of a FCM is discussed in the methodology section (see Section 2.5). The examination of contacts in the entire homologous set of proteins represents the uncertainty of residue-residue contact across a sample of related protein structures. The contact information of all the structures in a homologous set of proteins makes FCMs informative as compared to single contact matrices.

1.8.3 Contact Entropy

Entropy is a thermodynamic concept which refers to the degree of uncertainty or disorder within a system. Shannon entropy is an information theory interpretation of entropy [49] and is often used as a measure of unpredictability of a state. Shannon entropy has been used for various applications of bioinformatics where disorder needs to be analyzed. It is used to determine the information content of protein sequences, where the most probable protein sequences derived from the substitution of amino acids can be calculated [50]. In this study the contacts of structurally aligned residues are examined and the concept of Shannon entropy is used to measure the conserved contacts across the dataset. Shannon entropy for each structurally aligned site is calculated using the values of a FCM. Low entropy values suggest that the contacts of structurally aligned sites are similar in all the structures. On the other hand, high entropy values imply that contacts are non-existent in some structures and variable across the dataset. In other words, high contact entropy residues have high side chain freedom across the dataset in comparison to low contact entropy residues.

Predicting contact maps is a challenge in the field of bioinformatics. There are numerous machine learning algorithms or evolutionary approaches used to predict the contact matrix of a protein structure [39]. Some studies use the physicochemical properties of amino acids to predict the contact matrices such as hydrophobicity, polarity, charge, and size [13]. The prediction model in such studies is based on an evolutionary algorithm and consists of a set of rules. These rules obtained from the model impose a set of conditions on amino acid properties to predict the contacts [13, 39]. Similar to our method, these studies also use the Kyte-Doolittle hydropathy profile for the hydrophobicity of amino acids. The values of hydrophobicity are normalized to a range between -1 and 1. Such studies indirectly indicate a relationship between the contact matrices and physicochemical properties of amino acids. In this study, we also aim to analyze this relationship. We aim to observe the relationship between contact entropy values of residues and their properties. We hypothesize that a significant relationship exists between the contact entropy values of residues and their respective hydrophobicity and size.

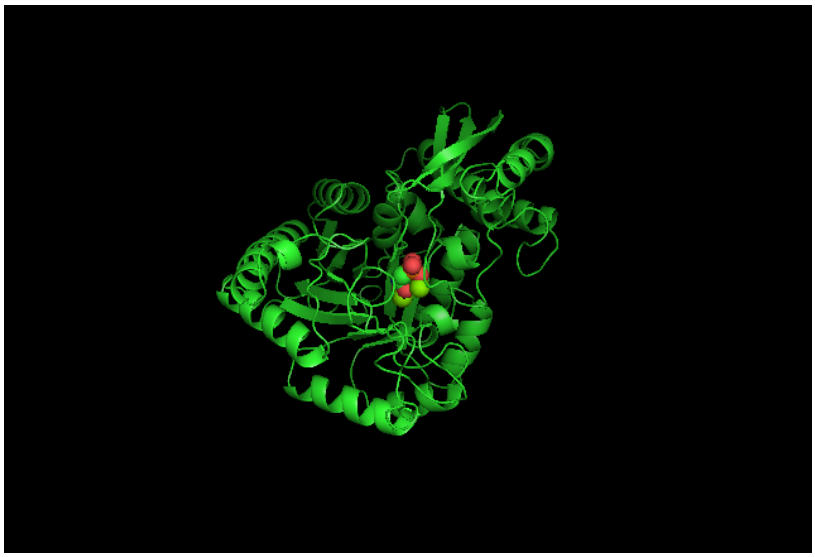

Figure 1.9: The structure of yeast enolase (PDB ID: 1EBG). The ligands (Mg^{2+} and Phosphonoacetohydroxamic acid) are shown in space-filling representation.

1.9 Ligands and Ligand-binding in Proteins

The function of a protein is dependent on its interaction with other molecules. Ligands are generally small molecules that form a complex with proteins. A ligand can also be another protein, or an inorganic ion such as manganese (Mn^{2+}) or magnesium (Mg^{2+}). Most ligands interact at specific sites on proteins, indeed often only one site per polypeptide chain. Figure 1.9 shows a ligand-bound to a protein. Ligands may bind at the surface of proteins or in deep clefts [14]. The sites where the ligand binds to a protein is called the binding site. Such binding sites are known as an active site if the protein is an enzyme and catalyses the chemical change of the ligand to the product. Thus, in enzymes these active site residues are where the enzymatic reaction takes place. The residues located at the active site are very specific [1]. The binding of a ligand may change the conformation of a protein's structure [21]. A ligand binds to a protein by intermolecular forces such as ionic bonds, hydrogen bonds, and van der Waals forces (distance-dependent interaction between atoms). Ligand binding can be characterized by binding affinity, and high-affinity binding can produce the binding energy required to effect a change in the conformation of a protein [1]. As stated in the previous section, protein contact networks are a direct implementation of the contact matrices. Variation observed in these contact networks are a reflection

of the variation in the contact matrix of the structure. It is shown through various studies that ligand binding affects the protein contact networks [21]. Studies have found that ligands act as network bridges in protein contact networks connecting components of protein contact graphs together [30]. These contact networks were used to identify functionally important residues using graph centrality. It is often found that the highly central residues are conserved and in close proximity to ligands [19, 20, 33]. In other studies, it was found that the central residues were often active site residues or residues in close proximity of active sites [15]. This implies that the most important residues in the contact network are the ones that interact with ligands. Ligands bound to an enzyme were also shown to affect the closeness centrality of the protein network [19]. Some studies have also suggested that ligand binding affects the flexibility of certain parts of proteins wherein some parts become stiffer than others [21] and therefore may alter residue contacts. The residue contacts in the stiff part of the proteins do not fluctuate significantly [21] and have fixed contacts. This analysis indicates that ligand binding makes some contacts in the contact map more stable than the other contacts hence changing the way a contact map looks for ligand-bound protein structure. Our study focuses on using residue contact entropy as a measure to determine the effect of ligand binding on the residue contacts with a goal of analyzing the conservation of residue contacts across the set of homologous protein structures. We hypothesize that the contact entropy value distribution of residues close to the active site or otherwise (see section 2.8) for ligand-bound and ligand-free structures are different.

1.10 Statistical Analysis using K-S test

The Kolmogorov-Smirnov or K-S test is a non-parametric test that compares probability distributions . I has been in used in various studies of structural bioinformatics suc as analyzing the protein network properties. For example, K-S test has been implemented to show the difference between the shortest path length and clustering coefficient for real protein networks and randomly generated networks [19]. In this study we also use K-S test to compare the difference between distributions. The

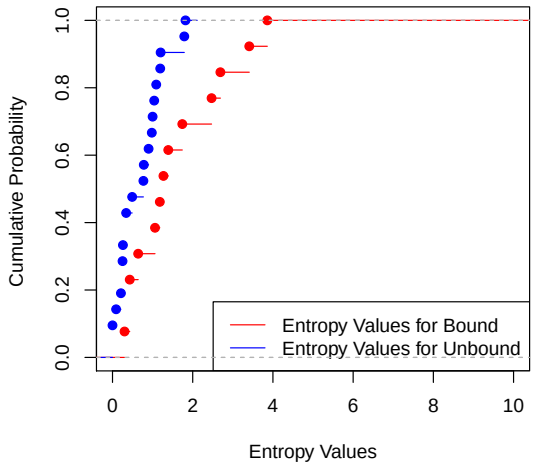

Figure 1.10: ECDF of two distributions shown in red and blue

reason of using K-S test for statistical analysis is that the distribution of residue contact entropy values of protein structures is not known and is continuous in nature. K-S test is a robust non-parametric test which does not depend on non normality of the distributions [27]. For a set of data points $x_1, x_2, x_3, ...x_N$ empirical cumulative distribution function (ECDF) is defined as:

$$E_N = n(i)/N \tag{1.1}$$

where $n(i)$ is the number of values that are greater than x_i. Figure 3.10 shows the ECDF for two distributions. K-S test is based on the maximum vertical distance between the two curves. The null hypothesis of the test is that the two distributions are identical, i.e., the sample distribution does not differ from the reference distribution. The p-value obtained from the test is then used to determine the significance of our results. A small p-value (typically < 0.05) indicates sufficient evidence to reject the null hypothesis [23].

1.11 Our Contributions

In this chapter, we have discussed the basics of protein structures, homology, ligand-binding at the active sites, and contact matrices. Contact matrices may be used to

analyze a variety of characteristics related to protein structure and function, including the formation of protein contact networks, determination of functionally important residues, effects of ligand binding, protein-protein interactions, prediction of tertiary structures, etc. While contact matrices have diverse applications, our study aims to extend the use of contact matrices to FCM which incorporates the contact information of homologous sets of structures. In the subsequent chapter, we describe the calculation of the FCM and contact entropy. In Chapter 3, we show the results of applying this approach to the enolase superfamily dataset. We describe the use of frequency contact matrices to find relationships between residue contact entropy and different properties of amino acid residues. We also show effect of ligand binding on side chain freedom of conserved contacts using contact entropy as a measure. In the final chapter, we discuss possible extensions of our approach using alternative methods.

Chapter 2

Methodology

In this section, the compilation of data, calculation of contact matrices, and calculation of frequency contact matrices is described. This section also provides the details for the calculation of the residue contact entropy and the method to statistically analyze results. An explanation of the simulation of data that was conducted to test the sensitivity of the K-S test employed to determine the statistical significance of real datasets is also provided. An overview of the workflow is shown in Figure 2.1.

2.1 Dataset

2.1.1 Database

The enolase superfamily was chosen for study because it is a widely studied dataset, comprises multiple subgroups and structural information is avalible for a number of its members. To choose the structures in the enolase superfamily and divide them into subgroups, the Structure-Function Linkage Database (SFLD) was used [3]. As described in Section 1.4 the SFLD is a manually annotated database which increases the reliability of datasets obtained from it. The enzyme structures are divided into subgroups that are homologous. The enzyme structures were downloaded from the RCSB PDB (http://www.rcsb.org/pdb/home/home.do) in PDB format and XML format (see Section 1.7).

2.1.2 Compilation of the Description File

XML formatted PDB structure files were parsed to retrieve a variety of properties characteristic of the proteins (see below). The information from parsing was saved into a CSV (comma separated values) file, description file. This file holds the description of each structure, subgroup, PDB ID of the structure, number and names of chains, organism to which the protein structure belongs, and the number and name of ligands

bound to each chain.

2.1.3 Division of Dataset

The enolase superfamily is divided into seven subgroups by the SFLD: mandelate racemase, mannonate dehydratase, glucarate dehydratase, *o*-muconate cycloisomerase, enolase, D-galactarate dehydratase, and β-methylaspartate ammonia-lyase. The number of structures in the three subgroups examined in the present work is shown in Table 2.1. The remaining four subgroups consisted of less than 45 quaternary structures and were therefore eliminated from the dataset. Most structures in each subgroup were quaternary structures and consisted of multiple polypeptide chains that may or may not be structurally similar to each other. The structures in each subgroup belong to different organisms. For each subgroup, the multimeric structures were divided into individual chains or monomers which were then treated as individual samples in the final dataset. Chains were segregated on the basis of the presence or absence of ligands. If the monomer had only Mg^{2+} and Mn^{2+} ion bound at the active site, the chain was categorized into a ligand-free $\mathcal{S}^{\mathcal{U}}$ set. However, other monomers or chains had other molecules bound to their Mg^{2+} and Mn^{2+} ions, and such structures were added to the set of ligand-bound structures $\mathcal{S}^{\mathcal{L}}$. The determination of ligands binding to the active site was done manually using RCSB PDB (http://www.rcsb.org/pdb/home/home.do) [18]. The ligand information for each structure stored in description file was also used to determine if the structures are ligand-bound or ligand-free (unbound).

Subgroups	Total no. of structures	No.chains in $\mathcal{S}^{\mathcal{L}}$	No. of chains in $\mathcal{S}^{\mathcal{U}}$
enolase	66	68	43
mandelate racemase	46	65	38
muconate cycloisomerase	77	78	63

Table 2.1: Number of structures in each subgroup of the enolase superfamily included in this analysis, where $\mathcal{S}^{\mathcal{L}}$ is the set of structures that have a ligand-bound at its active site and $\mathcal{S}^{\mathcal{U}}$ is the set of structures that are ligand-free.

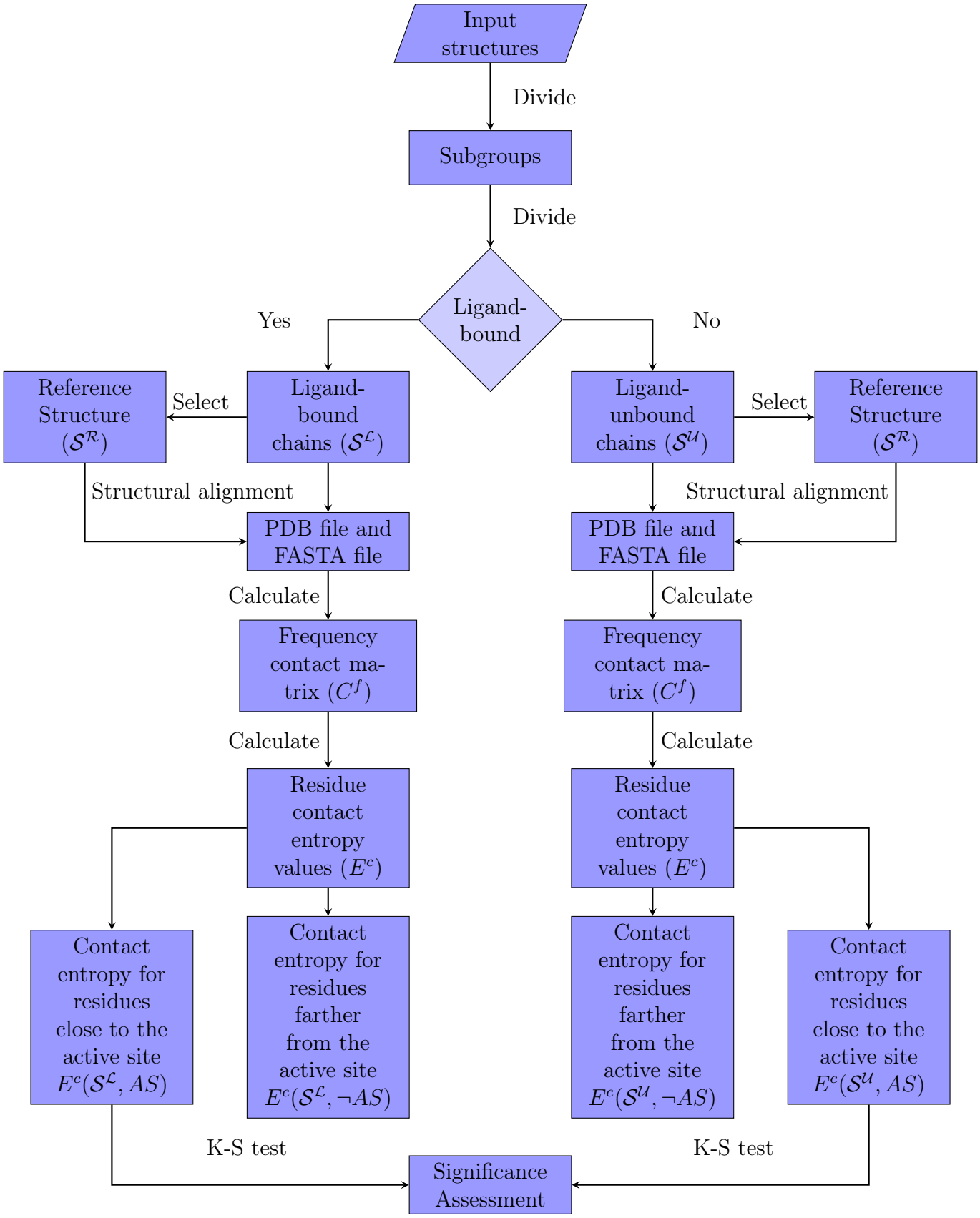
Input structures
Divide
Subgroups
Divide
Ligand-bound
Yes
No
Reference Structure ($\mathcal{S}^{\mathcal{R}}$)
Select
Ligand-bound chains ($\mathcal{S}^{\mathcal{L}}$)
Ligand-unbound chains ($\mathcal{S}^{\mathcal{U}}$)
Select
Reference Structure ($\mathcal{S}^{\mathcal{R}}$)
Structural alignment
Structural alignment
PDB file and FASTA file
PDB file and FASTA file
Calculate
Calculate
Frequency contact matrix (C^f)
Frequency contact matrix (C^f)
Calculate
Calculate
Residue contact entropy values (E^c)
Residue contact entropy values (E^c)
Contact entropy for residues close to the active site $E^c(\mathcal{S}^{\mathcal{L}}, AS)$
Contact entropy for residues farther from the active site $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$
Contact entropy for residues farther from the active site $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$
Contact entropy for residues close to the active site $E^c(\mathcal{S}^{\mathcal{U}}, AS)$
K-S test
K-S test
Significance Assessment

Figure 2.1: Overview of the our method. Input structures were divided into subgroups and then into $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$; the structures in each of the dataset were structurally aligned using MATT. The contact matrix and frequency contact matrix for each dataset is calculated. The residue contact entropy was then measured for reference structure in each dataset. These residue contact entropy values were compared using the K-S test.

2.1.4 Determination of the Location of the the active site

This study aimed to analyze the side chain freedom of residues in the structure that are close to the active site and farther from the active site. Thus it is important to determine the location of the active site in each structure. The position of either magnesium ion (Mg^{2+}) or manganese ion (Mn^{2+}) was used to determine the coordinates of the active site since these metal ions are essential for catalysis in the enolase superfamily. To maintain consistency the structures that did not have the above stated metal ions were eliminated from the dataset. Such structures constituted only 5-7% of the dataset.

2.2 Reference Structures

A reference structure was randomly chosen for the ligand-bound set $\mathcal{S}^{\mathcal{L}}$. The reference for $\mathcal{S}^{\mathcal{U}}$ was the ligand-free form of the chosen reference structure for $\mathcal{S}^{\mathcal{L}}$. Hence, for each subgroup, the reference structure for $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ was the same protein derived from the same organism in its ligand-bound and ligand-unbound form respectively. The reference structures are compiled in Table 2.2. One of the chains in these structures acts as the reference $\mathcal{S}^{\mathcal{R}}$ for structural alignment. In this research, chain A of all the structures was chosen as the reference structure.

2.3 Structural Alignment

For each subgroup, the protein structures for both the ligand-bound dataset $\mathcal{S}^{\mathcal{L}}$ and the unbound dataset $\mathcal{S}^{\mathcal{U}}$ were structurally aligned using the pairwise alignment tool

Subgroups	PDB ID $\mathcal{S}^{\mathcal{L}}$	PDB ID $\mathcal{S}^{\mathcal{U}}$
enolase	1ELS	1EBH
mandelate racemase	1MDR	2MNR
muconate cycloisomerase	2P8B	2P88

Table 2.2: PDB IDs of reference structures ($\mathcal{S}^{\mathcal{R}}$) of all subgroups for the ligand-bound and ligand-unbound datasets. These reference structures are used for structural alignments.

MATT [41] (see Section 1.5). This alignment generated the set of structurally aligned sites in $\mathcal{S}^{\mathcal{R}}$. The alignment produced a FASTA format file and a PDB format file. The PDB file contained the Cartesian coordinates for all the structures in the dataset. The FASTA file was used to determine structurally aligned sites, i.e., the residues that are located at structurally aligned positions.

