

**Role of Protein Microenvironment in Modulating Structure and  
Function of Cystine, Cysteine and Aspartic Acid**

**THESIS**

Submitted in partial fulfilment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

By

**AKSHAY BHATNAGAR**

ID number: 2012PHXF525H

Under the supervision of

**Dr. Debashree Bandyopadhyay**



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**HYDERABAD CAMPUS**

**2016**



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**  
**HYDERABAD CAMPUS**

---

**Certificate**

This is to certify that the Ph.D. thesis submitted by Akshay Bhatnagar, (ID number. 2012PHXF525H) entitled, “**Role of Protein Microenvironment in Modulating Structure and Function of Cystine, Cysteine and Aspartic Acid**” demonstrate original work done by him under my supervision.

Signature in full of the supervisor:

\_\_\_\_\_

Name in capital block letters:

**DEBASHREE BANDYOPADHYAY**

Designation:

**Assistant Professor,**

**Department of Biological Sciences**

Date:

\_\_\_\_\_

## Acknowledgement

The work presented in this thesis was carried out at the Biological Sciences department of BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, Pilani, Hyderabad Campus. I thank all the people who have helped me to accomplish this work. I would like to express my deepest gratitude and thanks to my advisor Professor **Dr. Debashree Bandyopadhyay** for her constant guidance, patience, supervision, suggestion and valuable advises throughout my research. You have been a great teacher and mentor for me and without your guidance, this would have never been possible. I would like to thank you for always encouraging me in my research and helping me to nurture my abilities to mature as a researcher.

I would also like to thank Professor Ramakrishna Vadrevu and Professor Durba Roy for serving as my doctoral advisory committee members. I am grateful to all the Faculty and HOD Biological Sciences Department, Prof. Suman Kapur, BITS-Pilani Hyderabad Campus for their valuable support. I would also like to thank all the lab technicians for their help and support. I deeply acknowledge BITS-Pilani, Hyderabad Campus for funding my fellowship. I would also like to thank Academic Research Division and Doctoral Research committee for their valuable support and backing.

I would also like to thank my friends and fellow research students. I would especially like to thank my father Mr. Gyanesh Prasad Bhatnagar and My mother Mrs. Sadhana Bhatnagar for their constant support and sacrifices done for me.

Akshay Bhatnagar

May 2016

## Table of contents

Title Page	i
Certificate	ii
Acknowledgement	iii
Table of contents	iv
Abstract	v
List of figures	vi-viii
List of tables	ix-xii
1. Introduction	1-22
2. Categorization of cysteine functions based on protein microenvironments	23-57
3. Modulating Effect of microenvironment on structure and function of disulfide bridged cystine residues	58-97
4. Effect of protein microenvironment on the protonation state of aspartic acid side chain using quantum chemical calculations	98-129
5. Conclusions and Future perspective	130-132
References	133-152
List of publications	153
Brief biography of the supervisor	154
Brief biography of the candidate	155

## ABSTRACT

Proteins are the functional units of any living system. Proteins play a vital role in many biological reactions like succinate acid dehydrogenase and pyruvate decarboxylase in respiration, Ribulose-1,5-bisphosphate carboxylase/oxygenase in photosynthesis etc. Protein function depends on the three-dimensional arrangement of amino acids. Amino acids are the building blocks of protein structures. Amino acid form peptide bonds through their main chain atoms to form protein structures. Side chains of amino acids arrange themselves to provide a stable three-dimensional protein structure. This stable structure of the protein has an interior hydrophobic core and a solvent exposed hydrophilic region. However, combination of all the amino acid side chains maintains a heterogeneous dielectric medium (protein microenvironment) within the protein structure. Protein microenvironment is defined here as the three-dimensional arrangement of atoms around a functional group of amino acid side chain. The orientation of neighboring amino acids within a protein microenvironment may dictate the function of an individual amino acid in a particular protein structure.

In this thesis, I have reported different microenvironment regions around the disulfide-bridged cystine and cysteine molecules in high-resolution protein crystal structures. Curation of these microenvironment regions has shown how microenvironment affects various functions of an amino acid in different protein structures. The conservation of microenvironment around cystine and cysteine molecules that are part of specific functional motifs implies the significance of microenvironment in deciding the functional form of such amino acids in protein structures. I have also studied how the protein microenvironment modulates the side chain protonation state of a titratable amino acid like aspartic acid. Aspartic acid model system has been applied here to understand the dielectric nature of different protein microenvironments by comparing it with different solvent systems.

## List of figures

Figures	Page no.
Figure 1.1: Schematic representation of microenvironment around side chain of ASP66 (in Ball and Stick model) in lysozyme (PDB ID: 2LZT).	5
Figure 2.1: The agglomerative hierarchical clustering of the 2070 cystines in 436 high-resolution protein crystal structures has resulted into four cysteine microenvironment clusters represented through individual colors.	31
Figure 2.2: Comparison of the extent of active cysteines (A), metal-binding cysteines (B), redox cysteines (C) and all the three functional cysteines (D) with the respect to the full cysteine residue distribution in terms of microenvironment property descriptor (rHpy).	33
Figure 2.3: Propensities of functional cysteines in different microenvironment clusters.	34
Figure 2.4: Normalized frequencies of different secondary structures in four cysteine microenvironment clusters.	40
Figure 2.5: Preference of the local secondary structure by the active site cysteine residues towards the cysteine microenvironment clusters	40
Figure 2.6: Normalized frequency of cysteines in different enzyme classes distributed across microenvironment clusters.	43
Figure 2.7: The structurally aligned C-x-x-C-H motif in the 12 cytochrome proteins is shown in different colors.	51
Figure 2.8: Local alignment of 11 cytochrome proteins showing conservation of C-x-x-C-H motif	51
Figure 2.9: Zinc binding cysteine motifs. Yellow balls represent sulfur atoms, cyan balls represent C-alpha atoms and the grey balls represent zinc ion. Remaining protein parts are shown in lines.	53
Figure 2.10: The aligned C-x-x-C motif from 9 proteins with zinc binding cysteines as part of C-x-x-C motif.	54

Figure 2.11: Frequency of active, metal binding, and redox cysteines in the buried-hydrophobic, buried-hydrophilic, intermediate and exposed hydrophilic clusters	56
Figure 3.1: Schematic representation of microenvironment around disulfide-bridged cystine residue.	61
Figure 3.2: Schematic representation of different dihedral angles observed in disulfide-bridged cystine structure.	65
Figure 3.3: Distribution of different microenvironment clusters (obtained from hierarchical clustering) around all the half-cystines from 175 different proteins in the microenvironment dataset.	69
Figure 3.4: Microenvironments around half-cystines (expressed in terms of Buried fraction and rHpy values) distributed in the entire microenvironment dataset.	71
Figure 3.5: Microenvironment of redox-active half-cystines with different secondary structures [ $\alpha$ -helix (red diamonds), coil (green cross) or turn (blue stars)] depicted in microenvironment (buried fraction, rHpy) space.	78
Figure 3.6: Structurally aligned regions of six oxidoreductases enzymes and one electron transport protein, all containing redox active Cystines within C-x-x-C motif. Half-cystines.	79
Figure 3.7: Local alignment of six oxidoreductases enzymes (except PDB ID 1JR8) that consists of half-cystines in alpha helix structure	82
Figure 3.8: Superimposed beta-alpha-beta region from the above six proteins, excluding the enzyme with PDB ID: 1JR8.	84
Figure 3.9: Structural alignment of six oxidoreductase enzymes (except PDB ID 1JR8) containing the redox active cystines in the C-x-x-C motif of their beta-alpha-beta fold.	85
Figure 3.10: Conservation of microenvironment around C-x-x-C motif from 6 proteins having redox-active disulfides in alpha-helical conformations.	85
Figure 3.11: Distribution of disulfide in different microenvironment clusters with respect to disulfide structure and strain energies.	93

Figure 3.12: Cystines representing the function of three microenvironment clusters.	96
Figure 4.1: Capped aspartic acid molecule. The side chain carboxylic “O-H” bond is highlighted within a box.	102
Figure 4.2: The DFT optimized capped aspartic acid in 16 different implicit models are shown. The superimposed structure of all the 16 optimized is also shown.	116-118
Figure 4.3: Aspartic acid side chain proton A) not dissociated in presence of water and B) dissociated in presence of NaOH, in vacuum ( $\epsilon=1$ ).	123
Figure 4.4(a): Microenvironment region of 4.5 angstroms around the Asp96 residue of bacteriorhodopsin protein (PDB ID: 1R2N). The aspartic acid side chain carboxylic “O-H” bond is highlighted within a box.	124
Figure 4.4(b): Microenvironment region of 4.5 angstroms around the Asp43 residue of toxin protein (PDB ID: 1ORL). The aspartic acid side chain carboxylic “O-H” bond is highlighted within a box.	125
Figure 4.5: Variation in bond order of carboxylic-OH in the aspartic acid side chain as a function of solvent dielectric constants. Inset pictures highlight the protons of aspartic acid in A) bacteriorhodopsin, B) toxin protein and C) capped aspartic acid in vacuum.	126



## List of tables

Tables	Page no.
Table 1.1: Six enzyme classes and the total number of enzymatic structures in each enzyme class reported in Protein databank	3
Table 2.1: Total number of proteins that belong to the six enzyme classes in the cysteine microenvironment dataset of 436 proteins	27
Table 2.2: List of proteins that are containing the metal binding cysteines bound to a particular metal ion directly, or through a ligand molecule	29
Table 2.3: Statistics of cysteine microenvironment clusters.	32
Table 2.4: Different functional cysteines in different microenvironment clusters	35
Table 2.5: Cysteines in exposed hydrophilic cluster with redox functions curated from the respective research articles.	36-38
Table 2.6: Distribution of secondary structures of all the 2070 cysteines among four microenvironment clusters.	39
Table 2.7: Analysis of secondary structures of active cysteines present in different microenvironment clusters.	41
Table 2.8: Distribution of cysteines in the enzyme class according to the microenvironment clusters. Total number of cysteines in each cluster is given in parenthesis	42
Table 2.9: rHpy values of the eight cysteines in transferase protein (PDB ID: 4NHW).	44
Table 2.10: Frequency of each enzyme class in different microenvironment clusters categorized according to cysteine functions are reported	45-46
Table 2.11: Description of cysteines those are part of the C-x-x-C-H motif in cytochrome proteins.	49
Table 2.12: Buried fraction and rHpy values of cysteines part of C-x-x-C-H motif found in cytochrome proteins	50

Table 2.13: Average rHpy and standard deviation values for the zinc binding cysteines found in different motifs.	55
Table 3.1: Different proteins present in all possible enzyme classes found in the current dataset. It is to note that none of the half-cystines in the dataset were found in protein part of Isomerases and Ligases enzyme class.	63
Table 3.2: Six representative proteins from each cluster pair for quantum chemical calculations	66
Table 3.3: Description of different microenvironment clusters around half-cystines, (S-S) <sup>1/2</sup> .	70
Table 3.4: Distribution of secondary structures in different microenvironment clusters, results obtained from DSSP analysis.	72
Table 3.5: Number of observations of (S-S) <sup>1/2</sup> (individual sulfur atoms from disulfide bonds) found in different SCOP classes	73
Table 3.6: Description of different microenvironment clusters populated with -(S-S) <sup>1/2</sup> - from different enzyme classes.	75
Table 3.7: Number of -(S-S) <sup>1/2</sup> - with different main chain conformations found in enzyme classes, hydrolases, and oxidoreductases.	75
Table 3.8: Buried hydrophobic microenvironment cluster: Secondary structures and strain energies of redox-active Cystines (part of C-x-x-C motif)	76-77
Table 3.9: Mean and standard deviation values (given in parenthesis) for buried fraction and rHpy of functional half-cystines from different enzyme classes.	78
Table 3.10: Cystines present in oxidoreductases, part of buried-hydrophobic cluster and not found in -C-x-x-C- motif region.	80
Table 3.11: Folds present in 11 oxidoreductases and 6 electron transport proteins present in the dataset.	80-81
Table 3.12: Patterns of atomic distribution in all the microenvironments around active half-cystines (all are part of C-x-x-C motif) with alpha-helical conformations in	83

oxidoreductases and electron transport proteins.	
Table 3.13: Buried hydrophilic microenvironment cluster: Cystines buried within protein interior (buried fraction >0.93) yet embedded in the hydrophilic microenvironment (rHpy >0.4).	87-88
Table 3.14: Exposed hydrophilic microenvironment cluster – Hydrolase enzyme: Cystines either part of the catalytic or ligand-binding site or part of the microenvironment of the catalytic site.	90-91
Table 3.15: Average strain energies and average chi3 angle of 700 cystine molecules present in the 6 microenvironment cluster pairs from high-resolution protein crystal structures	92
Table 3.16: Estimated strengths of the donor-acceptor interactions towards the disulfide bond are shown obtained through Natural Bond Analysis of the crystal structures using 631G** basis set	94-95
Table 4.1: Sixteen different types of implicit solvents used in quantum chemical calculations of capped neutral aspartic acid molecules.	104
Table 4.2: The number of steps and total energy for each optimization of capped aspartic acid molecule in 16 different solvents using Hartree-Fock method.	108-109
Table 4.3: Variation in bond lengths, bond order, bond distance and valence electron distribution on oxygen and hydrogen atoms of carboxylic “O-H” in the aspartic acid side chain in various implicit solvents using Hartree-Fock method.	109-110
Table 4.4: The number of steps and total energy for each optimization of capped aspartic acid molecule in 16 different solvents using Density Functional theory (B3LYP).	111
Table 4.5: Variation in bond lengths, bond order, bond distance and valence electron distribution on oxygen and hydrogen atoms of carboxylic “O-H” in the aspartic acid side chain in various implicit solvents using Density Functional Theory (B3LYP).	112

Table 4.6: The number of steps and total energy for each optimization of capped aspartic acid molecule in 16 different solvents using Møller-Plesset (MP2) method.	114
Table 4.7: Variation in bond lengths, bond order, bond distance and valence electron distribution on oxygen and hydrogen atoms of carboxylic “O-H” in the aspartic acid side chain in various implicit solvents using Møller-Plesset (MP2) method.	115
Table 4.8: The type of optimization methods and basis sets used for optimizing the system with capped aspartic acid in presence of single water molecule.	119
Table 4.9: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one water molecule using Hartree-Fock method.	120
Table 4.10: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one water molecule using Density Functional Theory (B3LYP).	120
Table 4.11: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one water molecule using Møller Plesset method.	121
Table 4.12: The type of optimization methods and basis sets used for optimizing the system with capped aspartic acid in presence of single NaOH molecule.	121
Table 4.13: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one NaOH molecule using Hartree-Fock method.	122
Table 4.14: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one NaOH molecule using Density Functional Theory (B3LYP).	122
Table 4.15: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one NaOH molecule using Møller Plesset method.	123
Table 4.16: Variation in aspartic acid side chain carboxylic “O-H” bond in bacteriorhodopsin (PDB ID: 1R2N, D96) and toxin protein (PDB ID: 1ORL, D43).	125
Table 4.17: Variation in aspartic acid side chain carboxylic “O-H” bond in bacteriorhodopsin (PDB ID: 1R2N, D96) and toxin protein (PDB ID: 1ORL, D43).	128

# CHAPTER 1

## Introduction

### 1.1) Amino acids in biochemical reactions

Amino acids are the building blocks of proteins. Proteins are critical for any biochemical reaction, in the form of enzymes, as they alter the rate of the reaction. The chemical reactions involving the enzymes as biocatalyst are known as biochemical reactions (Voet et al., 2008). The enzyme undergoing particular reaction involves conformational changes to efficiently convert substrate to product. Without enzymatic proteins, the biochemical reactions are slow and might not occur under physiological conditions (Cooper, 2000). The enzyme catalysis is a very specific interaction between the substrate and the enzyme. This involves the interaction of the substrate with the catalytic amino acids of the enzyme. The catalytic amino acids are found in the pocket-like structure, known as enzyme active site. Different enzymes are specific to different biochemical reactions involving specific catalytic residues for binding to the substrate. For example, serine proteases contain the catalytic triad of serine, histidine, and aspartic acid (Lehninger et al., 2008; Voet et al., 2008). They are involved in the hydrolysis of peptide bonds. Serine protease family has many proteins those undergo different types of substrate specific reactions. Second example; a hydrophobic amino acid in the binding pocket of chymotrypsin interacts with the hydrophobic side chains of the substrate (Cooper, 2000). Third example; aspartate at the binding pocket of trypsin interacts with basic amino acids (lysine and arginine) of the substrate (Berg et al., 2002). These examples demonstrate that in different biochemical reactions amino acid types are important (Cooper, 2000). Few more biochemical reactions happening in the cell are illustrated in the following section.