2.4 Calculation of the Contact Matrix

A contact matrix (C) is the representation of the 3D structure of proteins in the form of contacts that exist between each of the residues. There have been numerous studies conducted using contact matrices to find the functionally important residues of proteins [33]. The matrices can also be mapped back to the tertiary structure of the protein. Such matrices can be generated by parsing the PDB format file of the protein structure and extracting all the Cartesian coordinates, and then testing the coordinates for contacts. Contacts are determined by the distance between the residues. There are numerous ways to calculate the distance as described in Section 1.8.1. However, in this study residue i and residue j are said to be in contact if the distance $\mathcal{D}$ between any of their atoms is ≤ 4.5 [42]. The choice of 4.5 as a threshold was also empirically determined in the lab in past as being an optimal atom to atom distance. This is a more sensitive way of calculating the contacts. If the protein has r residues, then the contact matrix C will have the dimensions $r \times r$. After structural alignment, only structurally aligned residues or positions were taken into consideration. The structurally aligned sites or residues for each dataset are defined as strictly gap-less positions in the structural alignment. Hence, for a dataset $\mathcal{S}^{\mathcal{U}}$ with k aligned positions, all C will be $k \times k$ matrices. A contact matrix is a binary matrix.

All the residues are tested for contacts with each other; if residue i and residue j are in contact with each other then the value of $C_{ij} = 1$ or else $C_{ij} = 0$.

2.5 Calculation of the Frequency Contact Matrix (FCM)

A frequency contact matrix (FCM) is organized in a similar manner as that of the contact matrix, but each matrix entry contains the frequency of the contact for a set of structurally aligned residues. If there are n protein structures in a dataset $\mathcal{S}^{\mathcal{U}}$, a contact matrix C is calculated for each structure. The contact matrices are aggregated by adding all the respective contacts to create an aggregate contact matrix C^a using equation 2.1.

$$C^a = \sum_{i=0}^{n-1} Ci \tag{2.1}$$

The aggregate matrix C^a is normalized to [0,1] range to calculate the frequency contact matrix C^f in accord with equation 2.2.

$$C^f = \frac{C^a}{n} \tag{2.2}$$

2.6 Calculation of the Residue Contact Entropy

Residue contact entropy, E^c, is used to identify residues with an unusual level of side chain contact freedom. Shannon contact entropy of a residue i serves as an estimate of the uncertainty of contacts of that particular residue. The contact entropy for a residue i is calculated using equation 2.3, where n is the number of structures in the dataset and C^f_{ij} is the value of C^f for residue i and residue j.

$$E^c_i = -\sum_{j=0}^{n-1} (C^f_{ij}) log_2(C^f_{ij}) \tag{2.3}$$

High value of E^c_i indicates that the contacts of structurally aligned residue i are not stable across the dataset. Hence the sidechains of such residue forms different contact across the dataset.

2.7 Calculation of the Residue Sequence Entropy

The structurally aligned sites or structurally homologous sites are tested for their sequential variability using sequence entropy. The residues are represented by their one letter amino acid code in a FASTA format file. The set aa is a set of 20 amino acids $A,C,D.....Q,P,Y$. The sequence entropy E^s for each aligned position i in n number of structures is calculated using equation 2.4, where $count(aa,i)$ is the count of a particular residue at aligned site i.

$$E_i^s = -\sum_{aa}(\frac{count(aa,i)}{n})log_2(\frac{count(aa,i)}{n}) \tag{2.4}$$

High value of E_i^s indicates that the residue at structurally aligned site i are aligned to different types of residues.

2.8 Determination of the Residues Located Close to the Active Site

If the Euclidean distance $\mathcal{D}$ between the coordinates of either the Mg^{2+} or Mn^{2+} ion and the coordinates of residues in the structure ($\mathcal{S}^{\mathcal{R}}$) is less than or equal to a certain determined distance $\mathcal{D}^{\mathcal{A}}$, then the residue is said to be a component of the set of residues that are close to the active site. Alternatively, they become part of the set of residues located farther from the active site. The value of $\mathcal{D}^{\mathcal{A}} \leq 10$ was chosen by visually analyzing the surface area covered by various distances. Figure 2.2 shows the surface area covered by residues at a distance of 10 and 15 . 10 sufficiently covers the active site and its boundary confines to the residues within the TIM barrel. However, to understand the relationship of $\mathcal{D}^{\mathcal{A}}$ with contact entropy distribution, the values of $\mathcal{D}^{\mathcal{A}}$ ranging from 8 to 15 were examined.

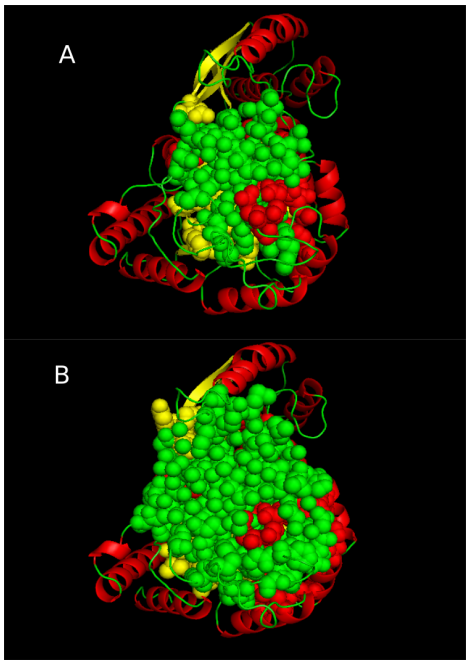

Figure 2.2: Surface area covered by the residues that are at a distance of **A:** 10 and **B:** 15 for yeast enolase (PDBID: 1ELS Chain: A). The structure is visualized using PyMOL.

2.9 Statistical Analysis

The difference between the entropy distributions was evaluated using the Kolmogorov-Smirnov test (K-S) test. A non-parametric test was used because the distribution of the entropy values was unknown. The distributions are considered significantly different from each other if the p-value is ≤ 0.05 (see Section 1.8).

2.10 Simulation of Data

To check the sensitivity of the K-S test, a controlled set of data was simulated. Simulated aggregate matrices $C^{a'}$ were generated as follows:

- The number of simulations was set to $r = 100$ and the standard deviation (σ) values ranged from 0 - 1 at an interval of 0.05
- Generate random values for $C^{a'}$

  - The value of $C_{ij}^{a'}$ between residue i and residue j is shown in equation 2.5, where $\lfloor gauss(0,\sigma) \rfloor$ is a random value from a Gaussian distribution with mean equal to 0 and a standard deviation equal to σ. If the value of $\lfloor gauss(0,\sigma) \rfloor$ is less than 1 then the value is rounded to 0 (using a floor function).
  - The value of $C_{ij}^{a'}$ must be greater than 0 and less than the number of structures in the dataset ($0 < Ca'_{ij} <$ n)

- Calculate FCM $C^{f'}$ from $C^{a'}$ using equation 2.2.
- Calculate the contact entropy values $E_i^{c'}$ from $C^{f'}$ by applying equation 2.3.
- For each σ value repeat r times

The contact entropy values calculated using the simulated aggregate matrices $C^{a'}$ close to the active site were compared to real entropy values for residues close to the active site in each dataset using the K-S test.

$$C_{ij}^{a'} = C_{ij}^{a} + \lfloor gauss(0,\sigma) \rfloor \tag{2.5}$$

2.11 Scripting

The experiments were performed using the Python language environment [52]. Simulations of the matrices were also done using scripts written in Python. PDBnet; a Python 2.7 library was used to organize PDB files of enzyme structures into easily accessible data structures [11]. PDBnet; a collection of Python objects intended to model and contain PDB protein data [11]. The K-S test was performed in the R language environment for statistical computing [46]. Plots were generated using a Python library called Matplotlib [32].

2.12 Visualization of Structures

All the structures in this study were visualized using multiple open source tools such as PyMOL [17] and Visual Molecular Dynamics (VMD) [31].

Chapter 3

Results and Discussion

In this section we demonstrate the utility of using the FCM and residue contact entropies to analyze a set of homologous proteins. A set of structures belonging to the enolase superfamily were divided into their respective subgroups using the SFLD. They were then segregated into monomers (individual chains), and categorized as ligand-bound or ligand-free as described in Section 2.1.1. The contact matrix, aggregate contact matrix, and FCM were calculated using the equations presented in Sections 2.4 and 2.5. For each dataset, contact entropies of all of the structurally aligned residues with respect to the reference structure $\mathcal{S}^{\mathcal{R}}$ were calculated. In this chapter we describe the experiments that were conducted on FCMs, the relationship between contact entropy and different properties of residues, and the application of contact entropy as a measure of sidechain freedom for residues close to the active site (or otherwise) after ligand binding.

3.1 Residue Contact Entropy Mapping

While contact matrices give information about the structural semantics of a single protein stucture, the FCM encompasses the information of the entire homologous set of structures. Contact entropy values calculated from the FCM for each aligned residue in a structure serves as a measure of the contact freedom of side chains. Values of the residue contact entropy were observed to be variable across the protein structure. Higher contact entropy values imply that the aligned site or residue has higher side chain freedom, i.e., the side chain of the residue forms contacts with different neighbours, whereas lower contact entropy values indicates otherwise. The distribution of contact entropy values across the structure helped us determine if any relationship exists between the contact entropy values and the 3D position of the residues in the structure. To show the position of high contact entropy and low contact

entropy residues, the contact entropy values were mapped onto structures using VMD (see Section 2.12). The temperature factor for each atom was replaced by its contact entropy value, which was then used to colour the structure. The mapping shows (Figure 3.1) that for some of the residues located in the interior of the protein, higher entropy values are observed as opposed to polar or surface residues. However, as the mapping in the Figure 3.1 shows, the difference is not consistent. Some of the surface residues also have high entropy values (indicated by green in Figure 3.1). Regardless of the location of high entropy residues, the mapping indicates that sidechain contacts across structurally aligned sites are not always same. Mapping of the other subgroups is shown in Appendix A.1

3.2 Contact Entropy Values for Hydrophobic and Hydrophilic Residues

The visual observation of the residues located in the core of protein structure appear to have higher contact entropy values. The visual observation however was not adequate to establish the relationship between contact entropy values of residues, their 3D location, and their properties. In this section, we examine the possibility that contact entropy values of residues may be related to the physical properties of the amino acid residues and their 3D location in the structure. For a set of 20 residues, the contact entropy values of the whole structure were compared on the basis of residue identity. Figure 3.2 shows the box plot of contact entropy values of amino acids in order of decreasing hydrophobicity. From the plot it appears that the hydrophobic residues have higher median values of contact entropies on average relative to other residues (see Appendix A.3 for the individual subgroup plots). For further analysis, the residues were divided into hydrophobic and hydrophilic residues using Kyte-Doolittle (KD) hydrophobicity values [34]. Residues such as isoleucine, valine, leucine, phenylalanine, cysteine, methionine, and alanine have positive KD hydrophobicity values (see Table 1.1); hence, they were categorized as hydrophobic residues while the others were categorized as hydrophilic. Hydrophobic residues are generally located in the core of protein structures [55]. A box plot was generated for the entropy distributions for both sets of amino acid residues. The median entropy values of the hydrophobic residues (2.3) is higher than the median entropy of the

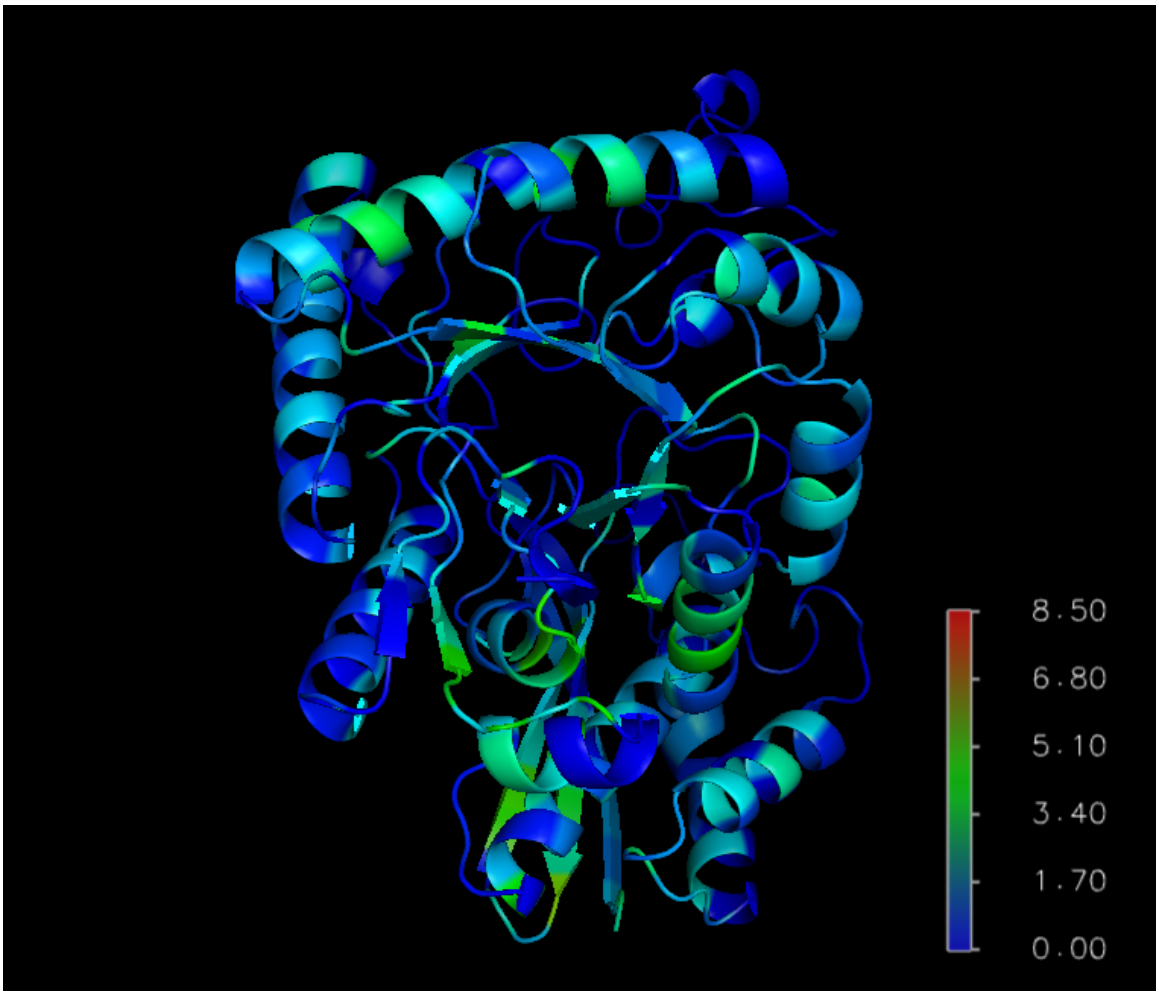

Figure 3.1: Contact entropy values (bits) mapped onto the reference structure in the enolase subgroup (PDB ID: 1EBG chain: A). The scale ranges from 0 bits to 8 bits (approx). Some of the high entropy values (green) are found in the interior of protein structure. The location of the high entropy residues is not necessarily in in the interior of the protein.

hydrophilic residues (1.5) (Figure 3.3). The box plot also shows that 50% of the data in hydrophobic residues is in the range of 1.5 to 3.9 while for hydrophilic residues the range of 50% data is 1 to 2.5. It can also be seen in the box plot that the upper whisker for the hydrophobic residues (3.5-6.5) is greater than that of the hydrophilic residues (2.5-4.5), which implies that the quartile group 4 (25% of values greater than upper quartile) is significantly higher for the hydrophobic residues. These results indicate that the contact entropy distributions of hydrophobic residues are not the same as those of hydrophilic residues.

The spread of the boxes do not statistically imply the difference in variance. To statistically analyze this we performed Levene's test which is a more robust form of F-test of equality of variance [36]. The F-test of equality of variance, which is normally applied to compare the variance of distributions, however, cannot be implemented on our distributions because it is very sensitive to non-normality of distributions and our data values are not normally distributed. Levene's test assesses the hypotheses that the population variances are equal. The p-value of Levene's test for hydrophobic and hydrophilic residue contact entropy values for aggregation of all the subgroups showed that the variance of hydrophobic contact entropy distribution is different from hydrophilic distribution ($p = 0.002$). When Levene's test was performed on individual subgroups, we found that for the enolase and mandelate racemase subgroup, the variance of hydrophobic and hydrophilic contact entropy values is not different with p-values equal to 0.79 and 0.46, respectively. The high p-value signifies that null hypothesis of Levene's test cannot be rejected and that the two distributions have same the variance. However, for the muconate cycloisomerase subgroup, the p-values are low ($p = 0.01$), implying a difference in variance. The aggregate dataset collected using all the subgroups is statistically more robust due to large sampling size. The box plots for the entropy values of hydrophobic and hydrophilic residues for each individual subgroup are shown in Appendix A.4.

To determine the difference in two entropy distributions we performed a Mann-Whitney U test. The test is used to test the distributions, especially central tendencies of the distributions (such as the median) [38]. One of the other reasons for using the Mann-Whitney U test instead of Wilcoxon signed-rank test was that the size of our distributions are not the same. For aggregate data values of all the subgroups

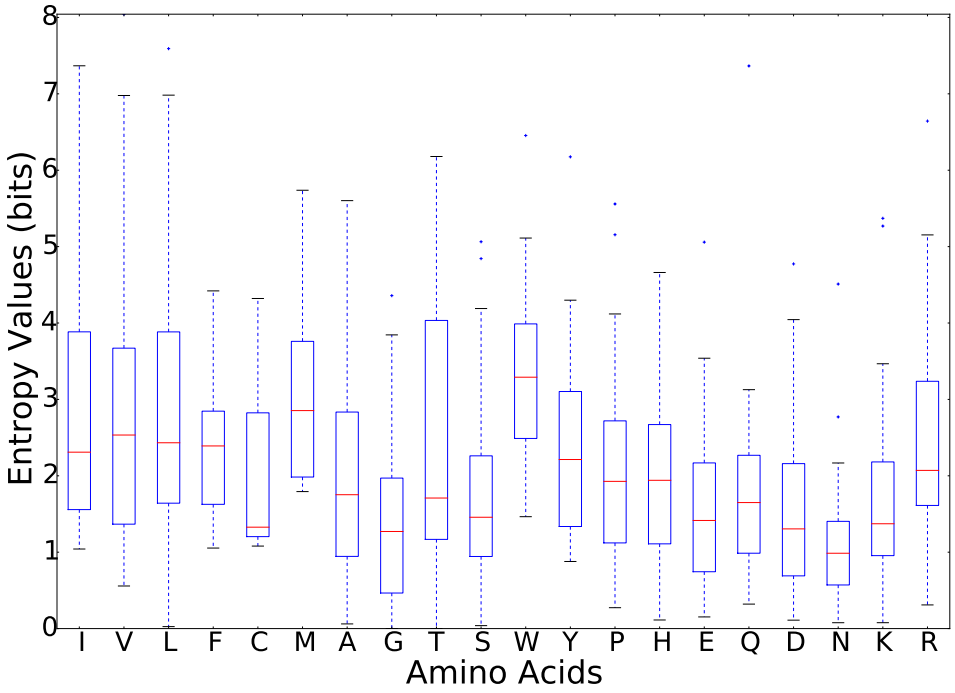

Figure 3.2: Box plot for the entropy values calculated for all the residues in all subgroups as a function of decreasing hydrophobicity. The residues I,V,L,F,C,M are mostly hydrophobic because of positive KD hydrophobicity values and have on an average higher median entropy values.

in the enolase superfamily, it was observed that the two contact entropy distributions were not the same (Mann-Whitney U; p = 0.001); hence, the two contact entropy distributions are not the same (see Appendix A.4 for other subgroups).

These results imply that there is more contact freedom in the side chains of hydrophilic residues which are generally located in the core [55]. This may be due to two reasons: firstly the structurally aligned sites are aligned to different residue types, secondly the residues in the interior of the protein generally have a higher degree of contacts in a contact network. Such residues due to their location in the interior of the protein structure tend to have more neighbours, which they may form contacts with.

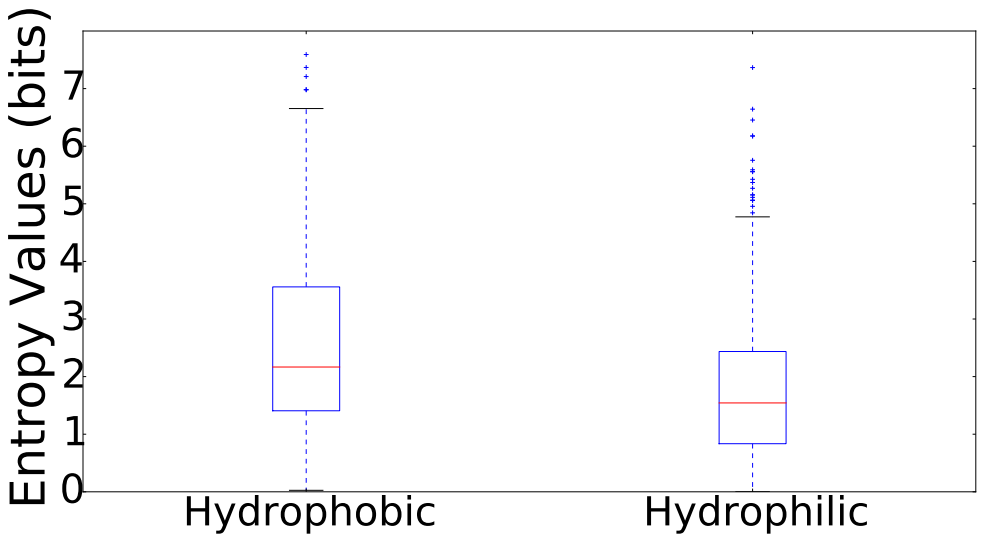

Figure 3.3: Box plot for the entropy values calculated for hydrophobic and hydrophilic residues for all subgroups in the enolase superfamily. Levene's test p = 0.002, Mann-Whitney U test p = 0.0001

3.3 Relationship Between Residue Contact Entropy and Residue Sequence Entropy

The high contact entropy values of some of the structurally aligned sites might be either due to the residue having more neighbours to interact with or the residue having high sequence entropy values (see Section 2.7). In this section, we analyze the relationship between contact entropy and sequence entropy values of structurally aligned sites. Sequence entropy of aligned sites or residues serves as an estimate for the number of different residues at that site. A high sequence entropy value of an aligned site indicates that the residue at the site is aligned to different residues, while low sequence entropy values indicates otherwise. Once the chains in the dataset were structurally aligned using MATT, the sequence entropy of each aligned site was calculated using equation 2.4. The two entropy values were plotted using a Python script. Figure 3.4 shows that there is a significant but weak relationship between the two entropy values (R^2 = 0.192, p = 0.007, slope = 0.260). Thus, the information obtained from the residue contact entropy values may be dependent on sequence entropy values. It also implies that a high value of contact entropy for some residues might be due to its respective sequence entropy. However, the high

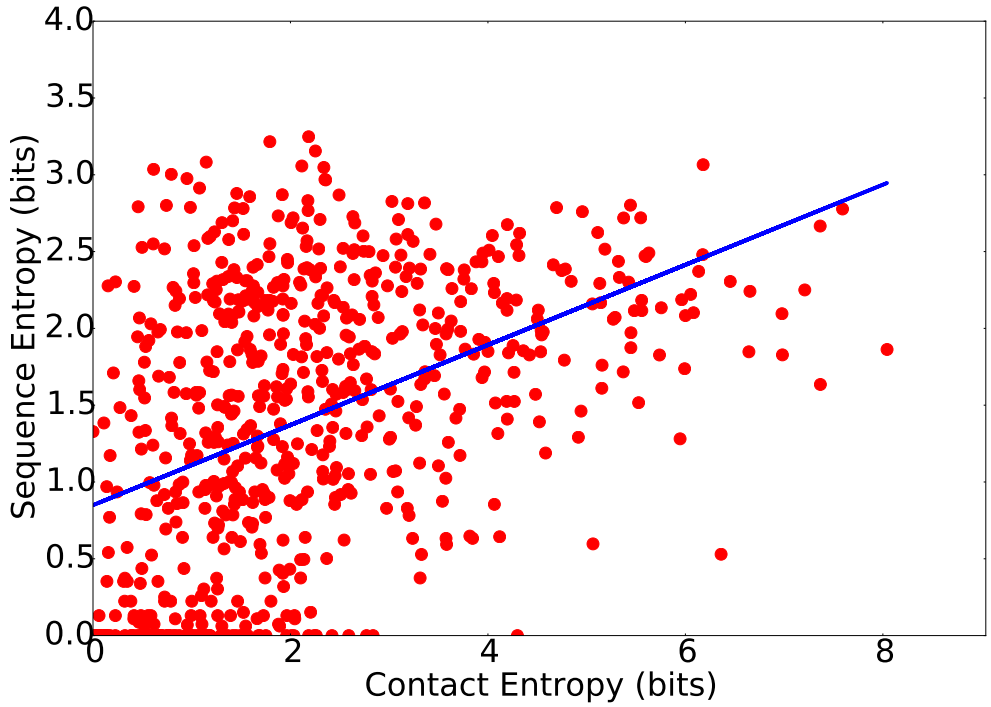

Figure 3.4: Relationship between the contact entropy and sequence entropy for all subgroups in the enolase superfamily. The plot shows a significant but weak co-relationship between the two entropies ($R^2 = 0.19$, p = 0.0006, slope = 0.260).

contact entropy value of residues may also be dependent on other factors such as the number of neighbours that the residue is surrounded by. We know that in the core of a protein, the residues are more tightly packed, which could also explain the high entropy values for such residues. The relationship between the contact entropy and the sequence entropy for the individual subgroups is shown in Appendix A.2.

3.4 Relationship Between the Contact Entropy and Properties of Residues

The mean contact entropy value of each amino acid was compared to the properties of the amino acids such as molecular weight and hydrophobicity [34]. The contact entropy values were divided on the basis of residue identity. A mean of all the contact entropy values was calculated for each residue and plotted against its hydrophobicity and molecular weight. As shown in Figure 3.5A, a significant correlation is observed

between the hydrophobicity and the mean contact entropy of residues; however, this correlation was weak ($p = 0.044$, $R^2 = 0.335$). The size of the amino acid residues also had significant but weak correlation with the mean contact entropy values ($p = 0.007$, $R^2 = 0.206$) (Figure 3.5B). The significant correlation implies that as the size and hydrophobicity of the amino acids increases, their mean contact entropy also increases. This relationship of contact entropy values derived from FCM are consistent with the relationship between contact matrix and residue properties assumed in various studies [13,39]. However, for the size of residues it is important to take the 3D location of residues into consideration, residues that have high molecular weight (size) such as tryptophan can have low contact entropy if they are located away from the core of protein structure. The low density of residues close to the surface of the protein structure gives such residues fewer number of neighbours to form contact with. Hence, the relationship between contact entropy and size of amino acids does not exist in void and is dependent on various other factors such as the location of residues. This may not be the case with hydrophobicity and contact entropy relationship because generally the hydrophobic residues are found in the core of protein structure [55]. All subgroups generate the same results and their plots are presented in Appendices A.5 and A.6.

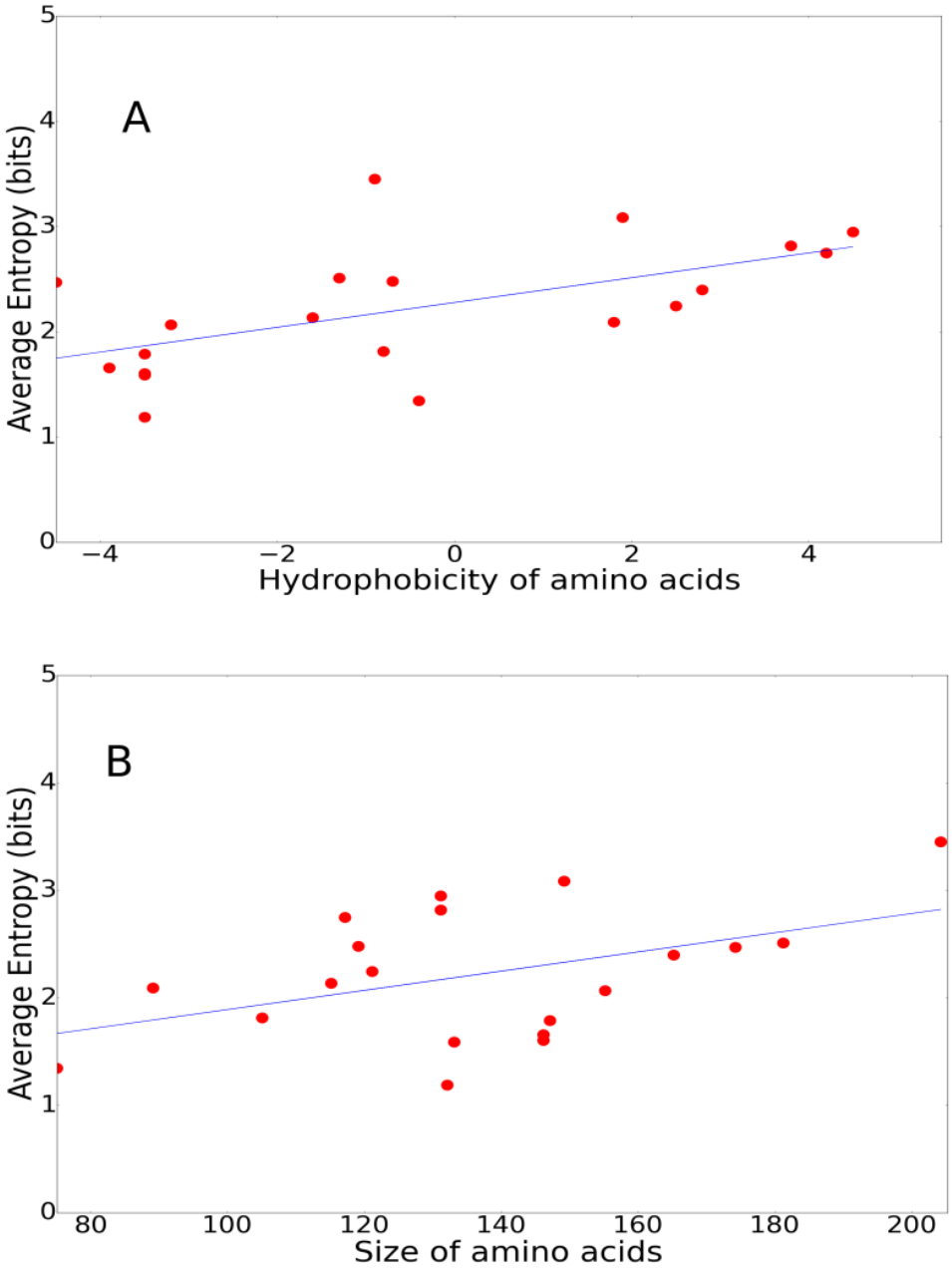

Figure 3.5: Relationship between mean residue contact entropy of amino acids and their respective **A:** hydrophobicity ($R^2 = 0.335$, $p = 0.044$), and **B:** molecular weight ($R^2 = 0.206$, p = 0.007) for structures of all three subgroups in the enolase superfamily. The correlation between size and entropy values and hydrophobicity and entropy values is significant but weak.

3.5 Difference Between Ligand-bound and Ligand-free Entropy Distributions

3.5.1 Residue Contact Entropy Along the Protein Sequence

Our study also aimed to examine the effects of ligand binding on protein structures. Ligands are molecules that bind at the active site of enzymes and are often substrates or inhibitors. Residues close to the active site are most affected by ligand binding through electrostatic effects and may alter the conformation of the proteins thereby affecting the semantics of the protein contact networks as a whole [21]. The relationships between the 3D location and the properties of amino acids and their contact entropy values were established in the previous section. In this section, we analyze the effects of ligand binding on contact entropy values. To achieve this, all three subgroups were divided into ligand-bound $\mathcal{S}^{\mathcal{L}}$ and ligand-unbound $\mathcal{S}^{\mathcal{U}}$ sets of structures as described in Section 2.1.3. For ligand-bound and ligand-free datasets, a reference structure was chosen as described in Section 2.2 for each of the three subgroups. The reference structures for $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ were structurally aligned using MATT to obtain structurally homologous sites. Figure 3.6 shows the running average (window size = 10) of the contact entropy values of all the aligned sites for the two reference structures in the enolase subgroup. The curves on the plot show that the contact entropy values along the sequence of the reference structures for $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ are not exactly the same. The peaks in the plot were investigated to find if they correspond to any secondary structural elements (α-helices and β-chains). The secondary structure identification scale was taken from RCSB and merged with the plot. In the plot, the upper scale belongs to ligand-free structure and lower scale belongs to ligand-bound structure. Some of the peaks correspond to α-helices (shown by red rods) while others point to β-sheets (shown by the yellow rods). The merging of secondary structure with the contact entropy plot might not be the optimal way of mapping the peaks to secondary structural elements. Hence, the peaks of the graph were mapped on ligand-bound structure using PyMOL and it was observed that for enolase subgroup these peak entropy values belong to $\alpha-$helices (as shown in Figure 3.8 in orange). We also observe that the high entropy residues are not in the immediate vicinity of ligands (Figure 3.8 shown in red). The contact entropy plots for ligand-bound and ligand-free

structures tend to show a similar pattern, however, for some of the residues, a clear difference is observed. To further investigate this we plotted the difference between the two curves. Figure 3.7 shows that the entropy values differ at every aligned site however the difference in some cases is not significant. Similar to the previous plot, the peaks and troughs (significant difference) of the plot were mapped to the protein structure and it was observed that for the enolase subgroup, the maximum difference is found in the loops. We also observe that some of the residues that have high absolute difference value are in close vicinity of ligands (Figure 3.9). Our results are parallel to previous studies in which the high centrality residues are in close vicinity of ligands and active site [19,20,33]. Residues close to the active site (residues marked by red) in the ligand-bound reference structure and the ligand-free reference structure were structurally homologous because reference structures for $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ were the same proteins belonging to same organism in its ligand-bound and ligand-free form. Reference structures that belong to other subgroups also show similar results (Appendix A.7).

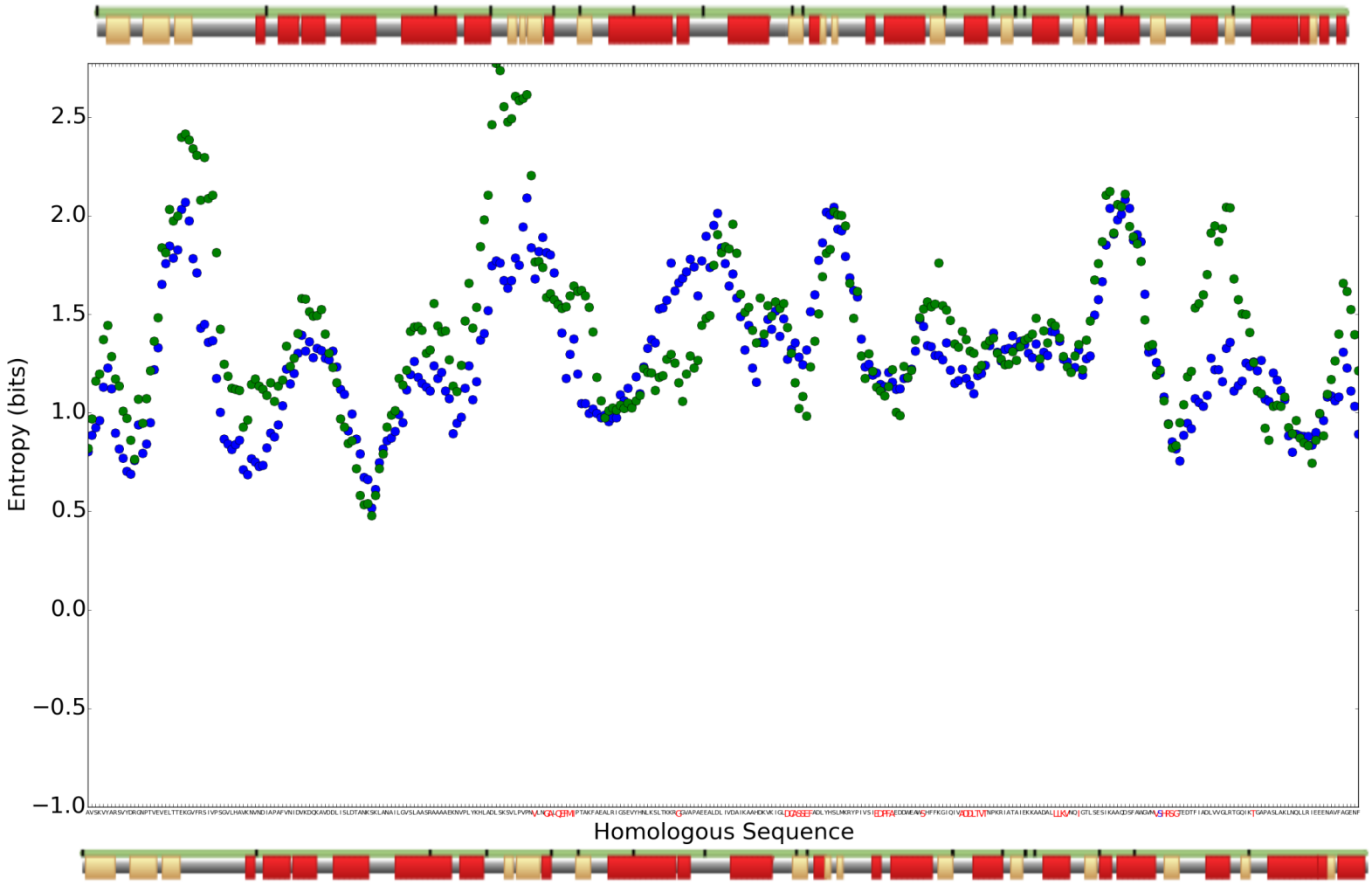

Figure 3.6: Running average (window size = 10) of contact entropy values along the sequence of the reference structure for $\mathcal{S}^{\mathcal{L}}$ (blue) and $\mathcal{S}^{\mathcal{U}}$ (green) (structurally aligned sites) in the enolase subgroup. The upper scale for secondary structure identification belongs to the reference structure in $\mathcal{S}^{\mathcal{U}}$ and lower scale belongs to the reference structure in $\mathcal{S}^{\mathcal{L}}$. The peaks in the plot correspond to different secondary structural elements such as alpha helices (red rods), beta sheets (yellow rods) and coil (grey thin rods). The residues close to the active site are shown by red coloured ticks.