### 1.2) Few biochemical reactions within the cell

Cell is the basic unit of life (Hooke, 1665; Alberts et al., 2002). Various cellular processes maintain cell viability. Important cellular processes are growth and development, active and passive

transport, signaling etc. (Alberts et al., 2002; Lehninger et al., 2008; Voet et al., 2008). These cellular reactions occur at different organelle level in the cell. For example, the photosynthesis reactions in C3 plants (plants that perform Calvin Cycle 3) occur at chloroplast, the cellular respiration reactions (glycolysis, oxidative phosphorylation, Krebs's cycle and electron transport chain) occur in cytoplasm and mitochondria, transportation and packaging of molecules by Golgi bodies and endoplasmic reticulum. These cellular reactions are regulated by the enzymes, involving specific amino acid residues.

### **1.3) Conformational changes induced in enzyme structures during biochemical processes**

The conformational change in the three-dimensional enzyme structure exposes the active site to the substrate to facilitate the biochemical reaction. For example, multiple phosphorylations at serine positions 8, 19, 31 and 40 on the regulatory domain of Tyrosine hydrolase enzyme expose the catalytic domain of the enzyme (Daubner et al., 2011). This conformational change in enzyme facilitates the conversion of tyrosine to dopamine.

Similarly, for substrate binding to cytochrome proteins, oxidation of exposed thiol group releases a proton to the catalytic histidine residue that interacts with the substrate molecule (Pokkuluri et al., 2004). Thus, the formation of a disulfide bridge between the two thiol amino acids is a necessary step towards the substrate-enzyme reactions.

Another example is phosphotransferase enzyme where the hydrogen bond between the threonine hydroxyl group and the HIS68 maintains two distinct protonation states on two nitrogen atoms of His68 (deprotonated and protonated states of N $\delta$  and N $\epsilon$  atoms of His68 respectively). This facilitates the catalytic histidine residue as a nucleophile for substrate binding (Liao et al., 1991). Similar acid-base chemistry is observed in proteases, like aspartic proteases those contain carboxylate dyads (Gutteridge and Thornton, 2005). Due to high pKa, one carboxylate is protonated and thus provide a proton to the substrate. The deprotonated carboxylate receives a proton from the neighboring water molecule to attain a nucleophilic hydroxyl group which attacks the substrate molecule.

The above-stated examples show that during an enzymatic reaction the interaction among the amino acid residues plays a major role for catalysis to occur. In the next section, we have discussed different classes of enzymatic reactions and role of amino acids there in.

#### 1.4) Different classes of enzymatic reactions

Enzymatic proteins are classified into six main classes according to International Union of Biochemistry and Molecular Biology (IUBMB) (Bairoch, 2000). The six main enzyme classes and the number of enzyme structures available in Protein databank (Berman et al., 2000) are shown in table 1.1:

Table 1.1: Six enzyme classes and total number of PDB (Berman et al., 2000) structures in each enzyme class are shown.

Enzyme class number	Enzyme class name	No. of PDB structures in enzyme class
1	Oxidoreductases	10844
2	Transferases	16900
3	Hydrolases	22493
4	Lyases	4207
5	Isomerases	2303
6	Ligases	2252

Specific type of enzymatic reactions belong to each enzyme class (Cuesta et al., 2015).

Oxidoreductases are involved in redox reactions. A specific example of this enzyme class is thiol oxidoreductases. Thiol oxidoreductases have thioredoxin-fold with a C-x-x-C motif. Based on the other two amino acids in the motif, the enzyme either perform reduction (as in thioredoxin) or oxidation (as in protein disulfide isomerases) (Varlamova et al., 2013). Transferases class of enzymes work by transferring a chemical group, for example sulfur transferases transfer the sulfur

containing groups (as in cysteine desulfurase) (Zheng et al., 1993). The hydrolase enzyme class, cleaves the chemical bonds by hydrolytic reactions. For example cysteine proteases hydrolyses the peptide bond to degrade proteins. Enzymes that belong to Lyases class break the chemical bonds without oxidation, reduction and hydrolytic reaction mechanisms. For example, cysteine lyase cleaves cysteine thiol bond to form cysteic acid molecule (Tolosa et al., 1969). Isomerases are involved in structural and geometric reorganization among the isomers. The ligases class of enzymes joins the molecules. For example, alanine racemase acts on alanine amino acid and is responsible for L-alanine and D-alanine interconversion (Watanabe et al., 2002).

### **1.5) Role of local environment (microenvironment) on structure and function of amino acids**

The interactions of an amino acid with the surrounding protein components and solvents (mentioned as microenvironment, throughout this thesis) play a significant role in maintaining the protein stability, particularly in small proteins (Tanford, 1962). Physical and chemical properties of amino acids also dictate the protein three-dimensional structure (Jones, 1975). Qualitatively, microenvironment refers to the local arrangement of amino acid residues in three-dimensional space, however, not necessarily a continuous stretch of amino acid sequence (Chen and Bahar, 2004). Mutual cooperativity among the amino acid residues significantly contribute towards the formation of secondary and tertiary structures in protein (Voet et al., 2008). Microenvironment was earlier described as the three-dimensional spatial arrangement of protein segments within six to eight Å distance cutoff around an amino acid, the region chose to be hydrophobic, polar, or non-polar according to the preference of central amino acid (Ponnuswamy et al., 1980). The local environment (first contact shell with radius of 4.25 Å) around an amino acid that characterizes the structure and function of amino acid in protein due its interaction with the protein and solvent atoms was referred to as the microenvironment (Mehler and Guarneri, 1999). The microenvironment defined by Mehler and Guarneri (1999) was updated and extensively tested for large number of high resolution protein crystal structures (Bandyopadhyay and Mehler, 2008). The radius of microenvironment varies in different literature reported works due to variation in their definitions of



the contact shell (Manavalan and Ponnuswamy, 1978; Bandyopadhyay and Mehler, 2008). A schematic of protein microenvironment around an amino acid is shown (Figure 1.1).

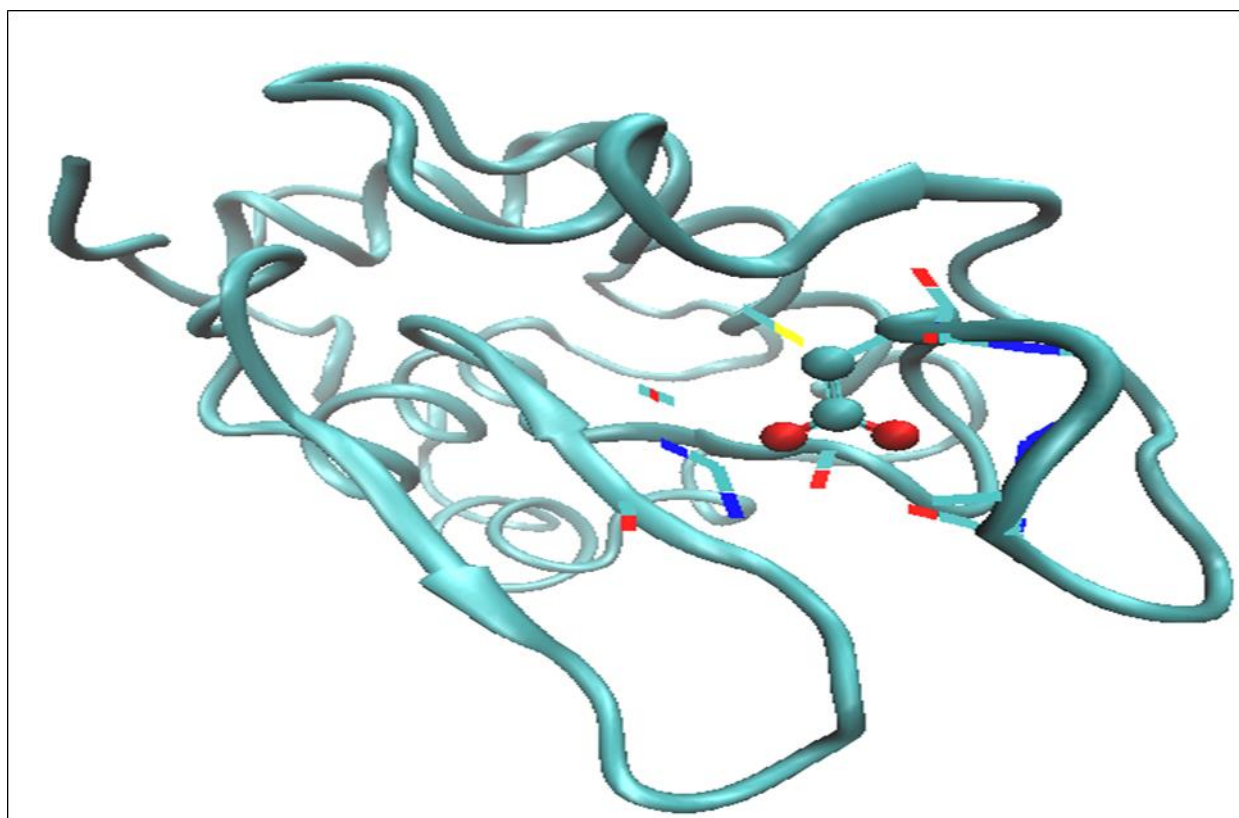


Figure 1.1: Schematic representation of microenvironment around side chain of in lysozyme (PDB ID: 2LZT). Red, cyan, blue and yellow colors represent oxygen, carbon, nitrogen and sulfur atoms. The microenvironment is shown in lines, whereas the aspartic acid 66 side chain is shown in ball and stick. The remaining protein is shown in new cartoon representation. The figure was generated using Visual Molecular Dynamics visualization software (Humphrey et al., 1996).

### **1.6) Role of microenvironment in enzymatic reactions**

Alteration in the microenvironment around the catalytic and ligand binding residues modulates the catalytic and binding efficiency of the amino acid residues (Pinitglang et al., 1997). Presence of amino acids that are part of binding site region of an allosteric enzyme, in the microenvironment region of each other is crucial for the enzyme function (Cornish-Bowden, 2014). This facilitates easy interaction of the active site residues and hence helpful in opening of the active site pocket for substrate binding. The interaction among the amino acid residues in the microenvironment plays a

crucial role in catalytic activity of an enzyme, for example: the catalytic triads of the 7 representative proteins from glycosidase hydrolase family were compared to show the conservation of position 6, where oxygen atom is found to be conserved irrespective of the residue present in the vicinity of the catalytic residue histidine so that it can form hydrogen bond to maintain the protonation state of histidine (Shaw et al., 2002). This shows that conservation of position plays an important role in catalytic action of the glycosidase hydrolase family proteins. Catalytic action of an enzyme is dependent on the local conformations of both catalytic residues and the residues in the microenvironment, for example: Asp192 in sucrose phosphorylase (PDB code: 1R7A) acts as a nucleophile, forming a covalent bond with the substrate. The nearby residue Arg190 ensures that Asp192 is unprotonated (Gutteridge and Thornton, 2005). Similarly, Amino acids in different microenvironments have different functionalities, for example; Exposed thiol groups are found at active sites (Pokkuluri et al., 2004). Hydrophobic side chains exposed to solvent are potential sites for protein-protein interactions (Ma et al., 2003).

### **1.7) Role of protein microenvironments in normal biochemical processes, diseases and their interventions**

Interruption in normal physiological process leads to diseases like diabetes, cancer, neurodegenerative disorders, obesity and many metabolic disorders. The analysis of biochemical and metabolic pathways provides the basis of medicine oriented research and to answer the disease related questions at molecular level. For example, mutations at amino acid level those lead to diseases, designing of drug molecules for specific targets etc. Therefore, the relationship between disease status and the biochemical pathway involved is crucial. For example; the distinction between diabetes mellitus and diabetes insipidus became more comprehensive after understanding its relationship with blood sugar level, which initially was detected in urine samples (Trowell, 1982). The metabolic process involves the role of nutrients like carbohydrates, fatty acids and amino acids. These help in maintaining the adequate homeostasis of a living system and a normal biochemical reaction (DeBerardinis and Thompson, 2012). Disruption in a biochemical process

may lead to diseases. As biochemical pathways and related diseases involve specific amino acids, alteration in their microenvironment could play profound role in respective diseases and their interventions.

### **1.8) Role of microenvironment in diseases caused due to protein aggregation**

Amino acids like aspartic acid, glutamic acid and tyrosine are involved in misfolding of proteins that leads to protein aggregation causing diseases like Alzheimer's, Parkinson's, etc. (Aguzzi and O'Connor, 2010). In case of Alzheimer's disease the beta amyloid protein ( $A\beta$ ) aggregates to form clumps (main amino acid involved is aspartic acid). These clumps known as amyloid plaques destroy neurons. The amyloid beta peptide is of two types: one of which has 40 amino acids and one of which has 42 amino acids. They are produced by two mechanisms using two types of proteins: alpha-secretase ( $\alpha$ -secretase) and beta-secretase ( $\beta$ -secretase, BACE1). They cleave Amyloid Precursor Protein (APP). If alpha-secretase cleaves APP, there is no formation of  $A\beta$ . If APP is cleaved by beta-secretase it can then be further cleaved by gamma-secretase ( $\gamma$ -secretase) to form either a 40 amino acid amyloid peptide ( $A\beta_{40}$ ) which is soluble & mostly innocuous — or a 42 amino acid peptide ( $A\beta_{42}$ ) which clumps together to form insoluble amyloid plaques. Alpha-secretase cleavage occurs at the cell surface, whereas beta-secretase acts at the endoplasmic reticulum. Gamma-secretase produces  $A\beta_{42}$  if cleavage occurs in the endoplasmic reticulum and  $A\beta_{40}$  if the cleavage is in the trans-Golgi network (Hartmann et al., 1997). Thus it shows that surrounding conditions affect the amino acid protonation state to alter the enzymatic reactions. Protein microenvironment around respective amino acids modulate the product formation.

### **1.9) Role of microenvironments in diseases caused due to point mutations, deletions and insertions**

Mutations are the permanent changes in the genomic DNA of an organism. Although mutations are rare but are generally harmful. An alteration in the single base pair in the nucleotide causes point mutations. For example; in sickle cell anaemia disease, the 6th codon of  $\beta$ -globin gene contains a point mutation that leads to substitution of glutamic acid by valine. Due to this change, the Red

blood corpuscles (RBC's) are of sickle shape instead of normal RBC's. Similarly, an insertion of base pairs in the codon sequences changes the whole codon read and leads to formation of a faulty protein. For example; an insertion of 38 base pair region in the factor VIII gene leads to inactivation of the factor VIII protein. This results in development of a fatal disease known as haemophilia. Like insertion, deletions of the base pairs are also responsible for many diseases. For example; a deletion of a single codon at position number 508 in exon 10 of the Cystic Fibrosis Trans-membrane Regulator (CFTR) gene leads to cystic fibrosis. This condition hampers the regulation of flow of chloride ions into the cells (Voet et al., 2008). Any mutations at nucleic acid or protein level affect the microenvironments of the wild amino acids in the protein. This change in microenvironment may relate to respective diseases.

#### **1.10) Role of microenvironment in diseases related to protein metabolism**

Disruption in the biochemical process involving enzymes and other structural proteins leads to many diseases. For example; defect in cystine and dibasic amino acid transporter leads to cystinuria disease, commonly known as kidney stones. Another kidney failure related disease is known as Fabry disease that is caused by the deficiency of alpha-galactosidase A. Major disorders related to amino acid metabolism include phenylketonuria, tyrosinemia, homocystinuria, non-ketotic hyperglycinemia, and maple syrup urine disease. In case of phenylketonuria, due to lack of phenylalanine hydroxylase enzyme the conversion phenylalanine to tyrosine is hampered. Tyrosine is a precursor of many hormones like melanin etc. (Alberts et al., 2002). From the above discussion it is clear that many of the metabolic disorders directly relate to the alteration in amino acids and its respective microenvironments.

#### **1.11) Effect of temperature on protein microenvironments around amino acids**

The microenvironment around an amino acid residue exclusively depends on the physiological conditions of temperature and pH values in the region. Alteration of the temperature may alter the microenvironment around the amino acid residue. For example: there is a decrease in microenvironment mobility around the tryptophan residue (W144) in mammalian tyrosyl-tRNA

synthetase protein with increase in temperature, however in elevated temperature (greater than 52°C), a complete disorder of the microenvironment that is; disruption of the surrounding atoms leading to denaturation of the protein was observed. In the temperature range of 37°C to 52°C a transition from a buried microenvironment to a polar microenvironment was observed (Kordysh and Kornelyuk, 2006).

## **1.12) Role of microenvironment in modulating certain physicochemical properties of amino acids in protein structures**

### **1.12.1) pKa around titrable residues**

The local microenvironment of an amino acid is a significant contributing factor towards the overall stability of the protein structure as the microenvironment involves the contribution of the hydrophobic interactions in the protein among the amino acids and with the solvent (Zhang and Kim, 2000). In protein structures, microenvironment is a significant factor towards modulation of pKa around the titrable amino acid residues (Mehler and Guarnieri, 1999). It was reported that microenvironment around titrable amino acid residues dictates the protonation state of those functional groups (Mehler and Guarnieri, 1999). A particular protonation state of a titrable amino acid is most likely important to the overall protein structure or function (Harris and Turner, 2002; Chen and Bahar, 2004; Gutteridge and Thornton, 2005) . Similarly, antigen-presenting dendritic cells maintain a reducing microenvironment around cysteine amino acids to activate the T-lymphocytes (Angelini et al., 2002).

Mutual interactions among several amino acids lead to a change in microenvironment and function of a particular protein. For example; phenylalanine-histidine interaction in the alpha-sarcin ribonuclease protein leads to an alteration in the pKa of the active HIS137 amino acid residue that is responsible for ribonuclease activity. This alteration in the pKa is a common phenomenon in all microbial ribonuclease enzymes. Removal of phenylalanine-histidine interaction leads to decrease in activity of ribonuclease enzyme (Pérez-Cañadillas et al., 1998).

### **1.12.2) Redox properties of cysteine and cystine residues**

Structural parts of the protein, those are covalently linked by disulfide bonds, are also influenced by surrounding microenvironment (Bagley and Altman, 1995). Modulation in the microenvironment leads to changes in the redox states of amino acids mainly thiol group containing cysteine amino acids. These changes in the redox states can be used for understanding the mechanism of homeostatic pathways during oxidative stress conditions and to study various redox-sensitive processes (Nair et al., 2007). Switching between two redox states in cysteines, thiol and disulfide modulate the function of lymphocyte cells and macrophage cells (Moriarty-Craige and Jones, 2004).

### **1.12.3) Hydrogen bonding in photoreceptor proteins**

Hydrogen bonding network in protein structures is also modulated by protein microenvironment. For example, rhodopsin II protein, known as photorhodopsin protein, transform from functionally inactive state to a functionally active state via a change in hydrogen bonding network during light-dependent transport reaction (Moukhametzianov et al., 2006). These modulations give rise to changes in the tertiary structure of the protein leading to a decrease in number of hydrogen bonds connecting helices C to G. This hydrogen bonding cascade alters the pKa values of the Schiff base and an aspartic acid (ASP75) leading to transfer of protons and subsequent redistribution of charges. This helps in the transition of the protein from an inactive K –state to a signaling M-state due to changes in the tertiary structure of the receptor (Moukhametzianov et al., 2006).

### **1.13) Annotation of protein local structure: comparison between microenvironment analysis and geometric consideration**

Microenvironment based analysis and annotation of the protein function and recognition of the recurring motifs is more accurate than the exploration of geometric core based methods (Wu et al., 2010). To identify novel 3D motifs; the cysteine amino acid microenvironment was clustered using K-means and hierarchical clustering (Wu et al., 2010). This resulted in the formation of 70 clusters with some of the clusters representing the motif like metal binding and phosphatase activity. The microenvironment chosen was of 7.4 angstroms around the cysteine amino acid residue.

Microenvironment based categorization of the amino acids may help to establish the structure-function relationship in novel proteins and an efficient characterization of the parameters that determine protein functions (Wu et al., 2010). Apart from the enzymatic proteins, the microenvironment also has a crucial part in protein-protein interactions. Residues involved in protein-protein interactions are referred to as the hot spot residues. The hot spot residues prefer only hydrophilic amino acid residues within their vicinity, such a microenvironment creates a hydrophobic microenvironment for the hot spot residues by interaction with the solvent water molecules and prevent the hot spot amino acid from solvent exposure and thus helps in maintaining the correct form of hot spot interfacial residues for efficient protein-protein interactions (Ye et al., 2014).

#### **1.14) Free energy change in a biochemical reaction and its relation with protein microenvironment**

The free energy for any biochemical reaction, even for interaction among the amino acids and with the solvent should be less than zero for it to occur spontaneously (Berg et al., 2002). Hydrophobicity estimates the free energy of partition of an amino acid in aqueous and in a non-polar solvent (Tanaka and Scheraga, 1976). Amino acids prefer a particular microenvironment region based on the hydrophobicity of the microenvironment which is imperative for its function in the proteins (Jha et al., 2010). The hydrophobic interaction between the side chain amino acid function groups and the atoms in the microenvironment of the amino acid have a major role in the protein stability and are quantified as hydrophobic moments (Eisenberg et al., 1986). These hydrophobic moments were based on the hydrophobicity of the atoms and were calculated from the respective atomic coordinates of the proteins. These hydrophobic moments were approximated for free energy calculations to estimate proteins stability.

#### **1.15) Literature reports on quantification of microenvironment**

Microenvironment around the catalytic residues in a protein plays a crucial role in deciding the protein conformations which directly implies to the catalytic activity of the enzyme and the stability

of that particular enzymatic reaction (Gutteridge and Thornton, 2005). This shows the significance of determination and characterization of the microenvironment. One of the first assumptions made to characterize microenvironment was done in 1962 by Tanford, where he proposed that charged residues are found in the solvent exposed region and the hydrophobic and non-polar residues will be mostly present in the protein interior (Tanford, 1962). Such an arrangement of atoms helps to maintain the stable native structure of the protein and prevents their denaturation. At the basic level of understanding, microenvironment represents the physical and chemical properties of the atoms around the functional group of an amino acid in the proteins. To explore the relationship between the physical and chemical properties of the protein structure, a statistical attempt was made in 1968 using six properties (bulkiness, polarity, RF, pI, pK1, and hydrophobicity) of amino acid side chains (Zimmerman et al., 1968). Similarly, few other statistical (Nagano, 1973; Chou and Fasman, 1974), pattern recognition (Denisov et al., 1973) and information theoretic approaches (Robson and Pain, 1972) were used to explore the sequence-structure relationship and the basic parameters for structure assessment in proteins. Although, these statistical and theoretical efforts provide an insight towards the impact of the microenvironment in protein structure and function but still they are based on the average results of datasets and may be incomplete to explain the exact nature of microenvironment in proteins. Thus, there was a need to understand microenvironment, quantitatively instead of a qualitative analysis (Zimmerman et al., 1968; Jones et al., 1975). Therefore, in 1976, Cyrus Chothia quantitatively differentiated the microenvironment into two major groups; the part of protein buried inside the core region and the solvent exposed region of proteins (Chothia, 1976). This quantification was done by calculating the accessible surface area of the different residues in the protein structures. The accessible surface area of proteins is known as the space around an atom where a water molecule can be adjusted without disturbing its van der Waals interactions with that atom and not probing into other atoms (Lee and Richards, 1971). It is important to note that the accessible surface area is directly related to the hydrophobicity of the atoms and contributes to the hydrophobic free energy of interaction within the involved atoms (Chothia, 1976). The accessible surface area represents the role of water solvent into the structural



and functional relationship of proteins, but the protein microenvironment contribution is also an important factor towards the same. To evaluate the protein microenvironment contribution, the van der Waals contacts among the amino acid side chains was calculated for all the 20 essential amino acids (Tanaka and Scheraga, 1976). The calculations showed that polar residues are found in the exposed region whereas the hydrophobic and non-polar amino acid residues mostly prefer the protein interior core to maintain the stability due to hydrophobic interactions. Although there are many exceptions to this observation due to the respective role of the protein (Tanaka and Scheraga, 1976).

To further analyze and determine the overall microenvironment around the amino acid residues in proteins, the microenvironment was quantified based on two parameters: (1) the distance range necessary for influencing the amino acid's functional group was estimated (Manavalan and Ponnuswamy, 1977). The arrangement of neighboring residues was obtained for a distance range between 3 to 10 angstroms from the C-alpha atom of the amino acid residue, whose microenvironment is to be quantified. (2) Identification of the propensity of a particular amino acid in the neighboring region (6-8 Å) of each amino acid. After 8 angstroms the frequency of occurrence within the microenvironment region decreases steeply and hence 8 angstroms distance was considered the optimal distance cut-off for studying the effect of surrounding residues. Based on these two parameters the preferred microenvironment (in terms of amino acid propensity) around each amino acid was calculated (Manavalan and Ponnuswamy, 1977). To further refine the microenvironment quantification; the effect of hydrophobic character of the surrounding residues in determining or modulating the chemical and physical properties of the amino acid were included (Manavalan and Ponnuswamy, 1978). The hydrophobicity of the surrounding residues was termed as 'surrounding hydrophobicity' and it is defined as the summation of the hydrophobicities of the amino acids within the 8 angstroms distance cut-off of the amino acid's functional group in the protein structure. Manavalan and Ponnuswamy analyzed the role of hydrophobic environment in protein folding using the average "surrounding hydrophobicity" from 14 crystal structures around

an amino acid (Manavalan and Ponnuswamy, 1978). Microenvironment determination was done computationally by using hydrophobic scales starting with Rekker fragmental system to calculate partition coefficients of various functional groups (Rekker, 1977). The Rekker fragmental constants were established on the basis of experimental water–octanol partition coefficients that varied a set of atomic or fragment parameters to calculate the partition coefficients of certain functional groups. However, Rekker’s fragmental constant was initially developed for small molecules, independent of protein structures. The first attempt to understand protein microenvironment was done with 21 protein crystal structures to develop a hydrophobicity scale for amino acid residues to understand the protein folding mechanisms in protein structures (Ponnuswamy et al., 1980). The next attempt to understand protein microenvironment was done in terms of solvation free energy. The hydrophobic free energy was computed based on the hydrophobic contributions and solvent accessibility of a particular amino acid to understand the energetics involved in protein folding (Eisenberg et al., 1986). The microenvironment represents the contribution from both water solvent and the protein. Therefore, it was also important to efficiently calculate the extent of any amino acid’s functional group buried inside the core or exposed to the water solvent. For such a calculation, GEPOL algorithm (Pascual-ahuir et al., 1994) was developed to more clearly understand the interactions between a solute and solvent. The GEPOL program offers more refinement to the accessible surface area calculations by adding two more parameters; Vander Wall surface and molecular surface. The algorithm calculates the Solvent-Excluding Surface by replacing an inaccessible area of the solvent with a set of new spheres, and thus iteratively, computes the extent of amino acid buried inside the protein core (Pascual-ahuir et al., 1994). Alternatively, a specialized grid-based system (Bagley and Altman, 1995) enclosing the characteristics of different types of atoms found in protein structures representing the microenvironment atoms around a given functional group of an amino acid in the protein structures was used to establish an efficient method of structure-function relationship. A grid was defined as the 3-dimensional cube. The length of the diagonal was taken as the length of carbon-oxygen single bond. This enables to have an edge length of 0.826 angstroms. The rationale for such dimension was to ensure that two nearby atoms rarely

occupy the same unit cell of the cubic grid. The grid axis was declared according to the coordinates of the atoms in the protein structures crystallized and stored in Protein Databank database. Different properties of the protein atoms determine the values in the grid cell. The microenvironment varies due to variation in the properties of the atom present in the vicinity of the amino acid. The results showed the (i) microenvironment based preference of amino acid residues as calcium binding sites, (ii) comparison between disulfide bridging and non-bridging environments and (iii) the conservation of microenvironment around histidine, serine, and aspartic acid catalytic triad in the serine protease family proteins (Bagley and Altman, 1995).