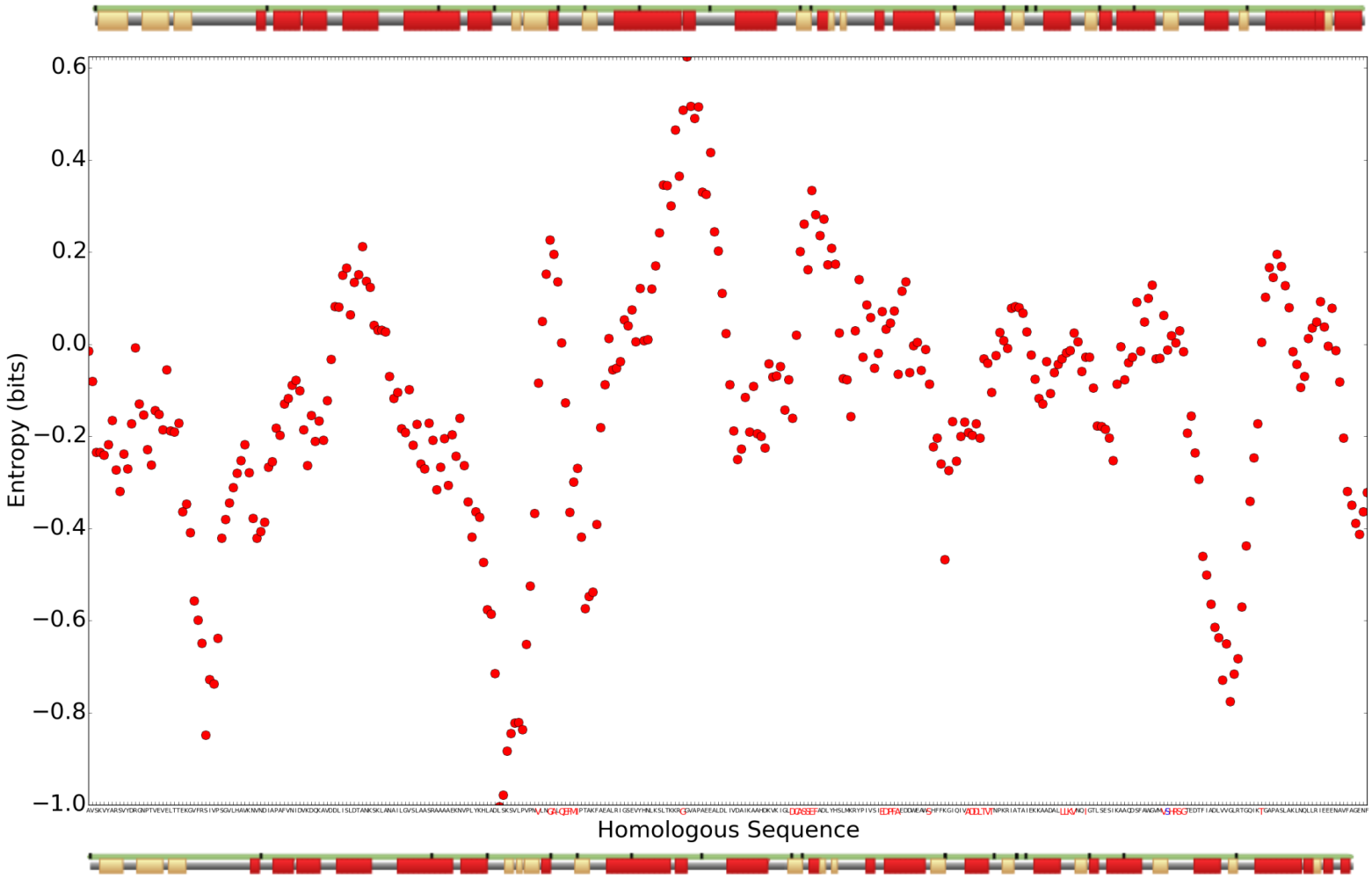

Figure 3.7: Difference between the running average of contact entropy values for structurally aligned sites in the reference structure for $\mathcal{S}^{\mathcal{L}}$ and the reference structure for $\mathcal{S}^{\mathcal{U}}$ in the enolase subgroup. The upper scale for secondary structure identification belongs to the reference structure in $\mathcal{S}^{\mathcal{U}}$ and lower scale belongs to the reference structure in $\mathcal{S}^{\mathcal{L}}$. The peaks and the troughs in the plot correspond to different secondary structural elements such as alpha helices (red rods), beta sheets (yellow rods) and coil (grey thin rods). The residues close to the active site are shown by red coloured ticks.

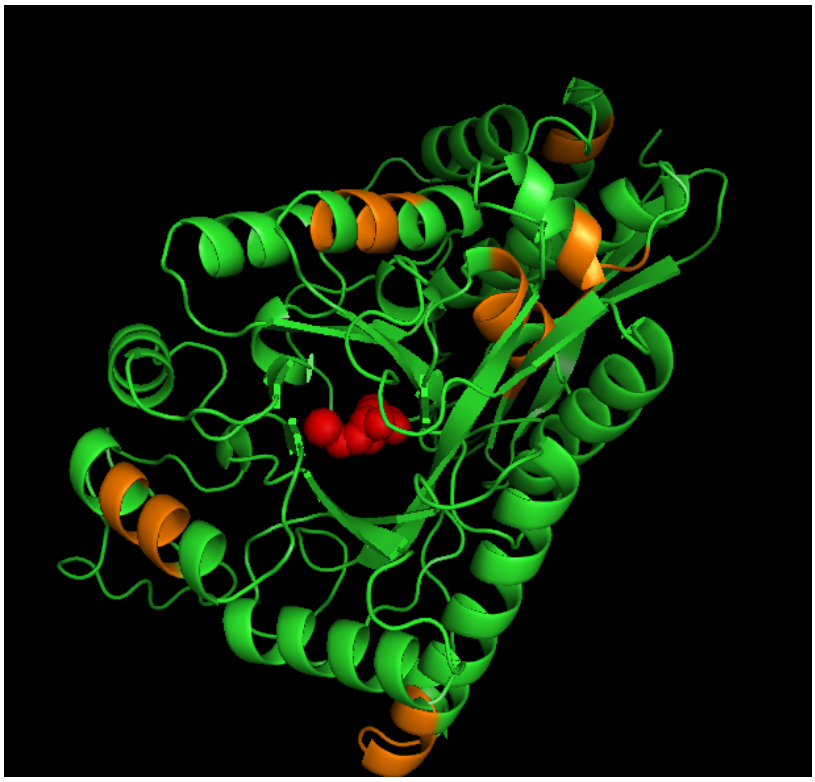

Figure 3.8: Mapping of the peaks (orange) of the contact entropy distribution for ligand-bound and ligand-free structures in enolase subgroup (Fig 3.6). The high entropy residues mostly correspond to α- helices on the surface away from the ligands (red spheres). The figure was generated using PyMOL (PDB ID: 1ELS Chain: A)

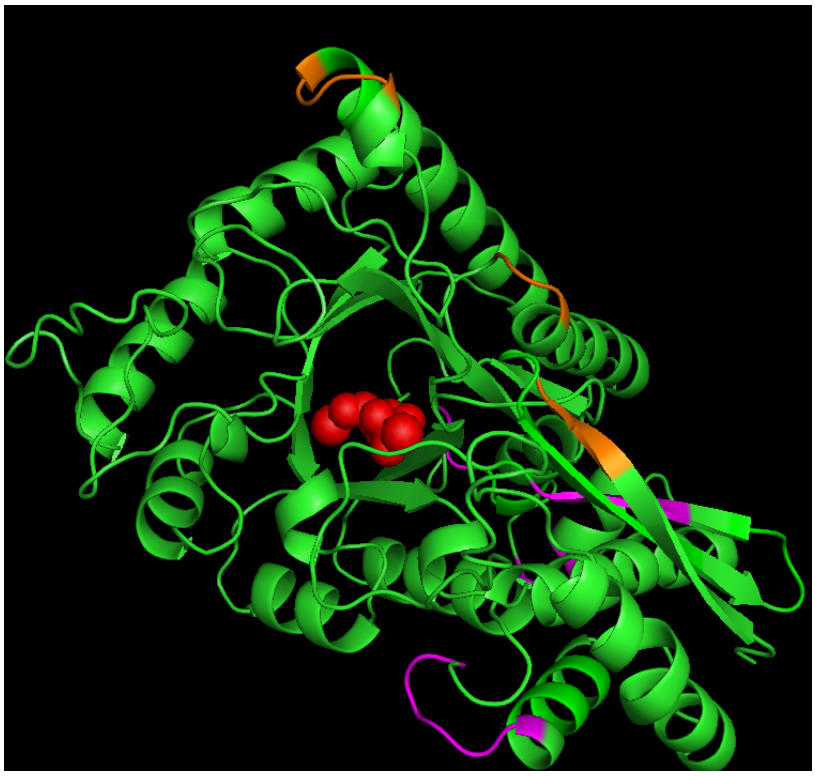

Figure 3.9: Mapping of the peaks (orange) and troughs (magenta) of difference between contact entropy values of ligand-bound and ligand-free structures in enolase subgroup (Figure 3.7). The high difference determines the highest effect of ligand binding on the structures and mostly the loops are found to have an effect of ligand binding. Some of the maximum difference residues are found in close proximity to the ligands (red spheres). The figure was generated using PyMOL (PDB ID: 1ELS Chain: A)

3.5.2 Difference Between Ligand-bound and Ligand-free Entropy Distributions Close to the the Active Site and Far from the Active Site

A difference in entropy distributions was observed between the ligand-bound and ligand-free reference structures. The results indicate that peaks in the corresponding plots do not map to a particular secondary structure or location. Further investigation was performed to understand the difference in entropy distributions of ligand-bound and ligand-free structures. For all datasets, the residues were divided into a set of residues close to the active site (AS; at $\mathcal{D}^{\mathcal{A}} \leq 10$ from Mg^{2+} or Mn^{2+}) and far from the active site ($\neg AS$) as described in Section 2.8. Let E^c be a set of contact entropy values. $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ are the contact entropy values close to the

active site for $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ structures, respectively. $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$ are the sets of entropy values of residues that are farther from the active site for the $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ structures, respectively. We tested the difference between the entropy distributions using the K-S test. The null hypothesis of the K-S test states that the two compared distributions are identical to each other (derived from same reference distribution). For each subgroup, four residue contact entropy distributions for $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$, $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$, $E^c(\mathcal{S}^{\mathcal{U}}, AS)$, and $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ were compared in the following pairs:

1. $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, AS)$
2. $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$
3. $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$
4. $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$

Empirical cumulative distribution function (ECDF) plots were used to visualize the differences in the distributions. As shown in Table 3.1, most of the p-values obtained from the K-S test for different distributions for all subgroups are greater than the a significance cut-off threshold of $\alpha = 0.0041$, which indicates that the null hypothesis cannot be rejected. The cut off threshold was taken to be $\alpha = 0.0041$ instead of $\alpha = 0.05$ for multiple test correction (Bonferroni correction α/n where n is the number of tests [54]). The only p-value obtained from the K-S test that was smaller than the threshold was for $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$ in the muconate cycloisomerase subgroup (Figure 3.10). This suggested that for the above stated distribution the null hypothesis is rejected and the distributions are different from each other. However, this result does not provide enough evidence to universally reject the null hypothesis since the p-values for other distributions are higher than the threshold. This implies that for most of the datasets, the contact entropy values of $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ structures are similar to each other or, in other words, the side chain contact freedom of residues close to the active site (or otherwise) does not change subsequent to ligand binding. The ECDF plots for other subgroups are shown in Appendix A.8. Biologically, ligand binding changes protein structures. Previous studies have shown that ligand binding affects the protein conformation [21]. The study conducted by Erman (2015) states

p-values	$E^c(\mathcal{S}^{\mathcal{L}}, AS)$ $E^c(\mathcal{S}^{\mathcal{U}}, AS)$	$E^c(\mathcal{S}^{\mathcal{L}}, AS)$ $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$	$E^c(\mathcal{S}^{\mathcal{U}}, AS)$ $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$	$E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$
Enolase	0.3457	0.8727	0.6811	0.3949
Mandelate racemase	0.9725	0.5525	0.8965	0.335
Muconate cycloisomerase	0.5907	0.0127	0.732	0.0009*

Table 3.1: p-values obtained from the K-S tests on the indicated pairs of residue contact entropy distributions. Significant differences are indicated by an asterisk (*)

that, when a ligand binds on activated CDC42 kinase 1, some parts of the protein contact network derived from contact matrix become stiffer than the others [21]. The closeness centrality of the contact network is also shown to be different prior to ligand binding [19]. As stated earlier, protein contact networks are an implementation of contact matrices. The changes in the contact networks are a reflection of changes in the contact matrices, which determines the effect of ligand binding on the contact matrices of protein structures. However, our results show otherwise, and there can be many reasons for our results being contradictory. One reason may be that the contact entropy within the enolase superfamily actually does not change due to ligand binding. It may be possible that the contact entropy calculated using FCM of the homologous proteins of other superfamilies shows a difference between ligand-bound and ligand-free structures. Since none of the above mentioned studies have used the enolase superfamily as the dataset, it will be informative in the future to analyze the contact entropy for other superfamilies. However, it is also possible that even the enolase as a superfamily may show difference in contact entropy values after ligand binding if the distributions compared are generated using different criteria. In this study, residues close to the active site are defined as those residues within a sphere of radius ≤ 10 from Mg^{2+} or Mn^{2+}. This may not be the optimal way to determine residues close to the active site. There may be an alternative method to determine the active site and an optimal distance from the active site. Homologous sets of protein structures might not be the best way to aggregate a FCM; for future studies, we can use a set of protein structures such as ones derived from molecular dynamic simulation to calculate a FCM. It is also possible that the concept of contact entropy might not be a satisfactory approach to determine the effect of ligand binding.

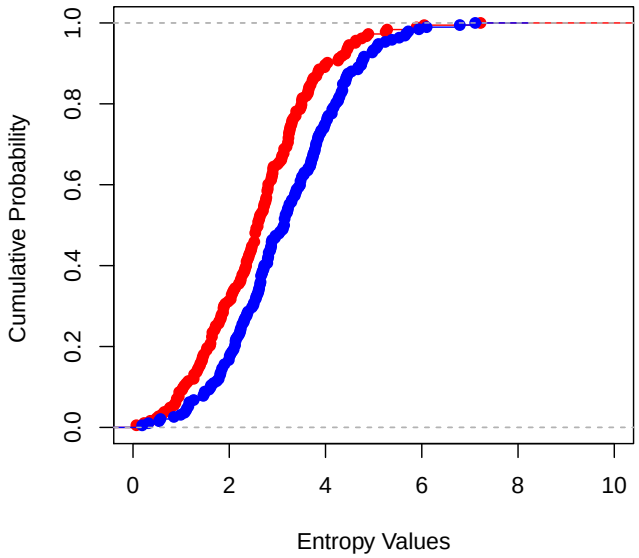

Figure 3.10: ECDF of the contact entropy values of residues for $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ (red) and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$ (blue) (p-value = 0.0009) for the muconate cycloisomerase subgroup.

3.6 Relationship Between p-values and $\mathcal{D}^{\mathcal{A}}$

The residues close to the the active site often have high centrality values in protein contact networks [19]. In this research, the distance from the active site $\mathcal{D}^{\mathcal{A}}$ is taken to be ≤ 10 to define the residues "close" to active site, as described in Section 2.8. In the previous section, it was shown that the distribution of contact entropies for residues close to the active site for $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ structures is the same. The p-values were observed to be higher than the determined threshold (> 0.004); hence, the null hypothesis that these distributions are the same could not be rejected. The choice of 10 as the distance from the active site was made by visual mapping of residues close to the active site (Section 2.8). It was important from the perspective of this study to examine the relationship between the results obtained and $\mathcal{D}^{\mathcal{A}}$. In this section, we investigate the relationship between p-values obtained from the K-S test and the active site distance $\mathcal{D}^{\mathcal{A}}$. This analysis is essential to determine if the contact entropy distribution depends on the value of $\mathcal{D}^{\mathcal{A}}$. Once the contact entropy of residues was

calculated, sets of $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ were collected for multiple values of $\mathcal{D}^{\mathcal{A}}$ (8 - 15). Any distance less than 8 would make the $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ too small to make statistically meaningful interpretation. This would lead to uneven partitioning of residues into "close to active site" and "farther from active site" making the results statistically weak. P-values were obtained by comparing residues at a distance of $\leq \mathcal{D}^{\mathcal{A}}$ from the active site for ligand-bound and ligand-free reference structures using the K-S test. These p-values were plotted as a function of increasing $\mathcal{D}^{\mathcal{A}}$. The results reveal that the curve of p-value versus $\mathcal{D}^{\mathcal{A}}$ does not show any obvious pattern or trend for the enolase and mandelate racemase subgroups (see Figure 3.11 A and B). However, for the muconate cycloisomerase subgroup, a significant relationship was observed where the p-values increase as $\mathcal{D}^{\mathcal{A}}$ increases (Figure 3.11 C) (R^2 = 0.825 p = 0.00004). For the muconate cycloisomerase subgroup, it was also observed that for $\mathcal{D}^{\mathcal{A}} \leq 8.2$ the p-values obtained from K-S test are lower than the threshold (0.004), implying a clear difference in the distributions. This implied that the p-values are sensitive to the distance from the active site for the muconate cycloisomerase subgroup. It may also be possible that for other subgroups, the p-values obtained from the K-S test of contact entropy values close to the active site ($E^c(\mathcal{S}, AS)$) or farther from the active site ($E^c(\mathcal{S}, \neg AS)$) are also sensitive to the distance from the active site when the $\mathcal{D}^{\mathcal{A}}$ is taken to be ≤ 8 . However, as mentioned earlier in this section, any distance below 8 would lead to less number of residues in the "close to active site" dataset. In our study, we observe that optimal value of $\mathcal{D}^{\mathcal{A}}$ differs from one subgroup to another in the same superfamily, which makes it possible that value of $\mathcal{D}^{\mathcal{A}}$ which is ideal for one superfamily may not be ideal for the other superfamilies. This reiterates our previous assumption that the results obtained in this study might be negative due to the lack of ideal parameter values and definitions.