### **1.16) Literature reports on experimental techniques to calculate microenvironment**

Experimental work using fluorescence emission spectrum for calculating the microenvironment of cysteine residues was done for estimation of dielectric constant of protein interior (Haque et al., 2000). Similarly, Fluorescence spectroscopy studies were performed to characterize the microenvironment around tryptophan residue (trp144) of the C-module of tyrosyl – tRNA synthase as consequences of temperature persuaded conformational modifications in the C-module 17 (Kordysh and Kornelyuk, 2006). Experimental determination of microenvironment for functionally active proteins is limited to only a few amino acids exposed on to the surface of the proteins having fluorescence properties (Lakowicz, 1983).

### **1.17) Limitations of experimental techniques to calculate microenvironment**

The most extensively used experimental method to calculate microenvironment is fluorescence spectroscopy. Fluorescence quenching has been used to calculate the surrounding microenvironment for few proteins (Haque et al., 2000; Kordysh and Kornelyuk, 2006). Although fluorescence spectroscopy can only be used for few residues due to the constraints of effective quenchers (Bandyopadhyay and Mehler, 2008). Due to such experimental constraints, computational methods for microenvironment computation are inevitable.

### **1.18) Literature report on computational techniques to quantify microenvironment**

Theoretical approaches to compute microenvironment was based on atoms (Dinner, 2000) and a group of atoms (Suzuki and Kudo, 1990). The most recent attempt to quantify microenvironment was based on the Rekker's fragmental constant and extended to protein structures (Bandyopadhyay and Mehler, 2008). Microenvironment has been characterized using three measures (i) Hydrophilicity or Hydrophobicity (Hpy) of residues or their fragments around a particular titrable group, as described in equation 1:

$$Hpy_A = \sum_a^{N_A} \sum_b d_b RFHC_b (r_{ab} \leq 4.25A^0) (B \neq A) \dots \dots \dots \text{(Equation 1)}$$

RFHC<sub>b</sub> is the contribution of atom b to the fragment's fragmental hydrophobic constant according to Rekker's fragmental constants (RF and Nauta, 1977). N<sub>A</sub> is the number of atoms. Moreover, d<sub>b</sub> will have a value equal to 1 if atom b has not been counted and zero if atom b has been counted. Furthermore, "a" and "b" are the number of atoms in the functional groups A and B respectively. r<sub>ab</sub> is the distance cut-off according to the definition of the microenvironment. (ii) Total Hydrophobicity Index (THpy<sub>A</sub>), that can be stated as the contribution from both solvent and protein depending on the part buried inside the protein and exposed to the solvent and is explicitly explained in equation 2:

$$THpy_A = BF_A Hpy_A + (1 - BF_A) Hpy_A^0 \dots \dots \dots \text{(Equation 2)}$$

BF<sub>A</sub> is the value of buried fraction (A is the functional group). Buried Fraction is the portion of the functional group's solvent available surface area that is buried inside the protein. Buried fraction (BF<sub>A</sub>) was calculated using GEPOL algorithm (Pascual-ahuir et al., 1994). As explained earlier, GEPOL quantifies the Hpy<sup>0</sup><sub>A</sub> is the value of Hydrophobicity Index (Hpy) in the solvent. (iii) The hydrophobicity of a particular functional group's microenvironment relative to what when the group is completely dissolved in water (rHpy). rHpy can be computed mathematically from equation 3:

$$rHpy_A = THpy_A / Hpy_A^0 \dots \dots \dots \text{(Equation 3)}$$

rHpy is the Quantitative property descriptor for microenvironment contribution. Using the rHpy values and the solvent accessibility, the partitioning of amino acid residues in different

microenvironments was performed for the development of a new hydrophobicity scale exclusively for protein structures to characterize the different microenvironments in which the amino acid side chains are found. rHpy represents the microenvironment around an amino side chain. With the increase in rHpy value, the hydrophobicity of the microenvironment decreases (Mehler and Guarnieri, 1999). The hydrophobic scale is used for calculating the free energy between two different states of a particular functional group (Eisenberg and McLachlan, 1986). Microenvironment computation and determination is used for free energy calculation for all amino acids, which leads to the formation of amino acid clusters and involves the development of network models for establishing the statistics among the test amino acid groups. In such a statistical attempt, 1654 proteins were used to understand the amino acid interactions in different environments to evaluate the hydrophobicity, based on amino acid propensities. Three environments for the secondary structures of all amino acids in the protein dataset were formulated for characterization of the functionality of amino acids in a particular environment. The authors have shown that contacts between the amino acids depend on the hydrophobicity of their surrounding environment (Jha et al., 2010). Another attempt involving the clustering of different microenvironments of cysteine residues was done to identify functional sites and also to predict the same for some of the candidate motifs (Wu et al., 2010). To understand microenvironment around any titrable amino acid residue, pKa was calculated for WT SNase mutant proteins to show that due to variation in the microenvironment wrapping a titrable amino acid residue, there is a huge deviation in the calculated pKa values (Shan and Mehler, 2011). In the most recent attempt, interfacial residues (protein-protein interface residues) were studied by developing the novel networks, microenvironment models and features to understand the role of microenvironment around these hot spot residues by determining the probable propensities for amino acid residues to be part of microenvironment around these hot spot residues (Ye et al., 2014). Similarly, the thermodynamics including the conformational free energy and conformational entropy of the interfacial residues was computed for the reaction involving the enzyme Nuclease A (NucA) and NuiA protein to demonstrate the dominance of interfacial residues in binding and stabilizing the

complex (Das et al., 2014). The interfacial residues were determined for 66 non-homologous proteins based on the surface analysis of the two interacting proteins and the comparison between the favorable energy between the unbound and the bound states using a method called as Optimal docking area (Fernandez-Recio et al., 2005). Similarly, in an attempt, 129 pairs of non-homologous complex-forming proteins were analyzed using neural networks to show that the propensity of hydrophobic amino acids is more than the charged amino acids except for arginine (Zhou and Shan, 2001). In another study, molecular dynamics approach was applied to SCOWLP (Teyra et al., 2006) database to show that the stability of the protein – protein complex increases if the interfacial residues have less solvent accessibility, which is reasonable as they are mostly hydrophobic residues (Samsonov et al., 2008).

### **1.19) Gaps in existing research**

Despite the acknowledged prominence of microenvironment role in deciphering the structure and function of protein structures, there is not enough data to support and elucidate the exact role of protein microenvironment on different functions of same amino acid. This is because, i) lesser number of protein structures in the previously studied datasets (Ponnuswamy et al., 1980; Kordysh and Kornelyuk, 2006), ii) not so well defined quantification of protein microenvironment, iii) lack of exhaustive calculation of protein micro-environments based on PDB structures and iv) correlation of amino acid functions with their embedded protein microenvironment

The microenvironment is directly related to the function of the protein (Chen and Bahar, 2004). Based on the surrounding microenvironment, amino acid structure and function modulation has also been studied. For example, identification of novel motifs around cysteines (Wu et al., 2010), and disulfide bridged cystines (Bagley and Altman, 1995). Similarly, variation in microenvironment modulates the pKa of a functional group (Mehler and Guarnieri, 1999; Mehler et al., 2002). According to Henderson-Hasselbalch equation, modulation in pKa value affects the protonation state of the amino acid residues (Equation 4) (Nielsen, 2007).

$$\text{pH} = \text{pKa} + ([\text{A}^-]/[\text{HA}]) \dots \dots \dots \text{Equation 4 (Henderson-Hasselbalch equation)}$$

Here, pH and pKa represents the pH of the solution and pKa of the functional group. HA and A<sup>-</sup> are the conjugate acid and base. Therefore, it is quite relevant to study how the change in surrounding microenvironment affects the side chain protonation state of amino acid by modulating the pKa shift.

It is known that polar solvents, like water ( $\epsilon=80$ ), have high dielectric constant values and they facilitate deprotonation of certain amino acid side chains, like aspartic acid etc. However, the effect of varying dielectric constant values on titrable amino acid side chains are not studied yet.

There have been many attempts to quantify protein dielectric medium (Gilson and Honig, 1986; Haque et al., 2000b; Pitera et al., 2001; Kukic et al., 2013). But, how the amino acid side chain protonation state varies with dielectric medium is unknown. There are no reports on understanding the dielectric nature of different protein microenvironments.

The amino acids used and the rationale to study the particular amino acid is discussed in the following section.

### **1.20) Selection of amino acids for studying applications of microenvironment**

It is already known that the cysteine residues are mostly found in the catalytic region after histidine. The pKa of the thiol group of cysteine is typically close to physiological pH (Harris and Turner, 2002); therefore, the ionization state of cysteine is highly sensitive to small changes within the local protein environment. It has been earlier proven that cysteine is a potent candidate for enzymatic turnover reactions (Harris and Turner, 2002; Gutteridge and Thornton, 2005). Moreover, apart from the functional attributes, a special property of thiol group containing a cysteine amino acid is the formation of disulfide bridged cystine residues. The cystine residues are involved in the structural stability of proteins (Voet et al., 2008). Therefore, the reduced and the oxidized forms of the cysteine amino acid are good model systems for analyzing the modulation of structure and function of amino acids based surrounding microenvironment. Although, cysteine residues were observed to be extensively used for such studies (Bagley and Altman, 1995; Chen and Bahar, 2004), yet

characterization of all types of microenvironment regions around cystine and cysteine residues are not done.

Similarly, aspartic acid being a titrable residue is a good candidate to monitor the variation in the protonation state of the side chain with its protein microenvironment. Aspartic acid is also found in many catalytic pockets along with histidine and cysteine (Gutteridge and Thornton, 2005), that makes it even more important for studying the variation in side chain protonation states with respect to its protein micro-environments. It is already known that dielectric medium of the protein hydrophobic core is much lower than that of the aqueous medium (Harris and Turner, 2002). Correlation of protein dielectric medium and aspartic acid protonation states are not being deciphered yet.

### **1.21) Objectives**

Based on the above-mentioned gaps in the existing research, the objectives of this study have been designed, as given below.

- 1) To study the modulating effect of microenvironment on structure and function of disulfide bridged cystine amino acid residues.
- 2) To annotate and classify the different functions of cysteine amino acids based on their surrounding microenvironments
- 3) To study the dielectric behavior of different protein microenvironments; by studying the aspartic acid side chain carboxylic “O-H” bond strength in different solvents and protein microenvironments.

In this work I have exploited previously developed microenvironment software (Bandyopadhyay and Mehler, 2008) to compute protein microenvironments around, cystine, cysteine, and aspartic acid residues. Microenvironments were computed based on PDB crystal structures. Microenvironments are correlated with respective functions of amino acids to understand the modulating effect of microenvironment on amino acid structure and functions.



## 1.22) Significance of this work

This work shows the application of microenvironment in modulating the structure and function of few amino acids like disulfide bridged cystine, thiol group containing cysteine and aspartic acid residues.

The already available microenvironment database for cystine and cysteine residues (Bandyopadhyay and Mehler, 2008) was used as the respective datasets. These datasets represent the amino acid in terms of buried fraction and rHpy. Therefore, each amino acid microenvironment value represents a single cluster. So, the clustering method used for identification of important functionalities was Agglomerative Hierarchical Clustering (AHC). AHC clusters by calculating the pairwise distance among the observations and then cluster them based on the similarity matrices. This procedure is iteratively applied to till the final observation to cluster the full dataset. The microenvironment for both the amino acids was clustered according to Agglomerative Hierarchical Clustering into individual clusters according to the similarity in the respective microenvironment for the amino acids in the protein structures. The Secondary structure analysis was performed through DSSP program (Joosten et al., 2011) using Kabsch and Sander algorithm (Kabsch and Sander, 1983). Calculation of cystine geometry; computing the five cystine side chain dihedral angles and the disulfide bond length for all the proteins (Katz and Kossiakoff, 1986). The specific functional roles were retrieved from the databases (Laskowski, 2001; Furnham et al., 2014) and literature mining using PERL and CSH scripts. Based on the secondary structure classification and functional role of each cysteine and cystine amino acid, the microenvironment based categorization of the structure and function of cysteine and cystine amino acids was performed. The conservation of cystine amino acids and their respective protein microenvironments in the thioredoxin protein family was established using both structural alignment (Maiti et al., 2004; Braberg et al., 2012) and local sequence alignment tools (Papadopoulos and Agarwala, 2007).

The microenvironment based modulation in cystine geometry was studied by calculating the cystine bond strength in terms of bond order using GAMES-US software (Schmidt et al., 1993). The

cystines were chosen in which the two half-cystine are found in different types of microenvironment regions as found by agglomerative hierarchical clustering. The results obtained were compared with cystine bond strength when present in two solvents; water and Dimethyl Sulfoxide (DMSO). These input structures were generated by performing Molecular Dynamics Simulations of the systems with cystine molecule dissolved in water and DMSO respectively.

The functional annotation of cysteine residues in unknown protein was performed based on three parameters: (i) Secondary structure energy function (Maiorov and Abagyan, 1998), (ii) microenvironment based energy function (energy associated with rHpy) (iii) Enzyme class.

The dielectric effect of different microenvironment was studied using the side chain carboxylic “O-H” group of the aspartic acid. A free aspartic acid was subjected to quantum calculations using 6-31G\*\* basis set with the diffusion function on the oxygen atom attached to the hydrogen in the side chain of aspartic acid using GAMESS-US software. The change in bond order was studied in 16 different implicit solvents with varying dielectric constant values ranging from 1 (vacuum) to 80.4 (water) (Lide, 2004).

## CHAPTER 2

### Categorization of cysteine functions based on protein microenvironments

#### 2.1) Introduction

Proteins are the main functional units in most of the biochemical reactions. Proteins that act as catalysts in biochemical reactions are known as enzymes. These enzymatic reactions play a significant role in metabolism and signaling in a living system (Howland, 1990; Wood, 1996; Lehninger et al., 2008). An enzyme binds to the substrate at the catalytic site; at the end of the enzymatic reaction, a product is released from the catalytic site. Residues involved at the catalytic site are assisted by other residues, termed as active residues (Howland, 1990; Alberts et al., 2002). These active residues maintain structural integrity and protonation state of catalytic site residues. The region containing catalytic and active site residues is known as active site pocket. According to induced fit theory, the interaction of the protein with the substrate molecule induces a structural reorganization in protein to allow fitting of the substrate molecule to the protein active site pocket to facilitate catalytic reaction (Koshland, 1995). This structural reorganization occurs by modulating the protonation states of the active site residues. The protonation state of a titrable amino acid is governed by the pKa of the amino acid side chain and the pH of the protein microenvironment around the amino acid. In general, aspartic acid and glutamic acid exists as nucleophiles in proteins due to their low intrinsic side chain pKa values (Harris and Turner, 2002). Surrounding pH plays a crucial role along with pKa of the side chain for an amino acid to act as an electrophile or a nucleophile. Amino acids with pKa values close to 7 can easily switch between their protonation states and facilitate enzymatic reactions. Therefore, histidine and cysteine residues are mostly found as the catalytic or active site residues in most of the proteins (Bartlett et al., 2002; Gutteridge and Thornton, 2005).

The side chain thiol group containing cysteine residues performs wide range of functionalities- like binding to metals, involved in catalysis, structural stability by undergoing thiol-disulfide exchange

(Miseta and Csutora, 2000; Jacob et al., 2003). The intrinsic value of pKa of side chain thiol (-SH) group of cysteine amino acid residues in proteins and peptides is 9.1 (Harris and Turner, 2002). In different proteins, the pKa of side chain thiol group is perturbed to lower values of 7.5 in thioredoxin, and 5.5, 4.5, and 3.5 for glutaredoxin, protein disulfide isomerase, and DsbA, proteins respectively (Harris and Turner, 2002). The alteration in protonation of side chain thiol group is due to change in protein conformation, thus participate in various biochemical reactions within protein, for example, disulfide bond formation, metal binding, catalytic reactions and other redox reactions (Miseta and Csutora, 2000; Giles et al., 2003; Jacob et al., 2003; Jacob et al., 2006; Nagahara, 2014; Akabas, 2015). Wide functional range of cysteine residue is due to multiple oxidation states of sulfur atom in the side chain thiol group, ranging from + 6 to -2 (Jones, 1949; Pauling, 1988; Woo et al., 2005). Oxidation state of sulfur in disulfide bond is -1 and that in thiol group (-SH), is -2. Reactive thiol group of cysteine participates in many cellular processes like preventing oxidative damage (Kiley and Storz, 2004), acting as switches for oxidative stress (Bhandary et al., 2012), stabilization, metabolism and signaling of proteins like actin, cJun etc. (Martínez-Ruiz and Lamas, 2007; Tew, 2007; Townsend, 2007) This property of reactive thiol in cysteine is exploited to treat many cancer related disorders and signaling disorders (Finkel, 2003; Rhee et al., 2005; Lei et al., 2008). For example, thiol group of cysteine acts as redox-switches to detect the elevated level of metal ions like zinc, copper, iron, nickel (mainly the d-block elements, those having multiple oxidation states) in cells (cysteines are mostly found in the metalloproteins) (Haigi, 2013). One of the major redox activity of thiol group in cysteines is thiol/disulfide exchange in protein structures. The standard reduction potential for thiol/disulfide redox couple vary from -0.21 volt to -0.34 volt (Jocelyn, 1967). The redox potential in enzymatic proteins depends on the enzyme class and region around the reactive thiol (-SH) group. The reduction or oxidation of thiol group is dependent on the effect of various alterations in the local chemical environment around the thiol group-containing cysteine amino acids (Wu et al., 2011). This thiol to disulfide exchange is important for structure stabilization of various proteins (Wilkinson and Gilbert, 2004), regulatory redox process homeostasis (Depuydt et al., 2011), cellular reduction and oxidation processes

(Hwang et al., 1992; Fass, 2012; Messens and Collet, 2013; Nagy, 2013). The thiol form of cysteine is preferred in the reducing cellular environment, whereas, the disulfide form is generally found in the oxidizing environment of extracellular region. For example, integral membrane proteins present in the extracellular region generally contain high number of disulfide bonds and (van Geest and Lolkema, 2000), (Feige and Hendershot, 2011), (Winther and Thorpe, 2014) so also the hair keratins (Wilson and Tobin, 2010). Important to note that the above-mentioned variety of chemical processes occur within the protein structure, cellular or extra-cellular region; where the scope of change in the reaction condition is mainly limited to the changes in pH (within a certain range). The change of pH within the protein structure occur via the heterogeneous composition of different types of amino acids. It has been shown earlier that protein local environment has control over the protonation – deprotonation of the thiol group of cysteine residue in proteins (Jones et al., 1975), (Houk et al., 1987). It has also been shown that upward pKa shifts of acidic amino acids were induced due to the embedded hydrophobic microenvironment (Mehler and Guarnieri, 1999). Protein microenvironment was described as the spatial arrangement of atoms around the functional group of an amino acid (Bandyopadhyay and Mehler, 2008)

In this work, microenvironments around the thiol group of cysteines were computed, in high-resolution protein crystal structures. Four different types of cysteine (thiol) functions were reported: metal-binding, catalytic, active and redox. Relationship between cysteine functions and protein microenvironments is explored in current chapter. To find and establish the above-mentioned relationship, the protein microenvironments around cysteines were clustered using agglomerative hierarchical clustering that produces four different microenvironment clusters, namely, buried-hydrophobic, buried-hydrophilic, intermediate and exposed-hydrophilic. Role of microenvironment on thiol function embedded in several protein structures was analyzed and discussed.

## **2.2) Methodology**

The cysteine microenvironment database was derived from high resolution protein crystal structures with resolution of better than 2Å and sequence similarity less than 25% (Hobohm and Sander,

1994), (Bandyopadhyay and Mehler, 2008). There were 436 high resolution protein crystal structures in the microenvironment database containing 2070 cysteine residues. These cysteines were distributed in the microenvironment space. The microenvironment space is the two-dimensional region containing cysteine amino acid residues. This space was divided into different clusters using agglomerative hierarchical clustering.

### **2.2.1) Agglomerative Hierarchical Clustering of cysteine microenvironment database to cluster the similar individual observations into dissimilar cysteine microenvironment clusters**

The cysteine microenvironment space was formulated, based on two parameters (i) Buried fraction and (ii) rHpy (Bandyopadhyay and Mehler, 2008). The buried fraction is the extent of the side chain thiol group buried inside the protein. Buried fraction (BF) ranges from 0 to 1 and with increase in buried fraction value the extent of the functional group being buried inside the core of the protein increases. Cysteines with buried fraction value of 0 are completely exposed to the solvent. Quantitative Property Descriptor (QPD) rHpy, quantifies the local microenvironment values of the functional group (Bandyopadhyay and Mehler, 2008). The increase in rHpy is directly proportional to the hydrophilicity of the microenvironment. The upper limit of rHpy value is 1; that denotes the cysteine is completely exposed to water microenvironment. There is no lower limit of rHpy, however, the lowest value of rHpy noted in our cysteine database was -0.4.

The 2070 cysteine residues in the dataset represent their respective microenvironment, in terms of a point denoted by buried fraction and rHpy values, in the entire microenvironment space. Initially, each point on the microenvironment space represents an isolated cluster. To identify the function of microenvironment regions, the similar microenvironments around the cysteine residues must be clustered together to generate dissimilar individual microenvironment clusters. To achieve this, similar clusters based on the buried fraction and rHpy were clustered together. Based on the distance proximity the similar clusters were combined to generate larger dissimilar clusters (Tryon, R. C. and Bailey, 1973). Therefore; agglomerative hierarchical clustering method was adopted to find the similar microenvironment clusters (Murtagh and Legendre, 2014) though ward's method

(Ward, 1963). The agglomerative clustering was performed using XLSTAT software (Addinsoft, 2014).

### 2.2.2) Secondary structures of cysteines in different microenvironment clusters.

The secondary structures of all the 436 proteins were calculated using Dictionary of protein secondary structure (DSSP) algorithm (Kabsch and Sander, 1983), (Joosten et al., 2011). DSSP calculates the protein secondary structure. The calculation is performed by the 3D- coordinates of the atoms in the protein. This leads to the calculation of hydrogen bond energy and the hydrogen positions. Based on the hydrogen atom positions the structure of the residue is defined.

The calculation of secondary structure using the DSSP program was automated via an in-house custom PERL script. In-house PERL script and C-shell script were also used to encrypt the secondary structures of cysteines present in different cysteine microenvironment clusters.

### 2.2.3) Identification of enzymatic proteins that belong to different enzyme classes in the dataset

The enzymatic proteins from six enzyme classes (International Union of Biochemistry and Molecular Biology (IUBMB) (Bairoch, 2000)) present in the cysteine microenvironment dataset were identified from PDB header files and categorized into different microenvironment clusters (in-house PERL script were used) Total 183 enzymes were obtained (Table 2.1).

Table 2.1: Total number of proteins that belong to the six enzyme classes in the cysteine microenvironment dataset of 436 proteins

Enzyme class	Total number of enzymatic proteins	PDB ID
Oxidoreductases + Electron Transport Proteins*	57	1CP2,1D4O,1DCS,1DQI,1HD2,1HXX,1JF8,1JFB,1JTV,1JUB,1K3I,1KV7,1LJ8,1LU4,1MLD,1N4W,1N8K,1NXU,1O2D,1O7N,1OAA,1ODM,1P4C,1PBY,1QKS,1QX4,

		1R4U,1SU8,1SYY,1T2D,1UBK,1US0,1USC,1USP,1UZ B,1VJU,1VKN,1VL7,1VYR,1VZI,1CIF,1CZP,1DJ7,1H3 2,1HLQ,1IQZ,1ISU,1IUA,1M2D,1M70,1O8X,1OQQ,1P LC,1R26,1SFD,2C8S,2FDN
Hydrolases	57	1AKO,1AY7,1D3V,1D5T,1DBX,1E1H,1E7L,1ES5,1G6I ,1GKL,1GPP,1GXU,1H2W,1H4G,1H9H,1HDK,1HDO,1 HT6,1HYO,1I0D,1JMK,1K5C,1K6K,1KA1,1KIC,1KWF ,1LO71NOF,1NYC,1O4Y,1O5F,1OC7,1OGO,1P6O,1PJ X,1PMH,1PYO,1Q7L,1QTN,1QTW,1QWY,1QXY,1QZ M,1S5U,1S95,1SC3,1UG6,1UWC,1UWK,1UXO,1VKH, 1VKP,1W2Y,1YAC,2HRV,2PTH,3SIL
Transferases	36	1AYL,1B5E,1BX4,1CXQ,1E5K,1E6B,1FG7,1FSG,1I12, 1ID0,1JYK,1KJQ,1KWA,1LD8,1M15,1MML,1MOQ,1N KI,1NN5,1O26,1O50,1O9G,1OOY,1P5Z,1PTQ,1QF8,1 QST,1RDQ,1RKU,1RO7,1RYQ,3HPD,1VHW,1VKB,2T PS,3CLA,
Lyases	21	1B93,1DCI,1DOS,1FX4,1GKM,1GVF,1IDP,1KD0,1LC 7,1LK9,1LUG,1N13,1N7H,1NP7,1OJR,1PQH,1QJ4,1Q OP,1TLU,1UGP,1UUY
Ligases	5	1BYI,1FS1,1JAT,1K92,1PFV
Isomerases	7	1GYX,1HZT,1K4I,1OH0,1PIN,1TQJ,1USL

#### 2.2.4) Curation of catalytic cysteines from literature and secondary databases

Catalytic cysteines were defined as the residues which directly bind to the substrate (Porter et al., 2004). Catalytic cysteines were extracted from Catalytic Site Atlas (CSA) database (Furnham et al., 2014).



The catalytic cysteines were extracted by using the database crawler that identifies and extracts the catalytic site cysteine in proteins (in-house PERL script was used). These catalytic cysteines are defined in the Catalytic Site Atlas database through literature and experimentation results from various sources (Furnham et al., 2014).

#### **2.2.4.1) Curation of active site residues from the literature and secondary databases**

Active cysteines were defined as the residues those were not directly involved in catalysis but constitute the active site pocket (de Beer et al., 2014). These active cysteines were extracted from the PDB header files through a custom PERL script. The header file contains such active site residues that might also involve the ligand binding cysteine amino acid residues in the “SITE” region of the PDB files of the proteins in the dataset.

#### **2.2.5) Identification of cysteines binding to the metal ions in the proteins in the dataset through literature based text mining and database search**

Metal-binding cysteines were defined as the residues where the sulfur atom of the thiol group interacts with the metal ion, either directly or via a ligand. Cysteine binds to metal ions like iron, copper, zinc and mercury. 170 metal-binding cysteines were identified in 47 proteins using in-house PERL script (Table 2.2). 5Å distance cutoff was used between the sulfur atom of thiol group and the respective metal ion to detect the metal binding cysteine. Selection of 5Å ensures the inclusion of direct and ligand-mediated metal-binding to the thiol group of cysteine.

Table 2.2: List of proteins containing metal binding cysteines bound to a particular metal ion directly, or through a ligand molecule (in 5-angstrom vicinity of the ligand, bound to any atom o the ligand, mostly a thioether bond).

Metal ion	Number of proteins	PDB ID
Iron	21	1BBH,1C75,1CIF,1CTJ,1DQI,1E29,1GU2,1H32,1I8O,1J0P,1JFB,1JNI,1M1Q,1M70, 1OS6,1PBY,1QKS,1RB9,1UBK,1VZI,2C8S

Zinc	19	1E1H,1E7L,1HXR,1J98,1LD8,1N8K,1NZJ,1OQJ,1P6O,1PFV,1PTQ, 1Q08,1QF8, 1QTW,1RYQ,1UW1,1VFX,2HRV
Copper	3	1KV7, 1PLC, 1SFD
Mercury	4	1CC8,1HDK,1LUG,1PVM

### **2.2.6) Structural and sequence alignment of the iron binding and zinc binding proteins to test the conservation of microenvironment in functionally similar proteins like cytochromes**

The C-x-x-C-H motif containing cytochrome protein in the dataset and the zinc binding proteins containing C-x-x-C motif regions were structurally aligned using Salign Server (Braberg et al., 2012) and the RMSD's of the aligned regions were calculated using Chimera software (Meng et al., 2006). The global multiple sequence alignment of these proteins was performed using t-coffee MSA tool (Li et al., 2015). The local multiple alignments were performed to show the conservation of the motif region using COBALT (Papadopoulos and Agarwala, 2007) online tool.

## **2.3) Results and Discussion**

### **2.3.1) Statistics of cysteine microenvironment clusters based on agglomerative hierarchical clustering**

Agglomerative hierarchical clustering employed on 2070 cysteines present in current dataset produces four clusters, buried-hydrophobic, buried-hydrophilic, intermediate and exposed-hydrophilic, based on buried fraction and rHpy values (Figure 2.1).