3.7 Data Simulation and Sensitivity of the K-S Test

Simulation may be used to assess the validity of a method, tool, or algorithm in a controlled fashion. To ensure that the K-S test was sufficiently sensitive to detect the differences between the datasets, we conducted tests using simulated datasets where the difference in a contact entropy distribution was known to exist. This experiment

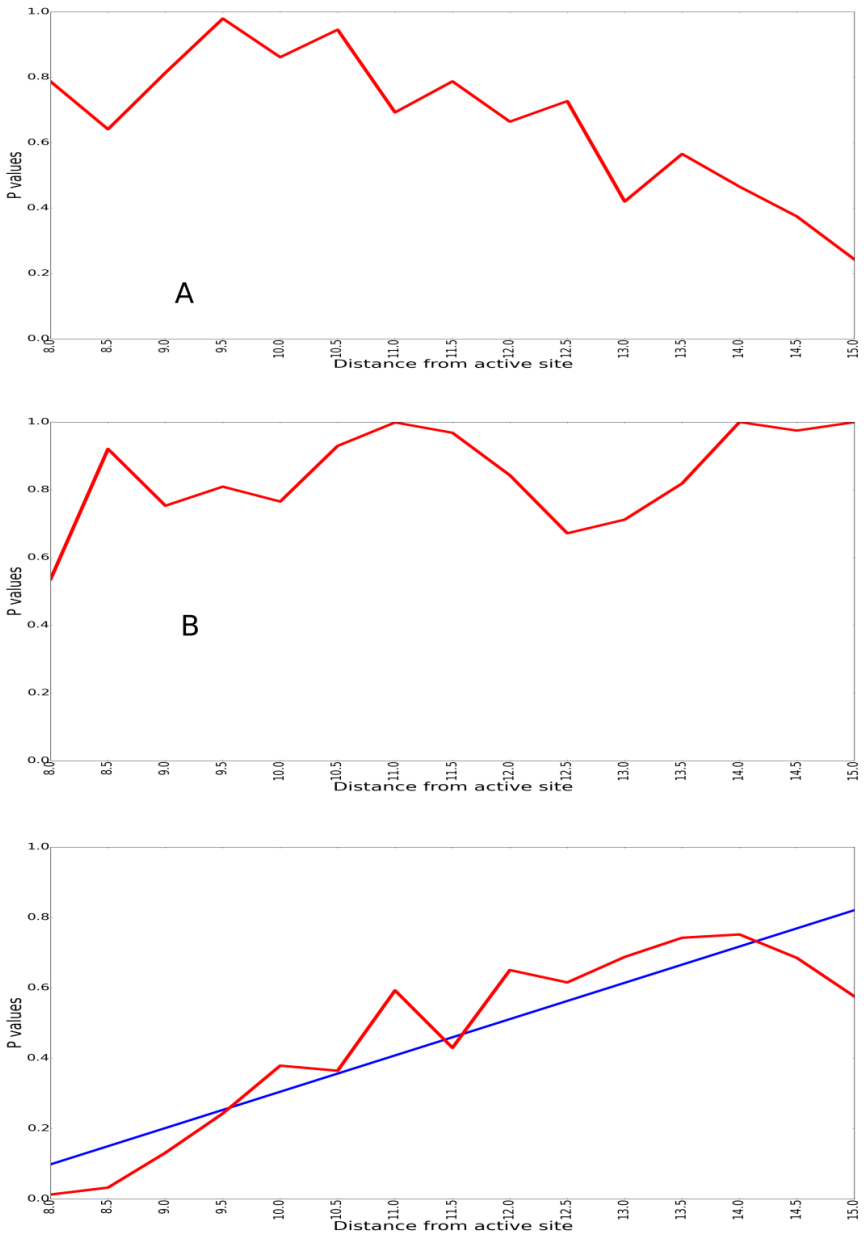

Figure 3.11: p-values after comparison of the contact entropy distribution of $\mathcal{S}^{\mathcal{L}}$ and $\mathcal{S}^{\mathcal{U}}$ structures in **A:** enolase **B:** mandelate racemase, and **C:** muconate cycloisomerase subgroups at multiple $\mathcal{D}^{\mathcal{A}}$ values (8 - 15). The plot shows no pattern for the enolase and mandelate racemase subgroup; however, in the muconate cycloisomerase subgroup a trend is evident (linear regression line) $R^2 = 0.825$ p $= 0.00004$.

was also performed to determine minimum detectable effect size, i.e., how different should the two distributions be for the K-S test to detect a difference. As described in Section 2.10, values in aggregate matrices $C^{a'}$ were simulated using equation 2.5. The simulation was approximated to actual data as much as possible. For each σ, 100 instances of randomized $C^{a'}$ were generated (see Section 2.10), which was later converted to $C^{f'}$. $E^{c'}$ was calculated from the simulated $C^{f'}$. The K-S test was performed on actual contact entropy distributions of residues close to the active site $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and 100 instances of the simulated contact entropy distribution of residues close to the active site $E^{c'}(\mathcal{S}^{\mathcal{L}}, AS)$. For each value of σ, we plotted the total number of p-values obtained from the K-S test that were < 0.05. As shown in Figure 3.12, for all simulated datasets the count of p-values < 0.05 increased as the value of the standard deviation σ increased. In other words, the number of simulated entropy distributions different from real distribution increases with increasing σ. The K-S test did not find the difference between the distributions when $\sigma < 0.5$. This may be due to the fact that the number of values in $C^{a'}$ different from C^a is quite low. Furthermore, we observed that at $\sigma < 0.5$, most of the values retrieved from the Gaussian distribution were less than 1, which were rounded to 0 (due to use of the floor function). As shown in equation 2.5, the simulated values of $C^{a'}$ are generated by adding the original value of C^a to the value retrieved from the normal distribution $\lfloor gauss(0, \sigma) \rfloor$; hence, the simulated data value = the original data value + 0. At $\sigma \geq 0.5$, 5% of the data values in $C^{a'}$ are different by at least 1 from the data values in C^a. The change in the remaining 95% of the data values is less than 1. This implies that if the two distributions have only 5% of their data values different from each other (0.5 σ), then the K-S test will detect the difference and the p-value obtained will be low. This suggests that the K-S test is sensitive to a 5% change in distributions and the minimum detectable effect size is 5%. Such low minimum detectable effect size indicated that the test was very sensitive and the number of false negatives was low.

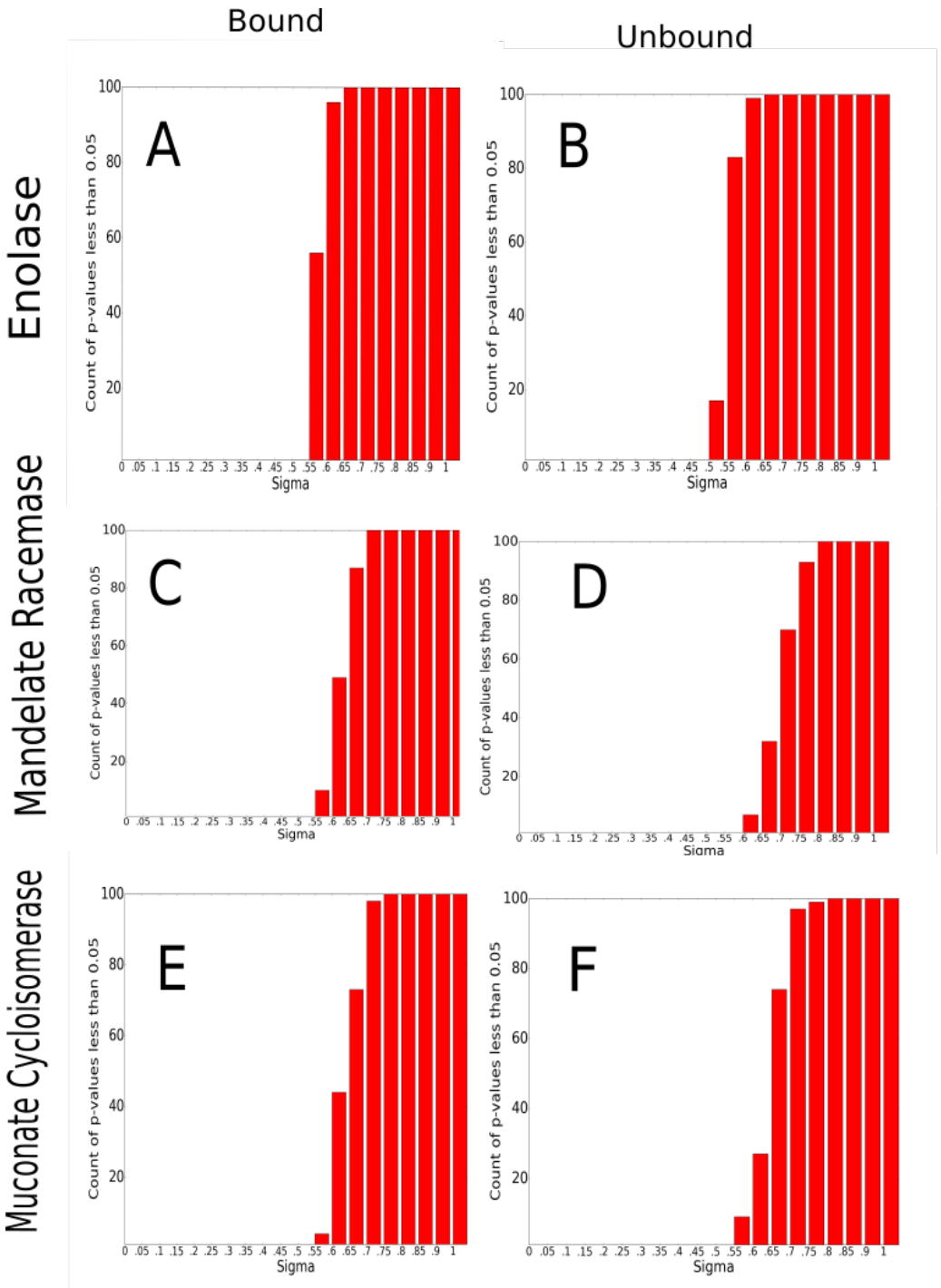

Figure 3.12: Count of p-values < 0.05 when a real distribution of frequency contact matrix at a $\mathcal{D}^{\mathcal{A}} \leq 10$ is compared to simulated distribution at sigma ranging from 0 to 1. For all datasets, the count of p-values < 0.05 obtained from the K-S test increases drastically at a σ of 0.5, which implies that the null hypothesis of the K-S test is rejected if 5% of the data values of the two distributions are different from each other.

Chapter 4

Summary and Conclusions

Primary structures of proteins are long polypeptide chains of individual amino acids called residues. This sequence of amino acids folds into secondary and tertiary structures. Some residues that are not consecutive in the primary sequence of a protein may interact with each other in the tertiary structure. The "interaction" between residues, often known as contacts, may be determined by the physical distance between the residues. Two residues are said to be in contact if the distance between them is less than a particular threshold value. There are multiple ways to calculate the contacts; one of the most sensitive ways is to calculate the distance between all the atoms of each residue. These contacts can be used to construct protein contact matrices, which may be used to understand various processes of protein folding pathways [29] and protein-protein interactions [16]. Contact matrices act as adjacency matrices for the formation of contact networks. Contact networks are small world networks and may be used to determine some important structural properties of proteins such as functionally important residues [19, 20, 33], residues that interact with ligands [7], and the location of the active site if the protein is an enzyme [5].

Homologous proteins are the set of proteins that reflect common ancestry and may have statistically significant similarity. The structural relationships inherent to a homologous set of proteins can be used to determine important information such as structural and sequential conservation of proteins. As stated above, the contact matrices of *individual* protein structures may be used for numerous applications, however examination of contacts with a homologous *set* of structures may provide more insight into the function and structure of these proteins. This study explores the use of frequency contact matrices (FCM) which are the normalized form of an aggregate of all contact matrices with the set of homologous protein structures. Structurally aligned sites or residues are determined by structural alignment of all the structures in the dataset. Low residue contact entropy values suggest that the structurally aligned

sites have conserved contacts throughout the homologous dataset, while high contact entropy values suggest that the contacts of structurally aligned sites are varied across the dataset.

4.1 Application to the Enolase Superfamily

This study aimed to extend the concept of contact matrices to frequency contact matrices and contact entropy. While contact matrices provide information about the structural semantics of an individual protein structure, we surmised that FCMs and residue contact entropy values, which encompass all the contact information for a set of homologous structures, would provide more insight into the protein structure for a related family of proteins. In this study, we implemented FCMs to investigate the relationship between residue contact entropy and the different properties of amino acids, as well as the effect of ligand binding on the conservation of side chain contact freedom across a set of homologous protein structures.

The proposed method was applied to the enolase superfamily. To identify the structural sets within the enolase superfamily, the structure-function linkage database (SFLD) was used [3] (see Section 2.1.1). The enolase superfamily is comprised of multiple subgroups, and members that belong to each subgroup are evolutionarily related and have more shared features than the superfamily as a whole. All the structures in each subgroup were structurally aligned using pairwise alignment (MATT) (see Section 2.3). The pairwise alignment generates structurally aligned sites. Each subgroup was divided into ligand-bound and ligand-free sets (Section 2.1.3). For each dataset, a reference structure was chosen in both the ligand-bound and the ligand-free form (Section 2.2). The residue contact entropy for each structurally aligned site was calculated (see Section 2.6). The residues were divided into a set of those residues that are close to the active site and that are farther from the active site on the basis of a threshold distance of ≤ 10 from the active site's Mg^{2+} or Mn^{2+} ion (Section 2.1.4).

The visual mapping of contact entropy values on residues in the protein structure showed that some of the residues buried inside the protein structure, which are generally more hydrophobic, had high contact entropy values (Figure 3.1). This figure

however was not enough to clearly determine the location of high entropy residues. This figure also established that for structurally aligned sites, an uncertainty of contact exists, i.e. FCM values across the dataset are not always equal to 1. After the division of residues into hydrophobic and hydrophilic on the basis of KD hydrophobicity, it was found that the median entropy values of hydrophobic residues was higher when compared to hydrophilic residues. Using Levene's test to analyze the variance showed that for aggregate data values of all subgroups, there was a difference between variances of hydrophobic and hydrophilic residues with p-values lower than the threshold (0.05). When contact entropy values for hydrophobic and hydrophilic residues were compared using the Mann-Whitney U test, the p-values obtained were low for all the subgroups. Our study also showed a significant relationship between residue contact entropy values and properties of the amino acids such as molecular weight ($R^2 = 0.206$) and hydrophobicity ($R^2 = 0.335$) (see Figure 3.5). Our study showed that when the hydrophobicity of residues increases, the contact entropy value also increases. It was also discussed that when size of the residues increases, it might not increase the entropy of residues. The location of residues in the protein structure plays an important role in determining the residue's entropy. The relationship between the FCM values of residues and properties was found to be in concordance with the relationship observed in the study conducted by Chamarro et. al (2011), where the values of contact matrix for protein structures were predicted using different properties of amino acids. The reasons for high contact entropy values of some residues were examined. It was observed that the sequence entropy value of high contact entropy residues was high. On analysis of residue contact entropy and residue sequence entropy, a significant relationship was found between the two (Figure 3.4). Hydrophilic residues which are generally present in the core of the protein were found to have higher contact entropy values as compared to hydrophilic residues. The high contact entropy of such residues may be also due to a higher density of residue packing at the protein core, which puts them in the vicinity of more neighbours for interaction. The contact entropy of ligand-bound and ligand-free structures were observed to be different with some peak entropy values. The peak entropy values were mapped on the actual structure and for some of the subgroups showed that the high contact

entropy values are located on $\alpha-$helices. For other subgroups, like mandelate racemase, the high contact entropy values were mostly found to be on loops whereas for muconate cycloisomerase no particular pattern was observed. The difference between the entropy values was also plotted. The highest absolute value of difference of a particular residue shows that the ligand binding has maximum effect on that residue. We observed that high difference values were in close vicinity of ligands when peaks of the entropy difference plot were mapped to the actual structure. The results agree with to previous studies where highly central residues were found interacting with ligands or as close neighbours to the active site [19, 20, 33]. To further analyze the effect of ligand binding, the difference in entropy values of residues close to the active site for ligand-bound and ligand-free structures was analyzed using the Kolmogorov-Smirnov (K-S) test (Section 2.9). The p-values obtained suggested that for two of the three subgroups the distribution of contact entropy values for ligand-bound and ligand-free structures was not significantly different; hence, the freedom of side chains in structurally aligned sites does not change across these datasets subsequent to ligand binding (Table 3.1). However, for the muconate cycloisomerase subgroup, there was a significant difference between the $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$ distributions of residue contact entropies. This, however, was not sufficient to universally support the hypothesis that subsequent to ligand binding the contact entropy of residues close to active site or otherwise changes. The literature in this field of study indicated otherwise. The study conducted by Erman et. al (2015) showed that ligand binding affects the protein's stiffness, i.e., some contacts in the contact matrix become more stable while as others fluctuate. Other studies also show that the centrality of protein contact networks is affected after ligand binding [19]. The reasons for the contrast between our results and previous studies were discussed. It was discussed that the dataset used in previous studies does not comprise of enolase superfamily which implied that the effect of ligand-binding might be different for different superfamilies. However, the FCM values of enolase superfamily members may also be affected by ligand binding if the parameters of the experiment like active site distance, active site definition, even different method of calculating FCM are optimized. There is also a possibility that the above stated parameters are optimal, but contact entropy

is not the right measure to differentiate between ligand-bound and ligand-free structures. The dependence of contact entropies on the distance from active site was also examined. It was shown that for the enolase and mandelate racemase subgroups, the results obtained were independent of the active site distance ($\mathcal{D}^{\mathcal{A}}$); however, statistical tests comparing distributions in the muconate cycloisomerase subgroup were sensitive to the choice of $\mathcal{D}^{\mathcal{A}}$. An optimal active site distance of ≤ 8.2 was observed. For the other subgroups, there may be a possibility of ideal active site distance to be < 8 . The variation of optimal active site distance in the same subgroup also signifies that there may be variation in different superfamilies. Simulation were conducted to check the sensitivity of the K-S test as a non-parametric test and to determine the minimum detectable effect size. Comparison between the real distribution and the simulated distributions suggested that the K-S test for two distributions with 5% difference between their data points, generates low p-values (Figure 3.12). The sensitivity of the K-S test was observed to be very high which decreases the chances of false positives.

4.2 Extensions and Future Work

At the beginning of this study two hypotheses were framed. One was that the contact entropy values are related to different properties of amino acid residues, and second was that the contact entropy can be used to determine the effects of ligand binding on the protein structure. The results have not indicated a strong support for either of the framed hypotheses. A significant but weak relationship was observed between contact entropy values of residues and its properties and no difference was observed between ligand-bound and ligand-free structures. The future work that succeeds this study should focus on identifying the reasons why there appears to be little difference in protein structure subsequent to ligand binding in three subgroups of the enolase superfamily studied in the present work. This can be done by addressing three questions:

1. Can a different method be used?

2. Is it possible that ligand binding does not affect the structure of a protein? Is this trait dependent on the superfamily?

3. Are homologous sets of structures the best way to assemble the dataset?

The methodological approach can be addressed in multiple ways. First, is to find the optimum number of structures required to test the hypothesis. Datasets with larger quantities of structures might provide a better estimate of entropy values, which could be more informative. Second, as stated earlier the active site may be redefined. In this study, the active site was determined by a sphere around the metal ion (magnesium and manganese). It would be informative to investigate a better way to define the active site. An optimal distance from the active site also needs to be assessed which may be different for different superfamilies. Also, the contact between two residues was determined by the least distance between any of their atoms. Other ways to calculate the contacts, such as centroid-based calculation could be implemented to examine if it is a more appropriate approach than the one used in this study. Third, the ligand-bound datasets used in this study were not homogeneous in nature, i.e., the ligands bound to structures in the same dataset are different and the protein structures belong to different organisms. The question that needs to be addressed is: Should the two structures that belong to the same subgroup, but having different ligands bound to them, be a part of the same dataset? Fourth, is there a better database other than the SFLD to organize our data and categorize the structures into different subgroups or families?

It might be possible that the basis of the hypothesis of this study is not valid for some superfamilies and ligand binding does not actually affect the protein structure in the enzyme of enolase superfamily. It is also possible ligand binding does affect the structure of proteins in enolase superfamily but contact entropy calculated over FCM (from homologous protein structures) may not be the right measure to determine the difference. The question then arises, what are the results of other superfamilies? In the future we would like to extend this method to other superfamilies and observe the effect of ligand binding on residues close to the active site.

This study leverages homologous sest of structures to calculate FCM because they give information about the structures, functions, and sequences of member proteins. Our work can be extended to sets of structures assembled in a different way, such as structures collected using molecular dynamic (MD) simulations. MD simulations can give a dynamic evolution of a protein folding into its native state. Two protein

structures can be selected in their ligand-bound and ligand-free forms. The datasets in the end will be comprised of "transient" protein structures, i.e., before the respective proteins fold into their native state. The time interval can be a microsecond or even a sub-microsecond between the structures. Such datasets could be used to calculate contact matrices, FCM, and contact entropies. The resulting contact entropy distributions could then be compared using the K-S test.