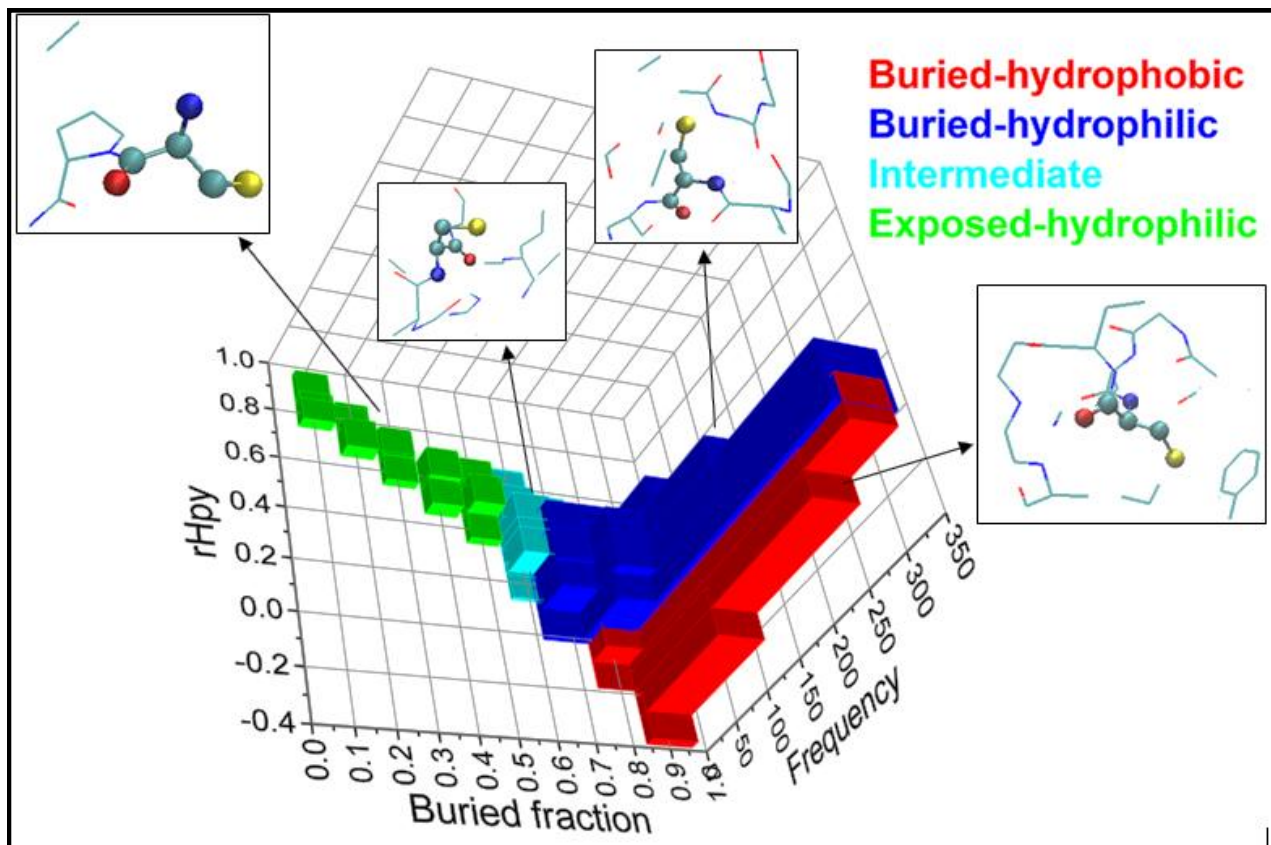


Figure 2.1: The agglomerative hierarchical clustering of the 2070 cysteines in 436 high-resolution protein crystal structures has resulted into four cysteine microenvironment clusters represented through individual colors. The x-axis represents buried fraction ranging from 0 to 1. rHpy is at y-axis and ranges from -0.4 to 1. Decrease in rHpy value indicates an increase in the hydrophobicity of the surrounding protein microenvironment around the thiol group-containing cysteine residues. The number of cysteines per cluster is represented by frequency on the z-axis. This graph was generated using Origin [Origin (OriginLab, Northampton, MA)].

Cysteines prefer buried microenvironment clusters compared to solvent exposed microenvironment clusters (1725 out of 2070 were present in buried clusters) (Table 2.3), this is in accordance with the previous observation (Bandyopadhyay and Mehler, 2008).

Table 2.3: Statistics of cysteine microenvironment clusters. The clusters are arranged according to descending order of hydrophobicity, measured by cluster center values. Decrease in rHpy value indicates an increase in the hydrophobicity of the surrounding protein microenvironment around the thiol group-containing cysteine residues.

Cluster type	Buried fraction	rHpy	Average distance to centroid (Å)	Within class variance	No. of cysteines in each cluster
Buried-hydrophobic	0.983	-0.015	0.073	0.007	798
Buried-hydrophilic	0.949	0.236	0.106	0.003	927
Intermediate	0.711	0.375	0.085	0.009	199
Exposed-hydrophilic	0.409	0.659	0.172	0.042	146

Here, clustering of various cysteine microenvironments and similarity of cysteine functions within a cluster are shown. The Buried-hydrophobic cluster and the Buried-hydrophilic cluster share similar buried fraction, however, they differ in terms of microenvironment (rHpy), secondary structure and function of the cysteine residues. Similarly, intermediate and exposed-hydrophilic cluster share similar buried fraction but differ in terms of rHpy, cysteine secondary structure, and functions. Structural and functional dissimilarities across different clusters are discussed in the following sections.

### 2.3.2) Categorization of cysteine functions based on microenvironment (rHpy value)

To identify the relationship between cysteine functions and the surrounding environment, the interactions of the cysteine molecule with the surrounding environment were studied and accordingly different cysteine functions were defined (Figure 2.2).

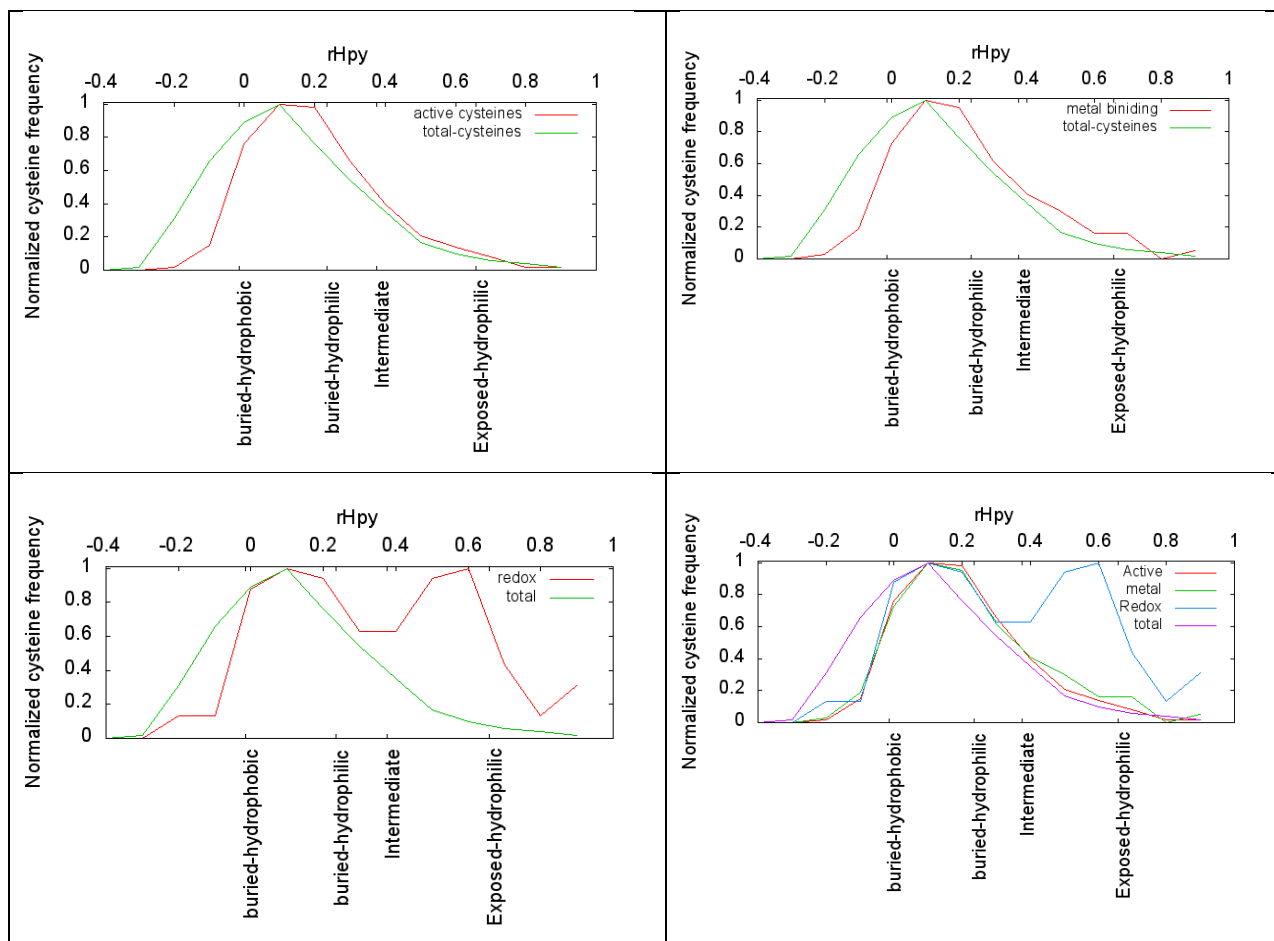


Figure 2.2: Comparison of the extent of active cysteines (A), metal-binding cysteines (B), redox cysteines (C) and all the three functional cysteines (D) with the respect to the full cysteine residue distribution in terms of microenvironment property descriptor (rHpy). The normalization was done by dividing the functional cysteines on the basis of rHpy values into small bins of size 0.1, ranging from -0.4 to 0.9 (rHpy values). Then, using the highest of cysteine population within these smaller bins, normalization was performed.

The metal binding cysteines are more prominently found in intermediate cluster whereas the redox cysteines are found in the exposed-hydrophilic region. The active cysteines were widespread into the full microenvironment space, but prefer buried-hydrophilic more as compared to other microenvironment regions. Due to the presence of an oxidizing environment in exposed hydrophilic microenvironment region the tendency for the thiol group-containing cysteine residues to undergo oxidation increases. Moreover, the active cysteines (not catalytic cysteines) are important residues

in protein functional mechanism, hence are buried into the protein core but still prefer a hydrophilic region to perform interaction with the substrate and surrounding residues more easily.

### 2.3.3) Distribution of different cysteine functions and secondary structures across four microenvironment clusters

Cysteine functions vary across the microenvironment clusters (Table 2.4). The frequency of functional cysteines increases with increase in the hydrophilicity of the microenvironment clusters (Figure 2.3).

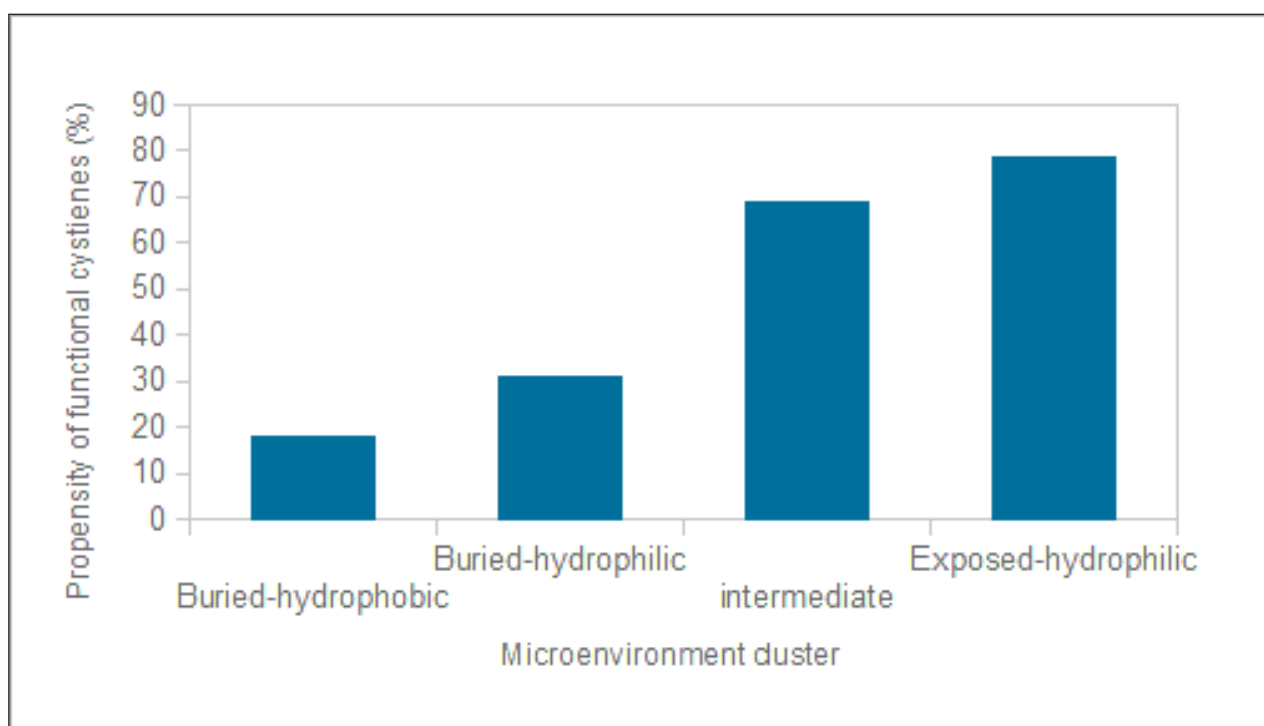


Figure 2.3: Propensities of functional cysteines in different microenvironment clusters.

Main cysteine functions are catalytic, active, metal binding and redox. Here onwards in the following section, functional cysteines will only include active, metal binding and redox cysteines. Catalytic cysteines are left out due to insufficient data (only 25 catalytic cysteines are reported; Table 2.4).

Table 2.4: Different functional cysteines in different microenvironment clusters. Number of functional cysteines in each cluster are shown in the first column within the parenthesis. As the absolute numbers depend on the size of the cluster, normalized values with respect to the cluster size (expressed in terms of percentage) are given within parenthesis of following columns representing the individual microenvironment cluster.

Cysteine functions	Buried-hydrophobic (798)	Buried-hydrophilic (927)	Intermediate (199)	Exposed-hydrophilic (146)
Catalytic (25)	4 (0.5%)	14 (1.5%)	5 (2.5%)	2 (1.4%)
Active (382)	86 (11.0%)	175 (19.0%)	79 (39.7%)	41 (28.1%)
Metal binding (177)	41 (5.1%)	64 (6.9%)	46 (23.1)	26 (17.8%)
Redox (114)	18 (2.25%)	40 (4.3%)	10 (5%)	46 (31.5%)

These observations presumably indicate that equivalent contributions of protein and solvent towards cysteine microenvironment play an important role in deciding reactivity of cysteine thiol. To note, active site cysteine assists in catalysis but does not directly participate in catalysis. It was reported that if two nearby cysteine residues are found in the hydrophilic microenvironment, the tendency of disulfide bond formation increases (Bagley and Altman, 1995).