Appendices

Appendix A

A.1 Protein Mapping

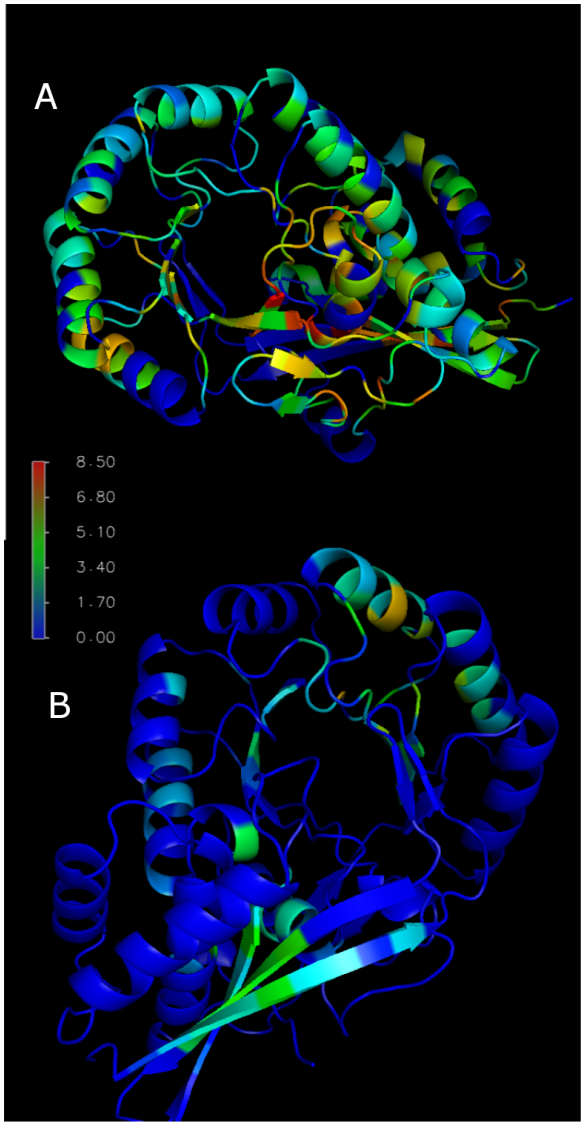

Figure A.1: Residue contact entropy mapping **A:** mandelate racemase (PDB ID: 3UXK chain B) **B:** muconate cycloisomerase (PDB ID: 3K1G chain A). The location of high entropy values is spread across the protein structures (red and yellow hotspots).

A.2 Contact Entropy and Sequence Entropy

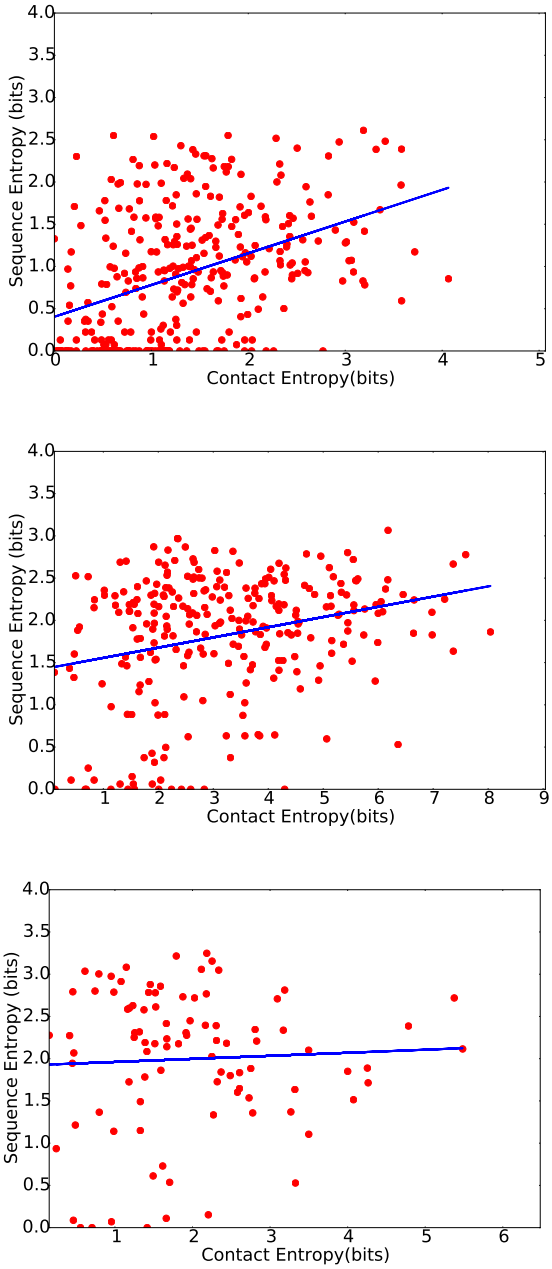

Figure A.2: Relationship between contact entropy and sequence entropy for enolase (top) ($R^2 = 0.16$, p = 0.0007, slope = 0.37), mandelate racemase (middle) ($R^2 = 0.07$, p = 0.0009, slope = 0.12), and muconate cycloisomerase (bottom) ($R^2 = 0.002$, p = 0.64, slope = 0.036) subgroups. The plots show that the two distributions are dependent on each other.

A.3 Contact Entropy Values of Individual Amino Acid Residues

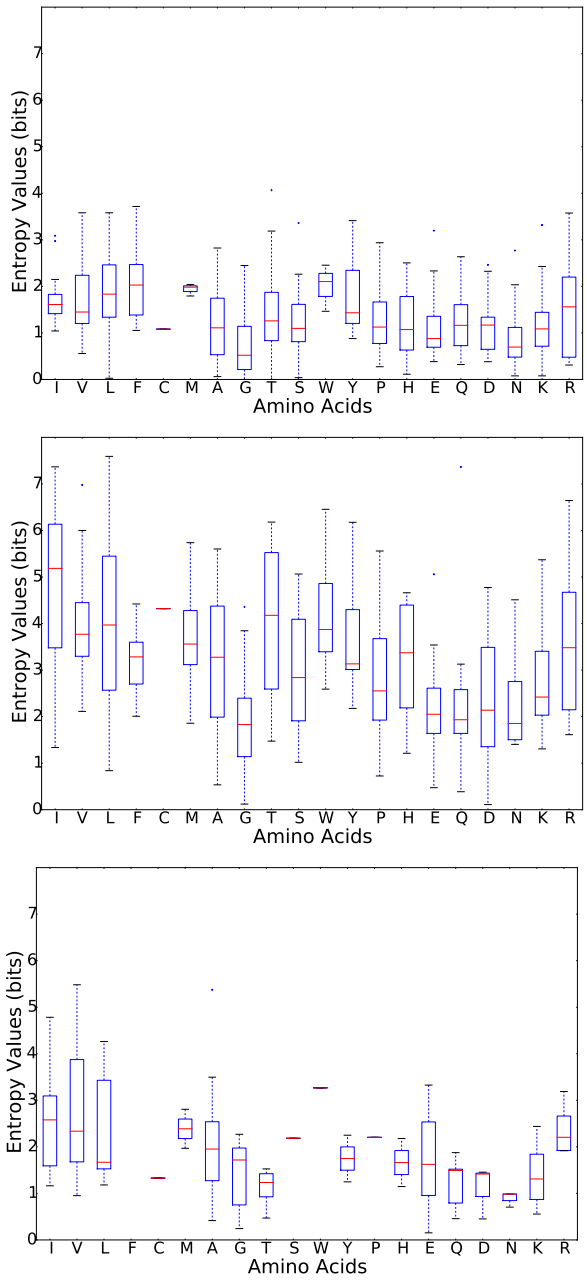

Figure A.3: Box plots for entropy values for all the residues in the enolase (top), mandelate racemase (middle), and muconate cycloisomerase (bottom) subgroups as a function of decreasing hydrophobicity. The median entropy values (marked by red lines) of hydrophobic residues (I,V,L,C,M) is higher than those of other amino acid residues.

A.4 Contact Entropy Values of Hydrophobic and Hydrophilic Residues

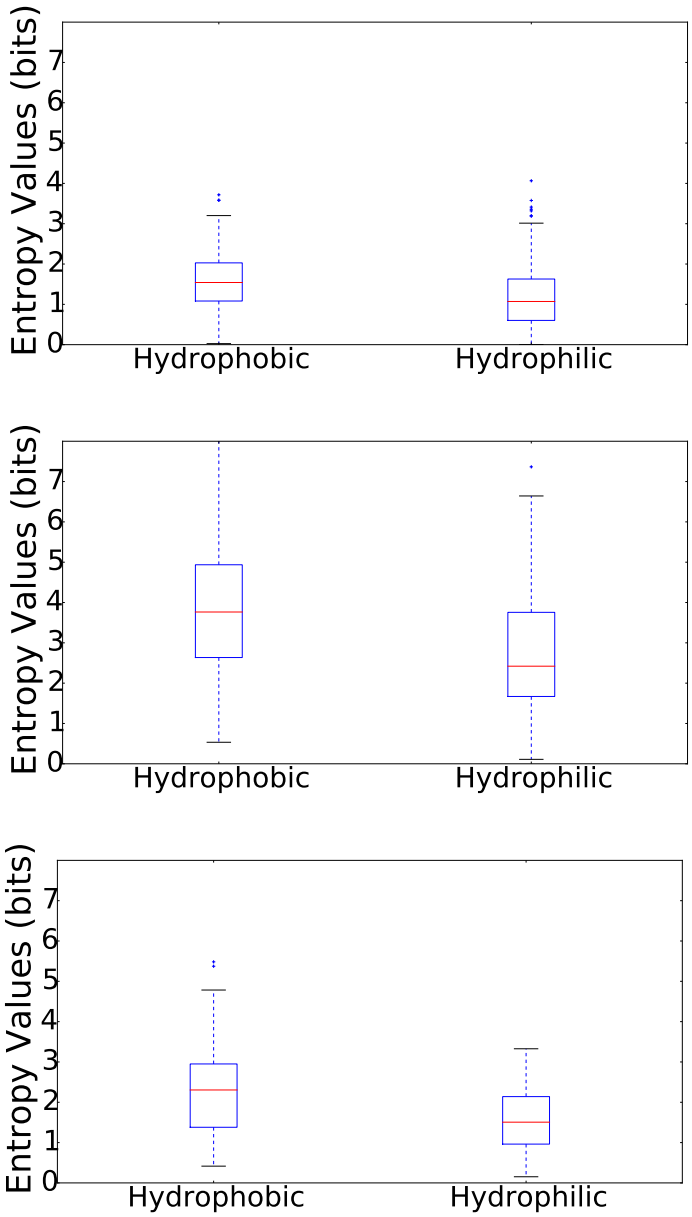

Figure A.4: Box plots for entropy values for hydrophobic and hydrophlic residues in the enolase (top) Levene's test p = 0.79, Mann-Whitney U test p = 0.0005, mandelate racemase (middle) Levene's test p = 0.46, Mann-Whitney U test p = 0.0001, and muconate cycloisomerase (bottom) Levene's test p = 0.01, Mann-Whitney U test p = 0.0005 subgroups. The median entropy values of the hydrophobic residues is higher than those for hydrophilic residues. The spread of the hydrophobic residue entropy distribution is larger showing higher variance than the spread of hydrophilic residues.

A.5 Relationship Between Residue Contact Entropy and Amino Acid Size

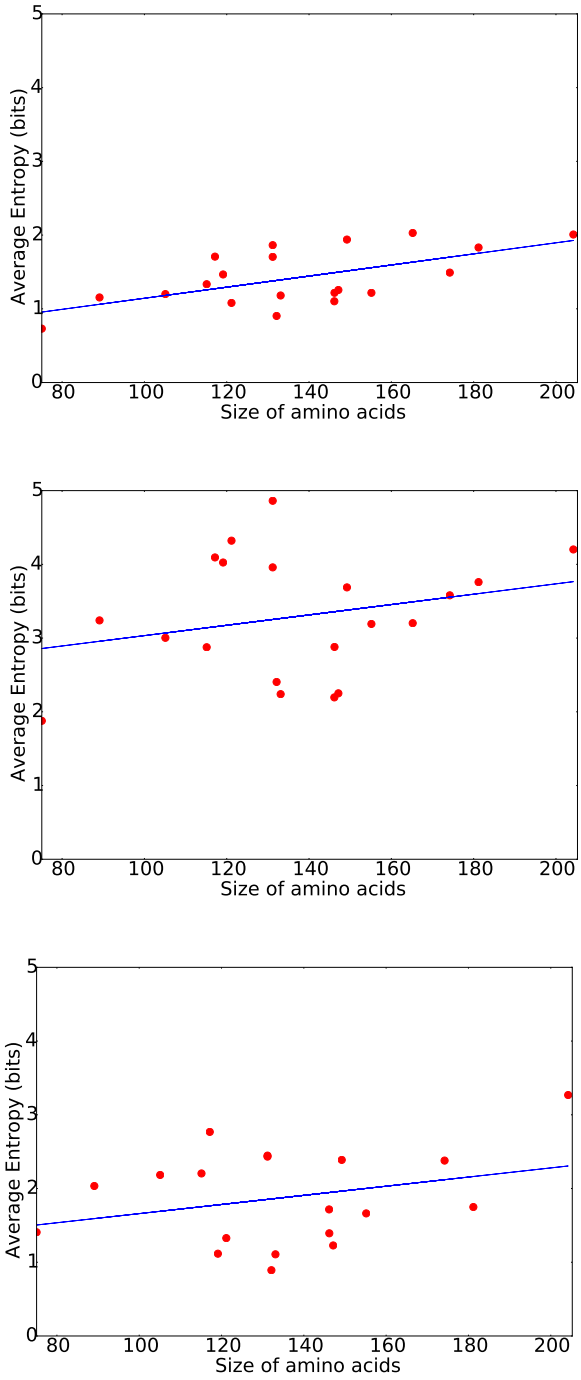

Figure A.5: Plots of average entropy values of residues as a function of amino acid size in the enolase $R^2 = 0.68$ (p = 0.004)(top), mandelate racemase $R^2 = 0.261$ (p = 0.265) (middle), and muconate cycloisomerase $R^2 = 0.297$ (p = 0.216) (bottom) subgroups. A significant but weak correlation is evident in all three subgroups.

A.6 Relationship Between Residue Contact Entropy and Amino Acid Hydrophobicity

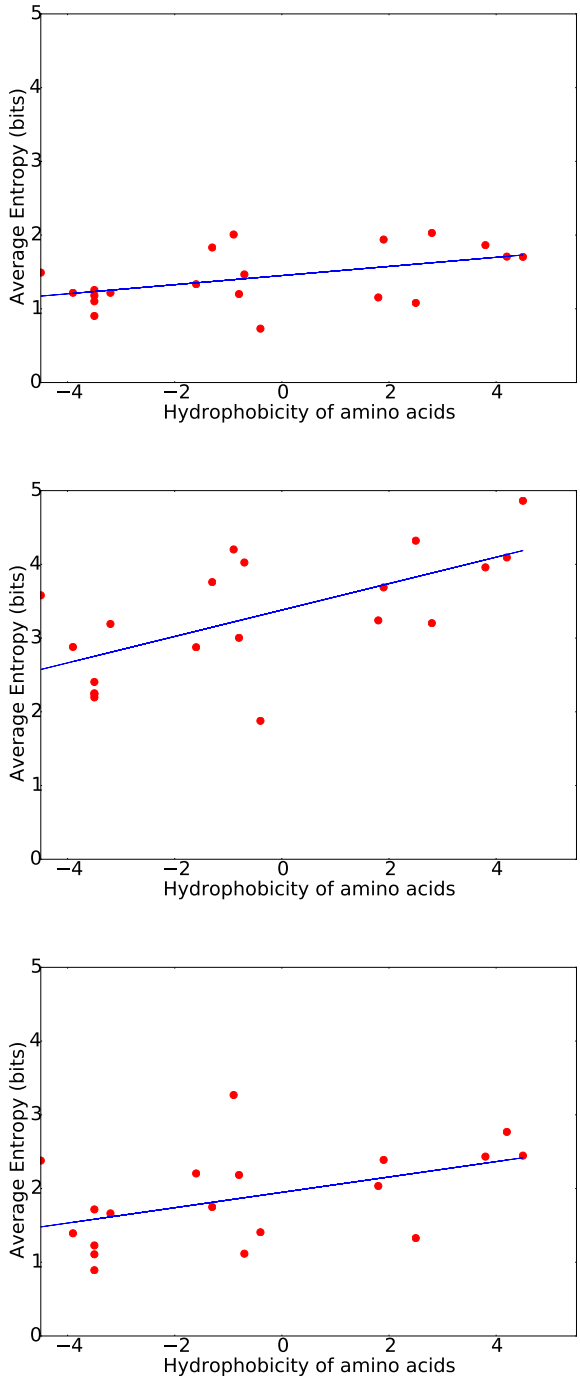

Figure A.6: Plots of average entropy values of residues as a function of hydrophobicity in the enolase $R^2 = 0.483$ (p = 0.031) (top), mandelate racemase $R^2 = 0.65$ (p = 0.002) (middle), and muconate cycloisomerase $R^2 = 0.48$ (p = 0.038) (bottom) subgroups. A significant but weak correlation is evident in all three subgroups

A.7 Residue Contact Entropy Along the Sequence

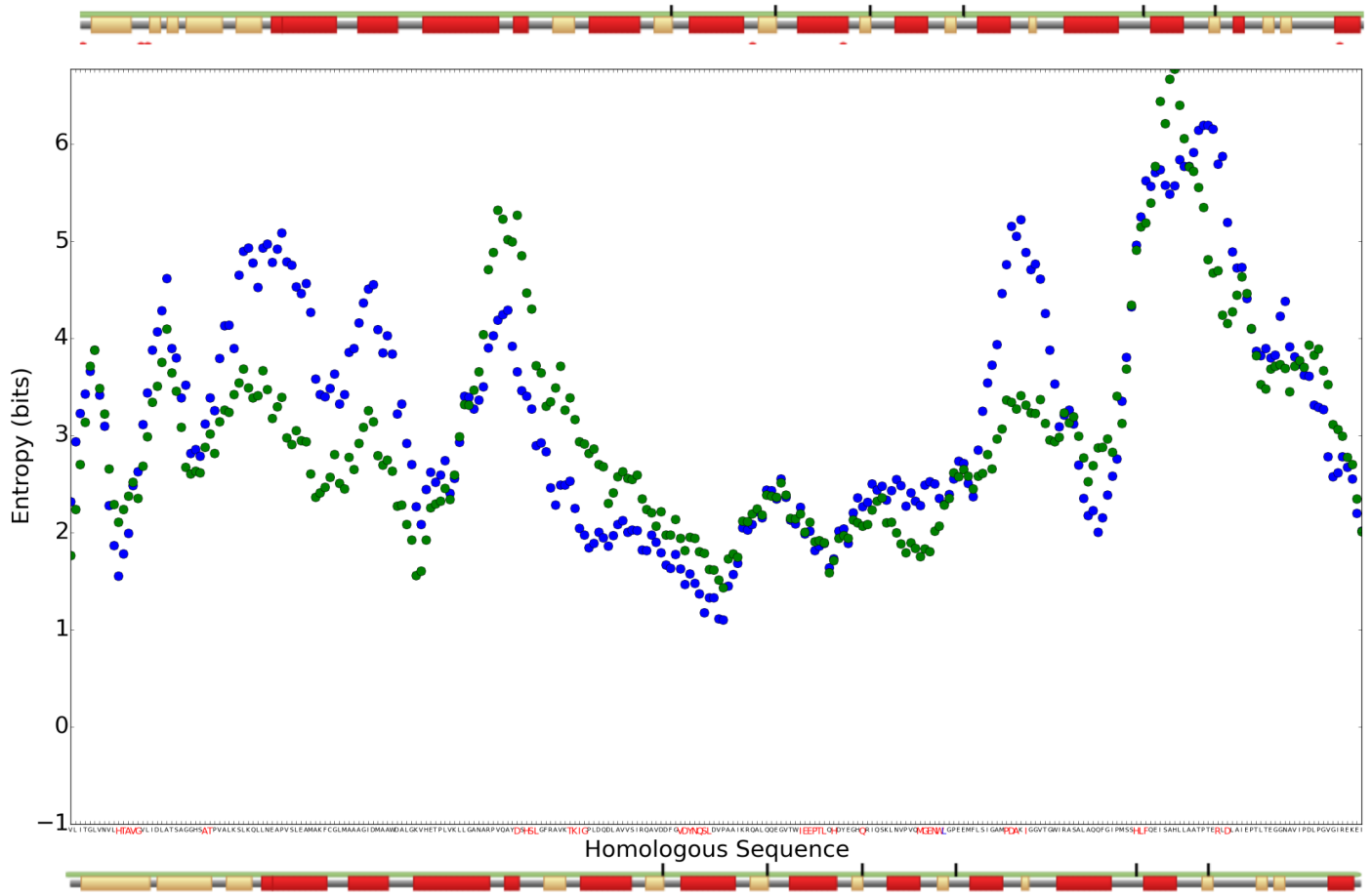

Figure A.7: Running average (window size = 10) of contact entropy values along the sequence of the reference structure for $\mathcal{S}^{\mathcal{L}}$ (blue) and $\mathcal{S}^{\mathcal{U}}$ (green) (structurally aligned sites) in the mandelate racemase subgroup. The upper scale for secondary structure identification belongs to the reference structure in $\mathcal{S}^{\mathcal{U}}$ and lower scale belongs to the reference structure in $\mathcal{S}^{\mathcal{L}}$. The peaks in the plot correspond to different secondary structural elements such as alpha helices (red rods), beta sheets (yellow rods) and coil (grey thin rods). The residues close to the active site are shown by red coloured ticks.

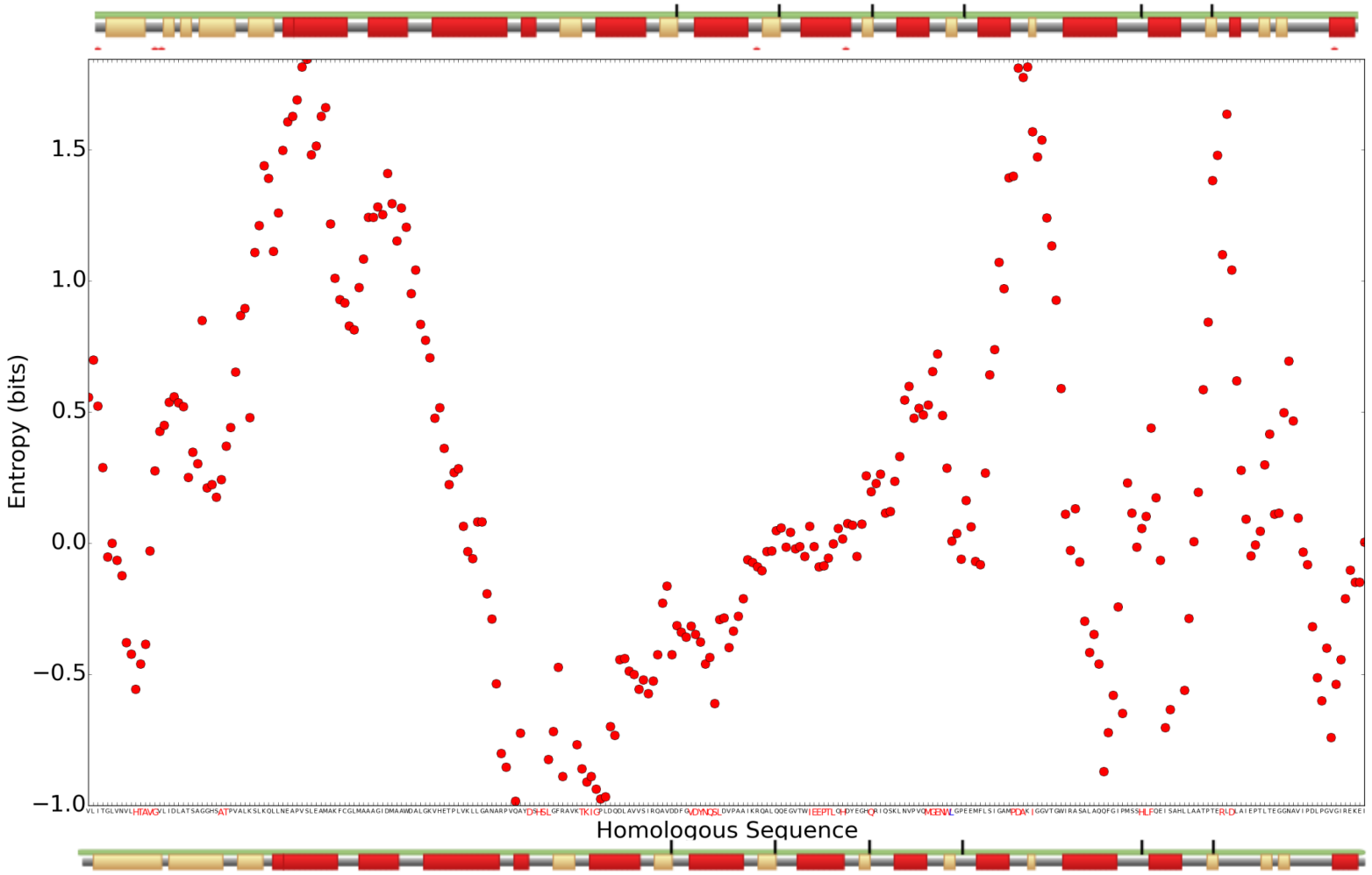

Figure A.8: Difference between the running average of contact entropy values for structurally aligned sites in the reference structure for $\mathcal{S}^{\mathcal{L}}$ and the reference structure for $\mathcal{S}^{\mathcal{U}}$ in the mandelate racemase subgroup. The upper scale for secondary structure identification belongs to the reference structure in $\mathcal{S}^{\mathcal{U}}$ and lower scale belongs to the reference structure in $\mathcal{S}^{\mathcal{L}}$. The peaks and the troughs in the plot correspond to different secondary structural elements such as alpha helices (red rods), beta sheets (yellow rods) and coil (grey thin rods). The residues close to the active site are shown by red coloured ticks.

Figure A.9: Mapping of the peaks (orange) of the contact entropy distribution for ligand-bound and ligand-free structures in the mandelate racemase subgroup (Fig A.7). The peaks mostly correspond to loops. Some of the high entropy residues are in the vicinity of ligands (red spheres). The figure was generated using PyMOL (PDB ID: 1MDR Chain: A)

Figure A.10: Mapping of the peaks (orange) and troughs (magenta) of difference between contact entropy values of ligand-bound and ligand-free structures in the mandelate racemase subgroup (Figure A.8). The peaks and the troughs mostly correspond to loops with some parts of helices and beta sheets. The high difference determines the highest effect of ligand binding on the structures and mostly the loops are found to have an effect of ligand binding. Some of the maximum difference residues are found in close proximity to the ligands (red spheres). The figure was generated using PyMOL (PDB ID: 1MDR Chain: A)

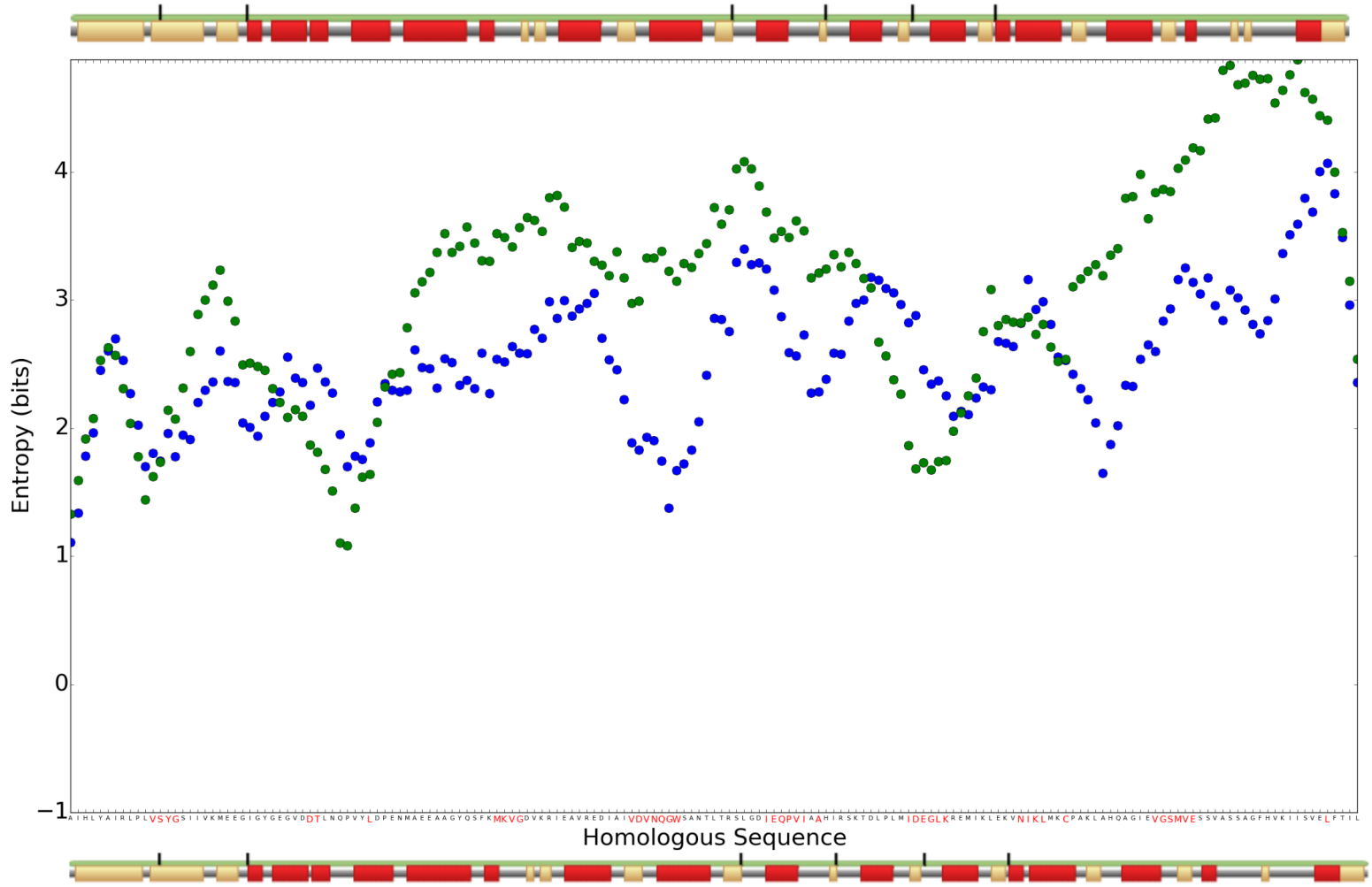

Figure A.11: Running average (window size = 10) of contact entropy values along the sequence of the reference structure for $\mathcal{S}^{\mathcal{L}}$ (blue) and $\mathcal{S}^{\mathcal{U}}$ (green) (structurally aligned sites) in the muconate cycloisomerase subgroup. The upper scale for secondary structure identification belongs to the reference structure in $\mathcal{S}^{\mathcal{U}}$ and lower scale belongs to the reference structure in $\mathcal{S}^{\mathcal{L}}$. The peaks in the plot correspond to different secondary structural elements such as alpha helices (red rods), beta sheets (yellow rods) and coil (grey thin rods). The residues close to the active site are shown by red coloured ticks.

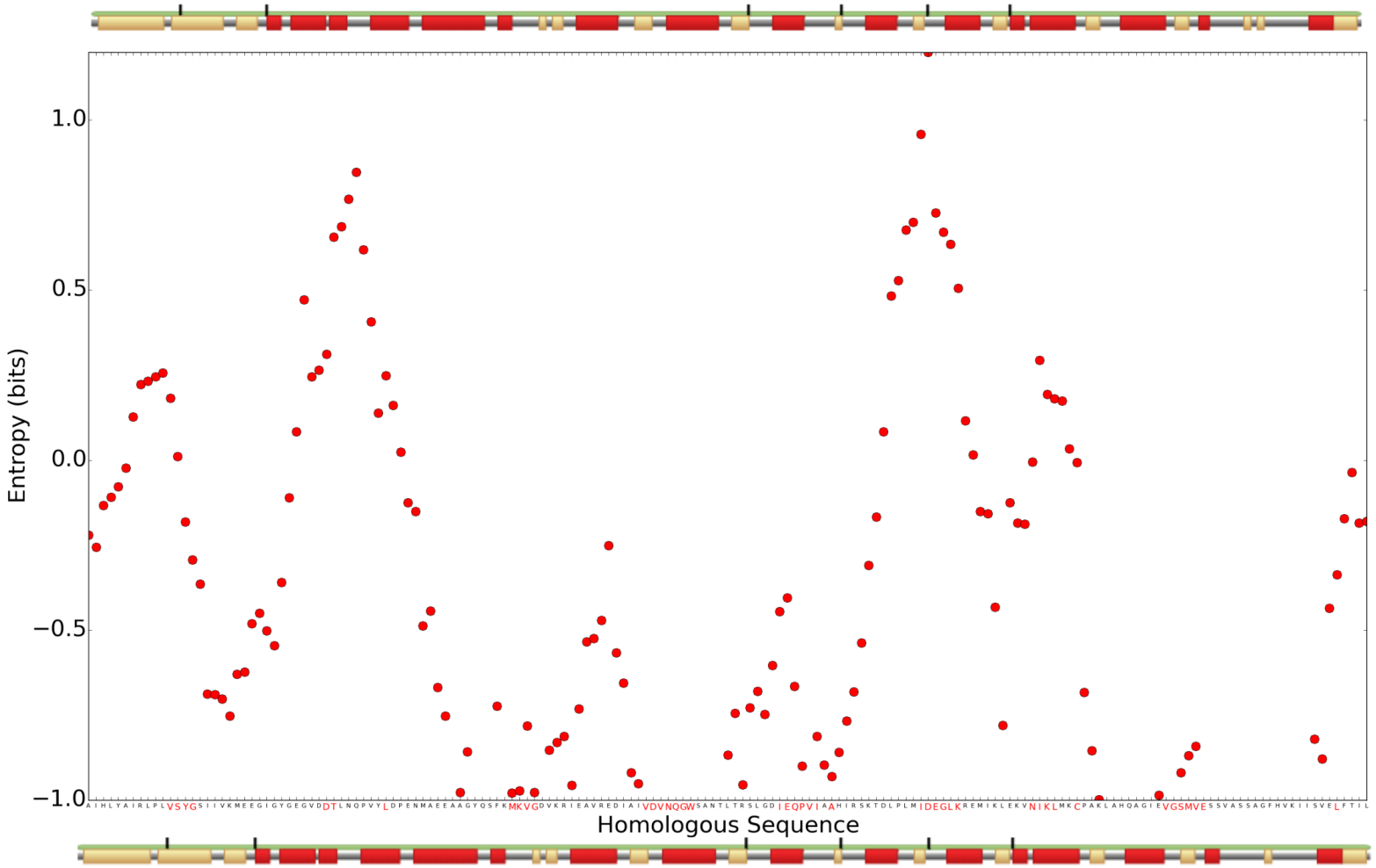

Figure A.12: Difference between the running average of contact entropy values for structurally aligned sites in the reference structure for $\mathcal{S}^{\mathcal{L}}$ and the reference structure for $\mathcal{S}^{\mathcal{U}}$ in the muconate cycloisomerase subgroup. The upper scale for secondary structure identification belongs to the reference structure in $\mathcal{S}^{\mathcal{U}}$ and lower scale belongs to the reference structure in $\mathcal{S}^{\mathcal{L}}$. The peaks in the plot correspond to different secondary structural elements such as alpha helices (red rods), beta sheets (yellow rods) and coil (grey thin rods). The residues close to the active site are shown by red coloured ticks.

Figure A.13: Mapping of the peaks (orange) of the contact entropy distribution for ligand-bound and ligand-free structures in the muconate cycloisomerase subgroup (Fig A.11). The peaks do not correspond to a particular type of secondary structure or location. Some of the high entropy residues are in the vicinity of ligands (red spheres). The figure was generated using PyMOL (PDB ID: 2P8B Chain: A)

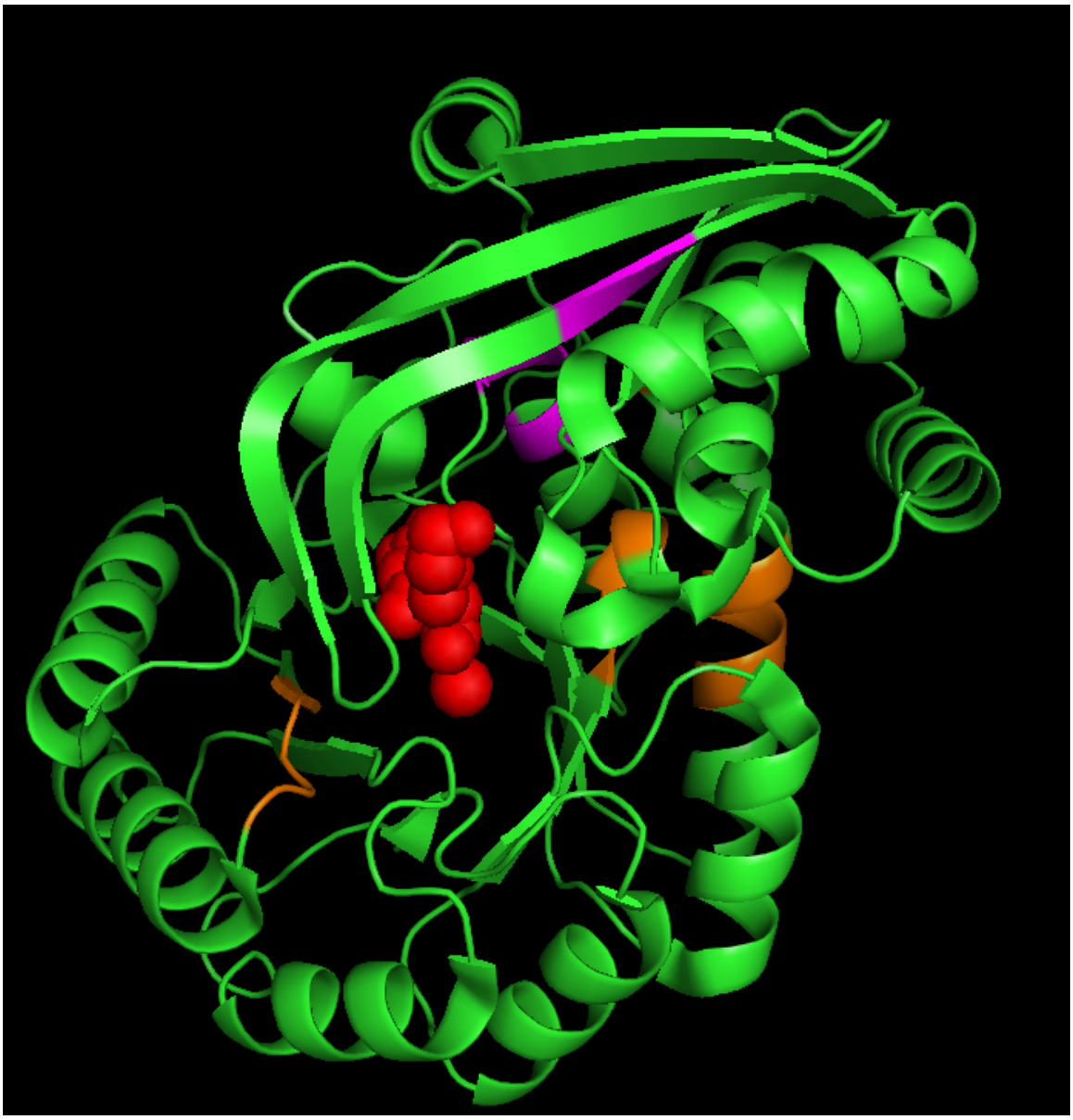

Figure A.14: Mapping of the peaks (orange) and troughs (magenta) of difference between contact entropy values of ligand-bound and ligand-free structures in the muconate cycloisomerase subgroup (Figure A.12). The peaks do not correspond to a particular type of secondary structure. The high difference determines the highest effect of ligand binding on the structures. The loop close to ligand (red spheres) is found to have an effect of ligand binding. The figure was generated using PyMOL (PDB ID: 2P8B Chain: A)

A.8 Comparison of Contact Entropy Distributions for Ligand-bound and Ligand-free Structures