This effect is more prominent in redox cysteines. Table 2.4 shows that maximum normalized frequency of redox cysteines was found in the exposed-hydrophilic cluster that has steeply risen from buried-hydrophobic to the exposed-hydrophilic cluster. To investigate the role of the hydrophilic microenvironment in the oxidation process of a thiol group, functions of all the 146 cysteines in the exposed-hydrophilic cluster were curated from the literature. Results showed that 46 out of 146 (~34%) cysteines were redox cysteines. Cysteines upon reaction with substrate were involved in disulfide formation (35 out of 146), thioether linkages (9 out of 146) or nucleophilic

reactions (2 out of 146) (Table 2.5). These observations illustrate the importance of hydrophilic microenvironment in functioning of redox cysteines. One specific example of redox cysteine is the C-x-x-C-H motif in cytochrome proteins where thiol groups undergo oxidation during electron transport chain in respiration and Calvin cycle in photosynthesis (Shirabe et al., 1991).

Table 2.5: Cysteines in exposed hydrophilic cluster with redox functions curated from the respective research articles

PDB	residue no.	Buried-fraction	rHpy	Cysteine function from literature (upon reaction with substrate)
1B5E	148A	0.59	0.545	Disulfide
1B5E	148B	0.592	0.596	Disulfide
1C75	35A	0.44	0.686	Disulfide
1CC8	15A	0.366	0.701	Disulfide
1CP2	94B	0.564	0.582	Disulfide
1CTJ	18A	0.501	0.523	Disulfide
1DG6	230A	0.476	0.554	Disulfide
1E29	40A	0.285	0.756	Disulfide
1G55	24A	0.153	0.868	Nucleophile
1G55	292A	0.627	0.628	Nucleophile
1GXU	7A	0.234	0.805	Disulfide
1H32	79A	0.412	0.645	Disulfide
1H32	114A	0.614	0.516	Disulfide
1H9H	128E	0.542	0.515	Disulfide
1HD2	151A	0.684	0.591	Disulfide
1HDK	29A	0.477	0.605	Disulfide
1JNI	101A	0.415	0.619	Disulfide



1K0M	89A	0.536	0.654	Disulfide
1K0M	89B	0.557	0.626	Disulfide
1LXJ	11A	0.504	0.544	Disulfide
1M1Q	15A	0.359	0.693	Disulfide
1M1Q	18A	0.558	0.507	Disulfide
1M1Q	38A	0.301	0.785	Thioether
1M1Q	61A	0.518	0.5	Thioether
1M1Q	78A	0.322	0.751	Thioether
1M70	122A	0.527	0.475	Disulfide
1M70	122D	0.486	0.525	Disulfide
1MXI	83A	0.424	0.611	Thioether
1OS6	27A	0.333	0.713	Disulfide
1OS6	30A	0.28	0.759	Disulfide
1OS6	65A	0.089	0.917	Disulfide
1PBY	14A	0.455	0.589	Thioether
1PYO	79A	0.407	0.638	Disulfide
1PYO	8C	0.058	0.943	Disulfide
1Q2H	27C	0.495	0.574	Disulfide
1Q2H	43C	0.41	0.697	Disulfide
1QGW	50C	0.463	0.601	Thioether
1QGW	50D	0.53	0.527	Thioether
1RGX	6A	0.064	0.938	Disulfide
1RGX	6B	0.249	0.753	Disulfide
1RGX	6C	0.41	0.626	Disulfide
1SU8	39A	0.413	0.607	Thioether
1UGP	189B	0.505	0.609	Disulfide

2C8S	68A	0.416	0.617	Thioether
2HRV	138A	0.056	0.952	Disulfide
2HRV	138B	0.038	0.967	Disulfide

The active (redox) cysteines in C-x-x-C-H motifs from cytochrome proteins are discussed in subsequent section.

Metal binding cysteines also prefer intermediate and exposed microenvironment over the buried ones (Table 2.4). Cysteine residues are reportedly bonded to the metal ions via thiol group in many proteins. This metal binding property of cysteine residue is facilitated by surrounding hydrophilic microenvironments around the thiol group of cysteine (Ebert and Altman, 2008). Metal binding cysteines mostly prefer solvent exposed hydrophilic region (starting from the intermediate cluster) due to presence water molecules (Frausto da Silva and Williams, 2001). The metal ion may further facilitate ionization of water that is important for catalytic reactions in many enzymatic proteins (Glusker et al., 1999).

Cysteine functions are significantly related to their respective secondary structures. For example, i) active cysteine (-S-S- functional group) in oxidoreductases prefer alpha helix for easy cleavage (Simone et al., 2006), ii) active cysteines (thiol group) in hydrolase enzyme class prefer alpha/beta fold for stability (Nardini and Dijkstra, 1999), iii) active cysteines (-SH) prefer alpha helical geometry for effective heme attachment in cytochrome proteins (Pokkuluri et al., 2004), iv) redox cysteines prefer bend secondary structures (Waring et al., 1996). Variation in number of cysteines with different secondary structure is reported for different microenvironment clusters (Table 2.6).

Table 2.6: Distribution of secondary structures of all the 2070 cysteines among four microenvironment clusters. Cluster sizes for respective microenvironment clusters are given within parenthesis.

Secondary Structures from DSSP	Buried-hydrophobic (798)	Buried-hydrophilic (927)	Intermediate (199)	Exposed-hydrophilic (146)
H = $\alpha$ -helix	252	218	63	34
B = $\beta$ -bridge	2	21	3	0
E = strand	322	280	31	14
G = 3(10)-helix	16	33	7	10
T = turns	56	65	30	25
S = bend	27	72	20	10
I = $\pi$ -helix	6	11	2	0
C = Coil	0	0	0	0

These numbers are extrinsic, that is, dependent on the cluster size. For ease of comparison across the clusters, normalized values are reported (Figure 2.4). Normalized frequencies of different secondary structures have indicated a steep decrease of the beta structure along with increasing hydrophilicity of the microenvironment clusters. For example, active cysteines, present in buried-hydrophilic cluster show slight preference towards beta sheet secondary structure (Figure 2.5). The active cysteines in intermediate and exposed-hydrophilic clusters are mainly present in turns and alpha helical geometries (Table 2.7). The normalized frequency of cysteines in turn structures doubled from buried-hydrophobic to exposed-hydrophilic cluster (Figure 2.4). Normalized cysteine frequencies of the alpha helix are more or less similar in buried-hydrophobic and exposed-hydrophilic clusters, although this value slightly varies in buried-hydrophilic and intermediate clusters. Bend structures remain almost invariant across different microenvironment clusters.

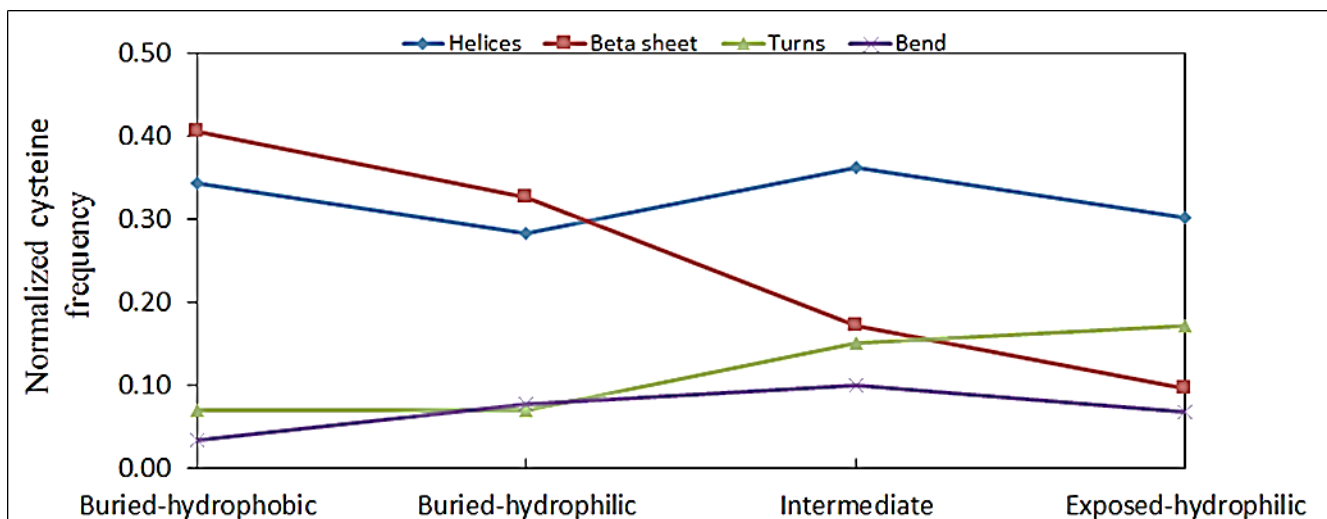


Figure 2.4: Normalized frequencies of cysteines present in different secondary structures in four cysteine microenvironment clusters. The y-axis represents the normalized frequency of cysteines present in a particular secondary structure in a cysteine microenvironment cluster with respect to the individual cluster size. Here, the helices include alpha-helical, pi-helix and 3(10) helical secondary structures. Similarly, the beta sheet includes both beta strand and beta bridge secondary structures. According to DSSP calculation, no cysteines were found in the Coil region, hence are not included in the secondary structure comparison of the cysteine microenvironment clusters.

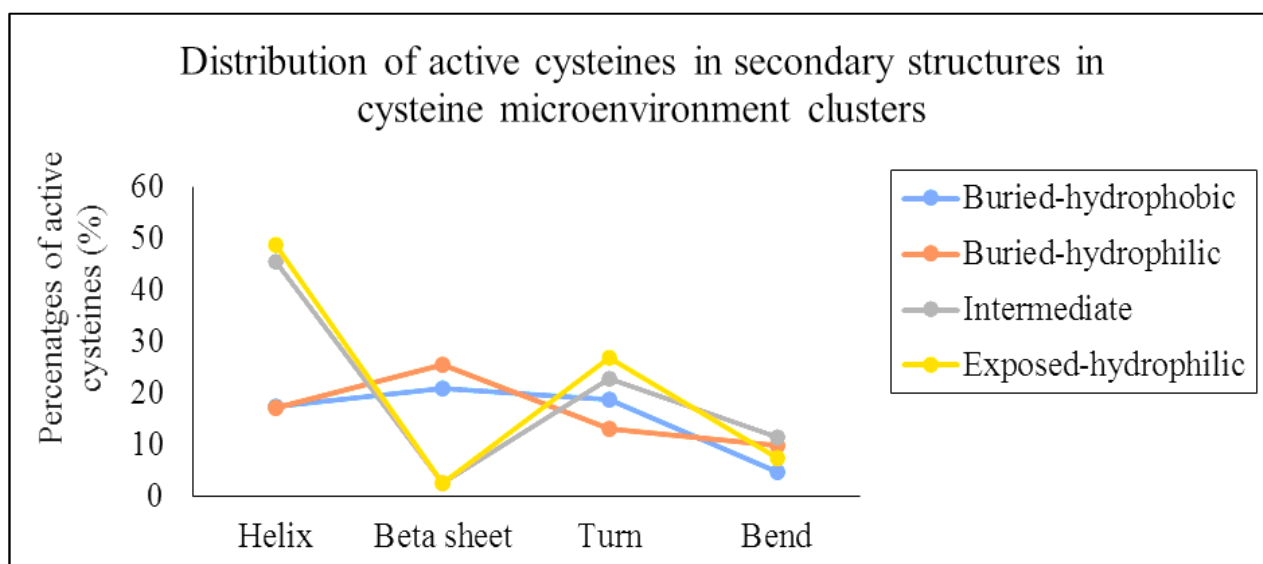


Figure 2.5: Preference of the local secondary structure by the active site cysteine residues towards the cysteine microenvironment clusters. This preference is established in terms of percentages of

active cysteine residues in different secondary structures (helix, beta sheet, turns and bends) present on the x-axis in a particular cysteine microenvironment cluster.

Table 2.7: Analysis of secondary structures of active cysteines present in different microenvironment clusters. Absolute values are shown in each column. The normalized values (with respect to the cluster size, expressed in terms of percentage) are given within the parenthesis.

Microenvironment cluster	Total number of active cysteines	Helix	Beta Sheet	Turn	Bend
Buried-hydrophobic	86	15 (17%)	18 (21%)	16 (19%)	4 (5%)
Buried-hydrophilic	176	30 (17%)	45 (26%)	23 (13%)	18 (10%)
Intermediate	79	36 (46%)	2 (3%)	17 (23%)	9 (11%)
Exposed-hydrophilic	41	20 (49%)	1 (2%)	11 (27%)	3 (7%)

As functional cysteines are mostly part of enzymes. We have performed analysis of enzyme classes in different microenvironment clusters, in the following section.

#### **2.3.4) Distribution of enzyme classes (involving cysteine) in different microenvironment clusters**

There are total 1307 cysteines in the enzymatic proteins present in the microenvironment dataset (Table 2.8). As cluster sizes differ, normalized frequencies of enzyme classes containing cysteines were compared across the microenvironment clusters (Fig 2.6). Oxidoreductase, including electron transport proteins and hydrolases, have shown significant changes in frequencies across the microenvironment clusters (Table 2.8).

Table 2.8: Distribution of cysteines in the enzyme class according to the microenvironment clusters.

A total number of cysteines in each cluster is given in parenthesis.

Enzyme classes	Buried- hydrophobic (798)	Buried- hydrophilic (927)	Intermediate (199)	Exposed- hydrophilic (146)
Oxidoreductases + Electron transport proteins	125	219	62	36
Transferases	115	122	18	18
Isomerases	20	23	7	3
Hydrolases	136	169	24	24
Ligases	12	11	5	2
Lyases	66	71	10	9
Total	474	615	126	92

Occupancy of isomerases and ligases are negligible compared to cluster sizes. Similarly, there is a slight decrease in the cysteine frequency in the lyases enzyme class while moving towards the hydrophilic microenvironment (Figure 2.6).