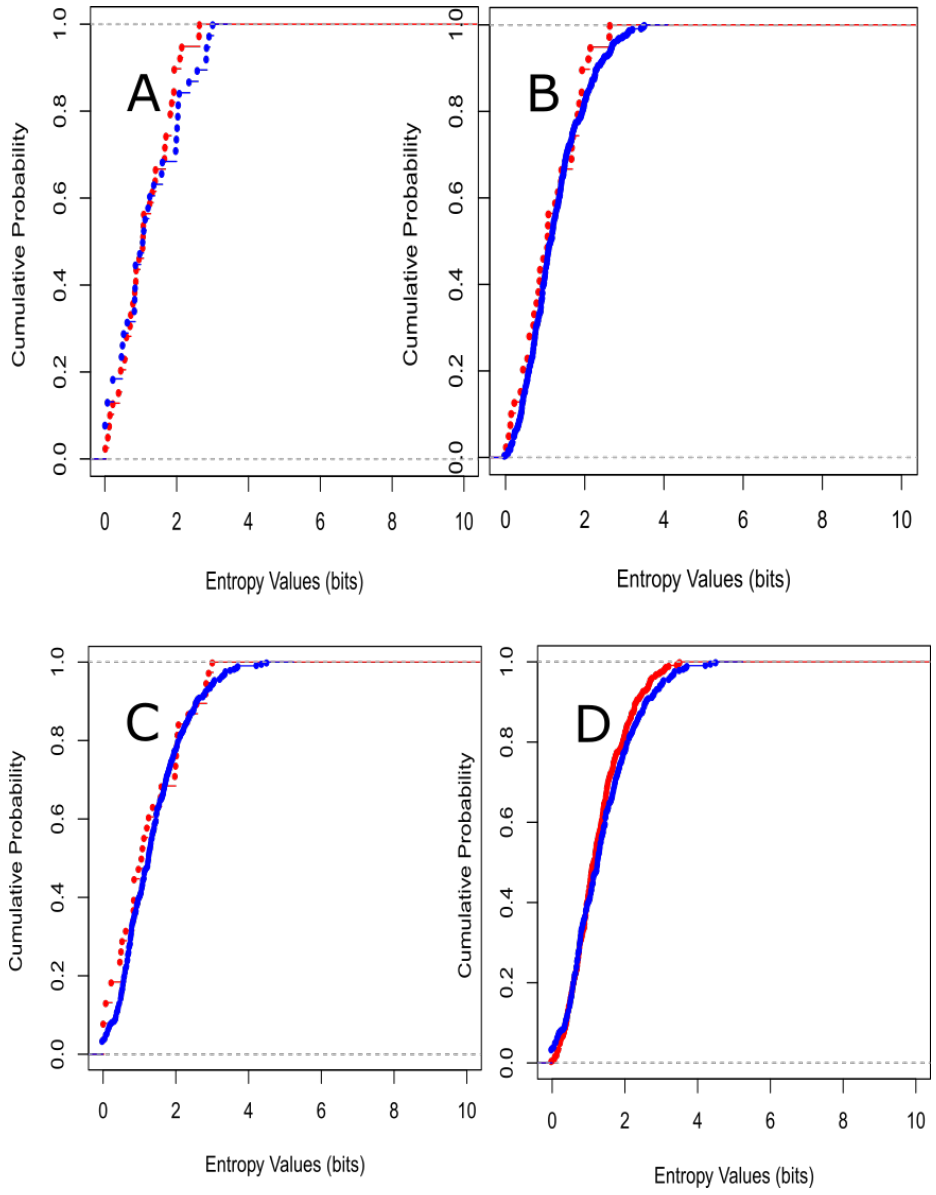

Figure A.15: ECDF for the enolase subgroup **A:** $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ **B:** $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ **C:** $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$ **D:** $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$, and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$. No apparent difference in the distributions is evident.

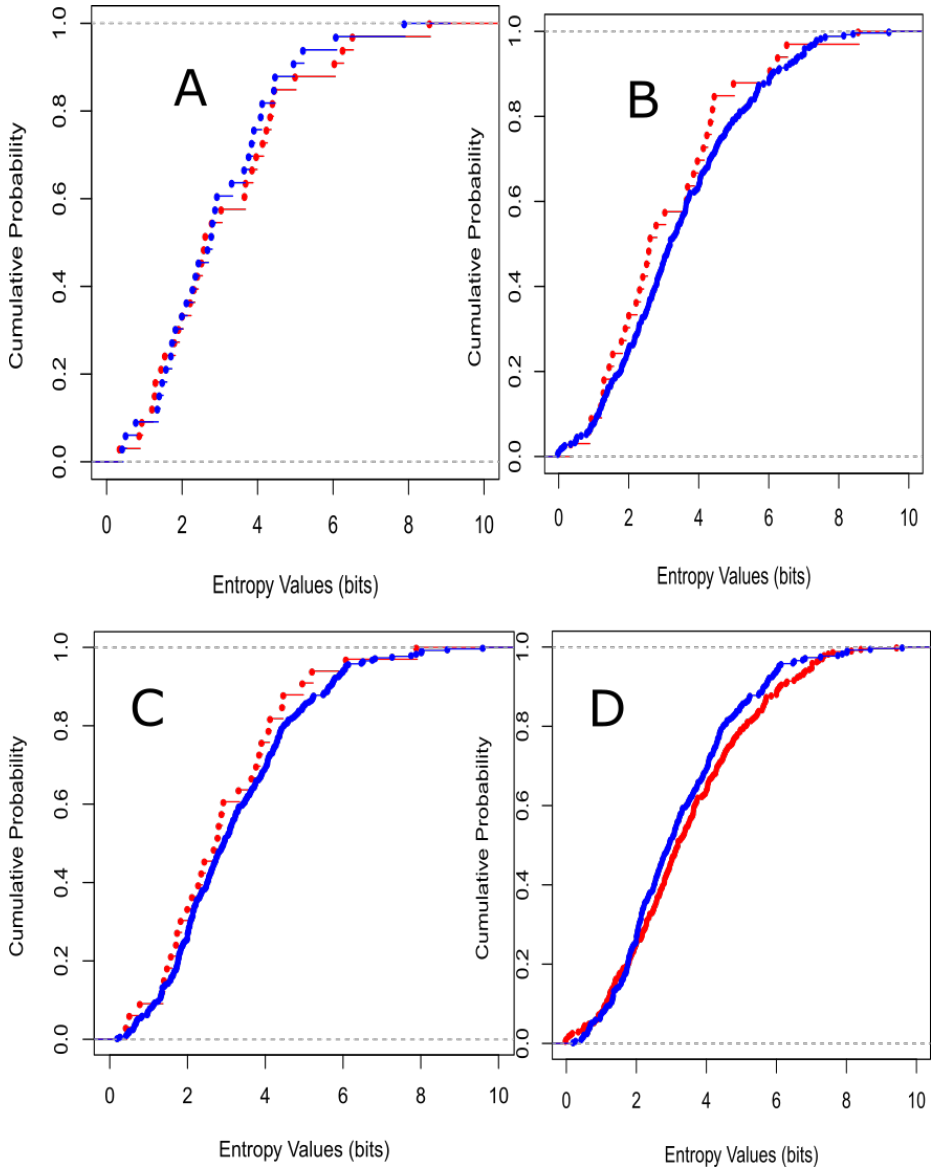

Figure A.16: ECDF for the mandelate racemase subgroup **A:** $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ **B:** $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ **C:** $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$ **D:** $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$, and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$. No apparent difference in the distributions is evident.

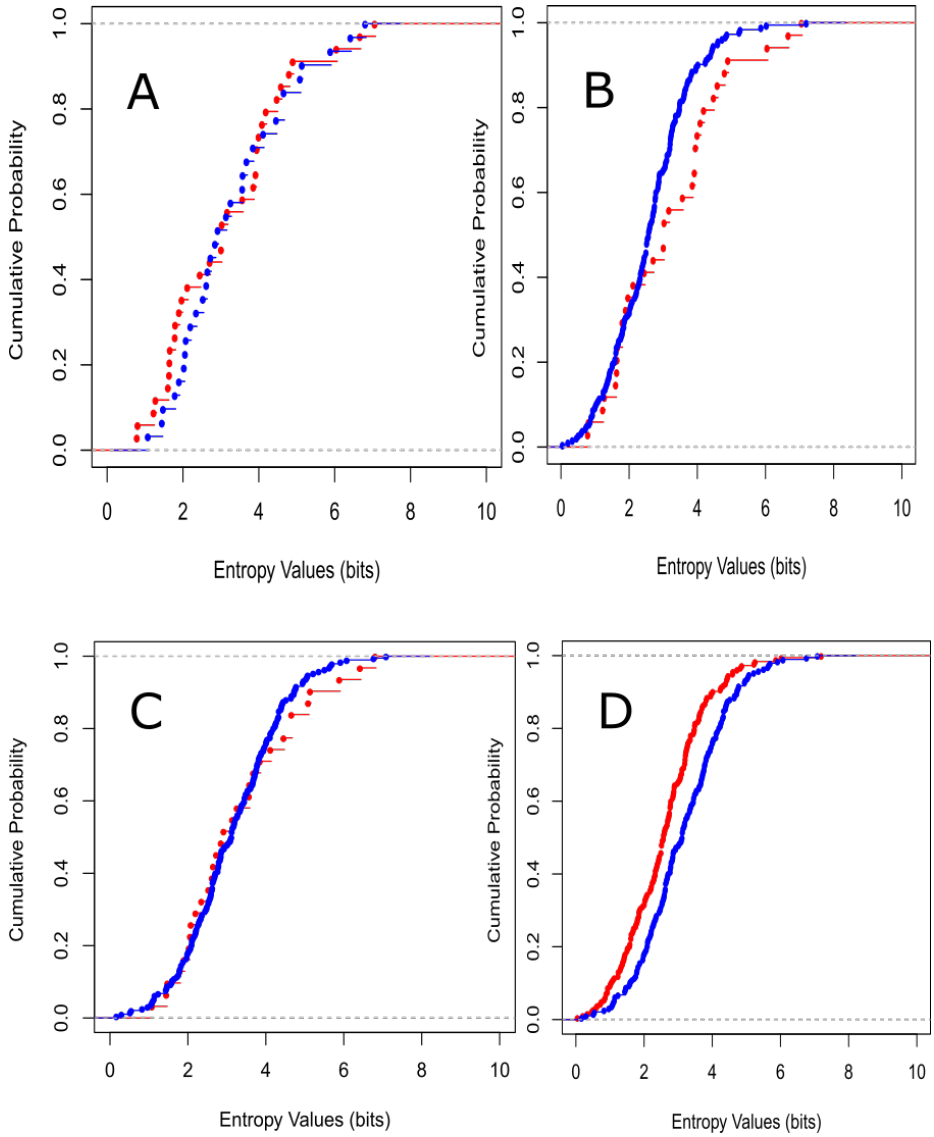

Figure A.17: ECDF for the muconate cycloisomerase subgroup **A:** $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ **B:** $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ **C:** $E^c(\mathcal{S}^{\mathcal{U}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$ **D:** $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$. No apparent difference in the distributions is evident except for $E^c(\mathcal{S}^{\mathcal{L}}, AS)$ and $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$ and $E^c(\mathcal{S}^{\mathcal{L}}, \neg AS)$, and $E^c(\mathcal{S}^{\mathcal{U}}, \neg AS)$.

Bibliography

[1] Ligand, 11 July, 2013 [Accessed: May 2017].

[2] Aftabuddin, M., and Kundu, S. Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophysical Journal 93*, 1 (2007), 225–231.

[3] Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., et al. The structure–function linkage database. *Nucleic Acids Research* (2013).

[4] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology 215*, 3 (1990), 403–410.

[5] Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology 344*, 4 (2004), 1135–1146.

[6] Andersen, N. H. Protein structure, stability, and folding. methods in molecular biology., 2001.

[7] Atilgan, A. R., Akan, P., and Baysal, C. Small-world communication of residues and significance for protein dynamics. *Biophysical Journal 86*, 1 (2004), 85–91.

[8] Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., and Gerlt, J. A. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids. *Biochemistry 35*, 51 (1996), 16489–16501.

[9] Bartoli, L., Fariselli, P., and Casadio, R. The effect of backbone on the small-world properties of protein contact maps. *Physical Biology 4*, 4 (2008), L1.

[10] Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic Acids Research 35* (2006), D301–D303.

[11] Blouin, C. Pdbnet computer software [v1.0]. `http://labblouin.github.io/LabBlouinTools/labblouin.PDBnet.html`.

[12] Bruhn, J., Lehmann, L. E., Röpcke, H., Bouillon, T. W., and Hoeft, A. Shannon entropy applied to the measurement of the electroencephalographic effects of desflurane. *The Journal of the American Society of Anesthesiologists 95*, 1 (2001), 30–35.

[13] Chamorro, A. E. M., Divina, F., and Aguilar-Ruiz, J. S. Evolutionary protein contact maps prediction based on amino acid properties. In *International Conference on Hybrid Artificial Intelligence Systems* (2011), Springer, pp. 303–310.

[14] Creighton, T. E. *Proteins: structures and molecular properties*, 2 ed. Macmillan, UK, 1993.

[15] del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Science 15*, 9 (2006), 2120–2128.

[16] del Sol, A., and OMeara, P. Small-world network approach to identify key residues in protein–protein interaction. *Proteins: Structure, Function, and Bioinformatics 58*, 3 (2005), 672–682.

[17] DeLano, W. L. The pymol users manual. *DeLano Scientific, San Carlos, CA* (2002).

[18] Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., et al. The rcsb protein data bank: a redesigned query system and relational database based on the mmcif schema. *Nucleic Acids Research 33*, suppl 1 (2005), D233–D237.

[19] Emerson, I. A., and Gothandam, K. Residue centrality in alpha helical polytopic transmembrane protein structures. *Journal of Theoretical Biology 309* (2012), 78–87.

[20] Emerson, I. A., and Louis, P. T. Detection of active site residues in bovine rhodopsin using network analysis. *Trends in Bioinformatics 8*, 2 (2015), 63.

[21] Erman, B. Effects of ligand binding upon flexibility of proteins. *Proteins: Structure, Function, and Bioinformatics 83*, 5 (2015), 805–808.

[22] Finn, R. D., Clements, J., and Eddy, S. R. Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research* (2011), gkr367.

[23] Frost, J. Interpretation of p-values, 17 April, 2014 [Accessed: May 2017].

[24] Fuhrman, S., Cunningham, M. J., Wen, X., Zweiger, G., Seilhamer, J. J., and Somogyi, R. The application of shannon entropy in the identification of putative drug targets. *Biosystems 55*, 1 (2000), 5–14.

[25] Gajiwala, K. S., Maegley, K., Ferre, R., He, Y.-A., and Yu, X. Ack1: activation and regulation by allostery. *PloS One 8*, 1 (2013), e53994.

[26] Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., and Almo, S. C. Divergent evolution in enolase superfamily: strategies for assigning functions. *Journal of Biological Chemistry 287*, 1 (2012), 29–34.

[27] Ghasemi, A., and Zahediasl, S. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism 10*, 2 (2012), 486.

[28] Hegyi, H., and Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology 288*, 1 (1999), 147–164.

[29] Hu, J., Shen, X., Shao, Y., Bystroff, C., and Zaki, M. J. Mining protein contact maps. In *Proceedings of the 2nd International Conference on Data Mining in Bioinformatics* (2002), Springer-Verlag, pp. 3–10.

[30] Hu, Z., Bowen, D., Southerland, W. M., Del Sol, A., Pan, Y., Nussinov, R., and Ma, B. Ligand binding and circular permutation modify residue interaction network in dhfr. *PLoS Compututational Biology 3*, 6 (2007), e117.

[31] Humphrey, W., Dalke, A., and Schulten, K. Vmd: visual molecular dynamics. *Journal of Molecular Graphics 14*, 1 (1996), 33–38.

[32] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering 9*, 3 (2007), 90–95.

[33] Isaac, A. E., and Sinha, S. Analysis of core–periphery organization in protein contact networks reveals groups of structurally and functionally critical residues. *Journal of Biosciences 40*, 4 (2015), 683–699.

[34] Kyte, J., and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology 157*, 1 (1982), 105–132.

[35] Lamport, L. Latex users guide and reference manual.

[36] Levene, H., et al. Robust tests for equality of variances. *Contributions to Probability and Statistics 1* (1960), 278–292.

[37] Lu, H., Lu, L., and Skolnick, J. Development of unified statistical potentials describing protein-protein interactions. *Biophysical journal 84*, 3 (2003), 1895–1901.

[38] Mann, H. B., and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* (1947), 50–60.

[39] Márquez Chamorro, A., Divina, F., Aguilar-Ruiz, J., and Asencio Cortés, G. An evolutionary approach for protein contact map prediction. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (2011), 101–110.

[40] Mehta, A. Primary structure of proteins, December 27, 2010, [Accessed: May 2017].

[41] Menke, M., Berger, B., and Cowen, L. Matt: local flexibility aids protein multiple structure alignment. *PLoS Computational Biology 4*, 1 (2008), e10.

[42] Mirny, L., and Domany, E. Protein fold recognition and dynamics in the space of contact maps. *Proteins: Structure, Function, and Bioinformatics 26*, 4 (1996), 391–410.

[43] Munks, B. P. Common amino acid structure, 2009, [Accessed: May 2017].

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

[45] Pearson, W. R. An introduction to sequence similarity (homology) searching. *Current Protocols in Bioinformatics* (2013), 3–1.

[46] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[47] Ryan, J. *Variability in Protein Residue Contact Matrices: A Geometric and Graph Theoretical Approach*. Halifax, Canada, 2016.

[48] Sakai, A., Fedorov, A. A., Fedorov, E. V., Schnoes, A. M., Glasner, M. E., Brown, S., Rutter, M. E., Bain, K., Chang, S., Gheyi, T., et al. Evolution of enzymatic activities in the enolase superfamily: stereochemically distinct mechanisms in two families of cis, cis-muconate lactonizing enzymes. *Biochemistry 48*, 7 (2009), 1445–1453.

[49] Shannon, C. E. A mathematical theory of communication, 2001.

[50] Strait, B. J., and Dewey, T. G. The shannon information entropy of protein sequences. *Biophysical Journal 71*, 1 (1996), 148–155.

[51] van der Lingen, R. Peptide bond, 17 January 2009 [Accessed: May 201].

[52] Van Rossum, G., et al. Python programming language. In *USENIX Annual Technical Conference* (2007), vol. 41.

[53] Vassura, M., Margara, L., Di Lena, P., Medri, F., Fariselli, P., and Casadio, R. Reconstruction of 3d structures from protein contact maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 5*, 3 (2008), 357–367.

[54] Weisstein, E. W. Bonferroni correction.

[55] Wertz, D. H., and Scheraga, H. A. Influence of water on protein structure. an analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules 11*, 1 (1978), 9–15.

[56] Westbrook, J., Ito, N., Nakamura, H., Henrick, K., and Berman, H. M. Pdbml: the representation of archival macromolecular structure data in xml. *Bioinformatics 21*, 7 (2004), 988–992.

[57] Wierenga, R. The tim-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters 492*, 3 (2001), 193–198.

[58] Zhang, E., Brewer, J. M., Minor, W., Carreira, L. A., and Lebioda, L. Mechanism of enolase: The crystal structure of asymmetric dimer enolase- 2-phospho-d-glycerate/enolase- phosphoenolpyruvate at 2.0 resolution. *Biochemistry 36*, 41 (1997), 12526–12534.

[59] Zhang, E., Hatada, M., Brewer, J. M., and Lebioda, L. Catalytic metal ion binding in enolase: The crystal structure of an enolase-mn2+-phosphonoacetohydroxamate complex at 2.4-. ang. resolution. *Biochemistry 33*, 20 (1994), 6295–6300.