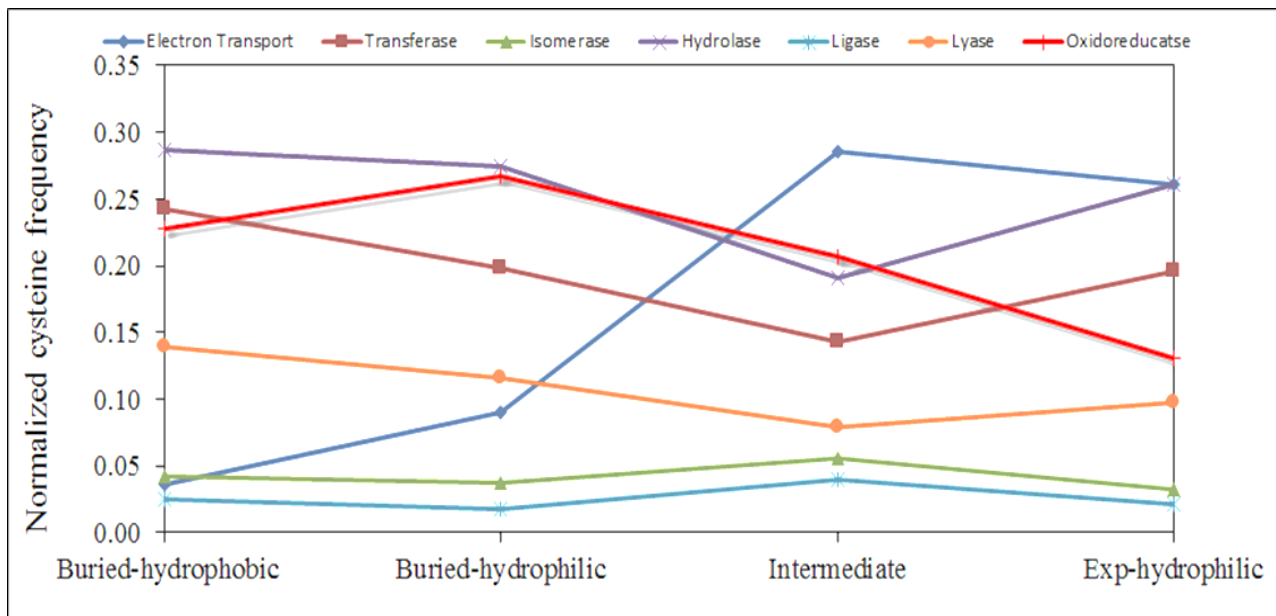


Figure 2.6: Normalized frequency of cysteines in different enzyme classes distributed across microenvironment clusters.

Normalized cysteine frequencies gradually increase from buried-hydrophobic to intermediate cluster for electron transport proteins. Whereas, normalized cysteine frequencies gradually decrease, from buried-hydrophobic to intermediate cluster for hydrolases and transferases. Normalized cysteine frequencies show a transition from intermediate to exposed-hydrophilic cluster for almost all the enzyme classes (except oxidoreductase). That could presumably indicate the larger contribution of solvent towards the microenvironment (compared to protein contribution) starting from the intermediate cluster (to the exposed-hydrophilic cluster). Average buried fraction of the intermediate cluster is 0.711; the value of 0.7 was empirically decided as the cutoff to mark the dominance of solvent contribution over protein contribution towards the microenvironment around an amino acid (Mehler et al., 2002).

Normalized cysteine frequencies indicate that different enzymes prefer different microenvironment clusters, (around cysteine residue) most probably to perform specific functions. Electron transport proteins, those involved in redox reactions, prefer intermediate and exposed hydrophilic clusters. Hydrolase enzymes, those involved in hydrolysis, prefer buried hydrophilic cluster, whereas the transferases were mostly found in buried (hydrophobic and hydrophilic) clusters. It has been shown

earlier that Chi-Class Glutathione S-Transferase proteins require a hydrophobic environment for proper folding (Pandey et al., 2015), that is also true as per our rHpy data. The Chi-Class Glutathione S-Transferase protein (PDB ID: 4NHW) was subjected to rHpy calculation. The rHpy of the eight cysteines in this protein were noted (Table 2.9). The average rHpy of these proteins was found to be 0.17.

Table 2.9: rHpy values of the eight cysteines in transferase protein (PDB ID: 4NHW).

Residue number	135	195	350	410	564	624	780	840
rHpy	0.202	0.4	0.275	0.051	0.165	0.041	0.192	0.41

It is to be noted that with an increase in rHpy value the hydrophilicity increases and the maximum limit of rHpy is 1 (Bandyopadhyay and Mehler, 2008).

To further understand the role of microenvironment on cysteine functions in different enzyme classes, we have combined cysteine microenvironment data with secondary structure and enzyme classification, in the following section.

### **2.3.5) Segregating cysteines according to microenvironment, secondary structures, and enzyme classes**

The four main functions of cysteine residues were identified as catalytic, active, metal-binding and redox cysteines (Table 2.4). The number of catalytic cysteines is too low to suggest any microenvironment role in the catalytic behavior of cysteines. Therefore, the cysteine function for a particular microenvironment cluster was only restricted to active, metal binding and redox cysteines (Table 2.10).



Table 2.10: Frequency of each enzyme class in different microenvironment clusters categorized according to cysteine functions are reported. The breakup of this number according to secondary structures are shown in parenthesis, [T stands for turn, H helix and E strand]. Number of different cysteine functions in each microenvironment cluster are shown in parenthesis. The secondary structure with the highest percentage is reported along with each enzyme class. The cysteines that were found in enzyme classes other than electron transport, oxidoreductases and hydrolases were few hence were not considered in this analysis.

<b>Buried-hydrophobic</b>	Active 86 (11.0%)	Metal-binding 41(5.14%)	Redox 18 (2.25%)
Electron Transport	25 (13T,4H, 8E) 52.0% T	4 (3T, 1E) 75.0% T	0
Oxidoreductase	20 (19T, 1E) 95.0% T	9 (8T,1H) 88.9% T	6 (6T) 100% T
Hydrolase	13 (7T, 4H, 2E) 53.8% T	9 (4T, 3H, 2E) 44.4% T	6 (3T, 2H, 1E) 50% T
<b>Buried-hydrophilic</b>	Active (175) (19.0%)	Metal-binding 64 (6.9%)	Redox 40 (4.3%)
Electron transport	54 (35T,12H,7E) 64.8% T	10 (9T,1H) 90.0%T	2 (2T) 100.0%T
Oxidoreductase	41 (19T, 8H, 14E) 46.3% T	17 (5T,2H,10E) 58.8% E	20 (11T, 3H, 6E) 55.5% T
Hydrolase	16 (11T,3H,2E) 68.7% T	7 (5T, 2E) 71.4% T	5 (1T, 3H, 1E)
<b>Intermediate</b>	Active 79 (39.7%)	Metal-binding 46 (23.12%)	Redox 10 (5%)

Electron Transport	42 (19T, 23H) 54.7% H	23 (9T, 14H) 60.9% H	2 (2T) 100% T
Oxidoreductase	15 (5T, 9H, 1E) 60.0% H	11 (3T, 8H) 72.7% H	2 (1T, 1H) 50% T
Hydrolase	4 (2T, 2H) 50% H	0	2 (2H) 100% H
<b>Exposed-hydrophilic</b>	Active 41 (28.1 %)	Metal-binding 26 (17.81%)	Redox 46 (31.5)
Electron transport	24 (7T, 17H) 70.8% H	21 (6T,15H) 71.4% H	18 (4T, 14H) 77.8% H
oxidoreductase	5 (4T, 1H) 80.0% T	52 (2T) 100.0% T	5 (5T) 100.0% T
Hydrolase	3 (3T) 100.0% T	1	7 (5T,2E) 71.4% T

### 2.3.5.1) Functional role of cysteines buried-hydrophobic cluster

The cysteines were found mostly in electron transport, oxidoreductases, and hydrolases proteins. The cysteines part of buried hydrophobic cluster are mainly found in turns and mostly were part of active site pocket. It has been observed that the cysteines in turns were more in oxidoreductases and electron transport protein as compared to hydrolases. The tendency of cysteines involved in any functional form (active, metal binding or redox) is more in oxidoreductases (Table 2.10).

### 2.3.5.2) Functional role of cysteines in buried-hydrophilic cluster

The cysteines in the buried-hydrophilic cluster have higher percentages of active, metal binding and redox cysteines, but the percentages of active cysteines in turns is lesser than those in the buried-hydrophobic cluster. Although, here also the active cysteines irrespective of the enzyme class were found mostly in turns. It was also found that active cysteines prefer hydrolases and electron transport proteins. Metal binding cysteines when part of oxidoreductases prefers beta-sheet structures. Similarly, redox cysteines when found in hydrolases prefer helical conformations (Table 2.10).

### **2.3.5.3) Functional role of cysteines in intermediate cluster**

The percentage of metal binding cysteines has increased considerably in the intermediate cluster, although active cysteines are still comparatively more within the cluster and prefer helical conformations in all three enzyme classes. The cysteines in intermediate cluster prefer mostly the helical conformation. Metal binding cysteines in intermediate cluster prefer only electron transport proteins and oxidoreductases. Redox cysteines when present in electron transport proteins and oxidoreductases prefer turns but if present in hydrolases than were found in helix (Table 2.10).

### **2.3.5.4) Functional role of cysteines in exposed-hydrophilic cluster**

The exposed-hydrophilic cluster has a clear preference towards the redox cysteines present in helical structures in case of electron transport proteins and in turns in case of oxidoreductases and hydrolases. The active cysteines and metal binding were also found in a similar distribution as for redox cysteines (Table 2.10).

Overall, it was found that the active cysteines were widespread into the cysteine microenvironment clusters. Buried-hydrophobic and buried-hydrophilic clusters prefer active cysteines in turn secondary structure irrespective of the enzyme class. Whereas, the intermediate cluster as compared to other clusters contains mostly metal binding cysteines and helical conformations. The exposed-hydrophilic cluster prefers redox cysteines in helical region of electron transport proteins and in turn structure of oxidoreductases and hydrolases (Table 2.10).

### **2.3.6) Conservation of microenvironment around cysteine residues in cytochrome proteins**

Cysteine residues present in Cytochrome proteins were of special interest because those were part of the conserved C-x-x-C-H motif (Shirabe et al., 1991). In the current study, 12 different types of cytochrome proteins were identified containing C-x-x-C-H motif. All the cysteines, part of the C-x-x-C-H motif of cytochromes, were embedded in the hydrophilic microenvironment (Table 2.11).

These cysteines are often involved in the thiol-disulfide exchange and in electron transfer (Meunier et al., 2004). The thiol-disulfide exchange was facilitated by the presence of two cysteine residues in the close vicinity along with a protonated histidine. Cytochrome proteins participate in electron transport chain in respiration and Calvin cycle in photosynthesis (Voet et al., 2008). Cysteine residues undergo disulfide bond formation by releasing protons to reduce the iron metal to react with the substrate (Cederbaum, 2015; Grabarczyk et al., 2015). The electron transfer, involving the C-x-x-C-H motif, follow the order - a) from substrate to the thiol groups of cysteines, b) from cysteine thiol to protonated histidine c) from protonated histidine to the heme group, reducing the iron from  $Fe^{3+}$  to  $Fe^{2+}$  d) from heme to the next reactant (Meunier et al., 2004). The present study has explained the role of hydrophilic microenvironment (Table 2.11) to stabilize the two nearby thiols in protonated form (avoiding disulfide formation). Here we have computed overall sequence similarity between 12 cytochrome proteins. That turned out to be 29%, based on multiple sequence alignment using T-COFFEE online tool (Li et al., 2015). The overall structural similarity between these 12 proteins was not reported by SALIGN server (Braberg et al., 2012). However, SALIGN server was able to produce structural alignment of C-x-x-C-H motif in all the 12 cytochrome proteins with a quality score of 5 (100%) (Figure 2.7). As Salign only provides the quality of the alignment in terms of scores, therefore, the root mean square deviation (RMSD) was calculated using SuperPose server (Maiti et al., 2004). The RMSD of 0.54 angstroms was calculated for the aligned motif regions of 12 cytochrome proteins. Despite low sequence similarity among all the cytochrome proteins, the C-x-x-C-H motif is strictly conserved (except, PDB ID: 1BBH) (Figure 2.8). The microenvironment around two cysteines in the motif is similar in all the cytochrome proteins (Table 2.12) The small standard deviation value (0.14 for CYS1 and 0.13 for CYS2, Table 2.12) indicates conservation of microenvironment around those residues.

Table 2.11: Description of cysteines those are part of the C-x-x-C-H motif in cytochrome proteins.

PDB ID, residue number are shown. Microenvironment clusters are written within parenthesis.

S.no	PDB	Cysteines with chain id (cluster name)
1	1BBH.PDB	121A,B (intermediate) , 124A,B (intermediate)
2	1C75.PDB	32A (intermediate) , 35A (exposed-hydrophilic)
3	1CIF.PDB	14A (buried-hydrophilic) , 17A (intermediate)
4	1CTJ.PDB	15A (intermediate) , 18A(exposed-hydrophilic)
5	1E29.PDB	37A (intermediate) , 40A (exposed-hydrophilic)
6	1GU2.PDB	49A and 52A (intermediate) , 49B and 52B (intermediate)
7	1H32.PDB	76A (intermediate) and 79A (exposed-hydrophilic) , 42B (buried-hydrophilic) and 45B (intermediate), 177 (intermediate) and 180 (intermediate)
8	1I8O.PDB	13A (intermediate) , 16A (intermediate)
9	1J0P.PDB	30A and 33 A (exposed-hydrophilic) , 79A and 82A (exposed-hydrophilic)
10	1M70.PDB	119A(buried-hydrophilic),122A(exposed-hydrophilic) , 119B(buried-hydrophilic), 122B(intermediate) , 119C(buried-hydrophilic) , 122C(intermediate) , 119D(buried-hydrophilic) , 122D(exposed-hydrophilic) ,
11	1QKS.PDB	65A and 68A (intermediate) , 65B (buried-hydrophilic) and 68B (intermediate)

12	2C8S.PDB	65A (buried-hydrophilic) , 68A (exposed-hydrophilic)
----	----------	--

Table 2.12: Buried fraction and rHpy values of cysteines part of C-x-x-C-H motif found in cytochrome proteins.

PDB	CYS1	Buried-fraction	rHpy	CYS2	Buried-fraction	rHpy
1BBH	121A	0.651	0.401	124A	0.638	0.437
1BBH	121B	0.651	0.433	124B	0.661	0.414
1C75	32A	0.764	0.212	35A	0.440	0.686
1CIF	14A	0.806	0.119	17A	0.659	0.394
1CTJ	15A	0.745	0.223	18A	0.501	0.523
1E29	37A	0.748	0.238	40A	0.285	0.756
1GU2	49A	0.751	0.33	52A	0.707	0.312
1GU2	49B	0.728	0.364	52B	0.707	0.353
1H32	76A	0.741	0.246	79A	0.412	0.645
1H32	42B	0.837	0.24	45B	0.598	0.457
1H32	177A	0.703	0.332	180A	0.701	0.311
1I8O	13A	0.766	0.334	16A	0.642	0.386
1J0P	30A	0.244	0.77	33A	0.397	0.642
1J0P	79A	0.521	0.53	82A	0.539	0.58
1M70	119A	0.805	0.197	122A	0.527	0.475
1M70	119B	0.797	0.208	122B	0.683	0.348
1M70	119C	0.801	0.202	122C	0.688	0.3
1M70	119D	0.786	0.217	122D	0.486	0.525
1QKS	65A	0.768	0.279	68A	0.625	0.463
1QKS	65B	0.784	0.27	68B	0.652	0.439

2C8S	65A	0.782	0.19	68A	0.416	0.617
------	-----	-------	------	-----	-------	-------

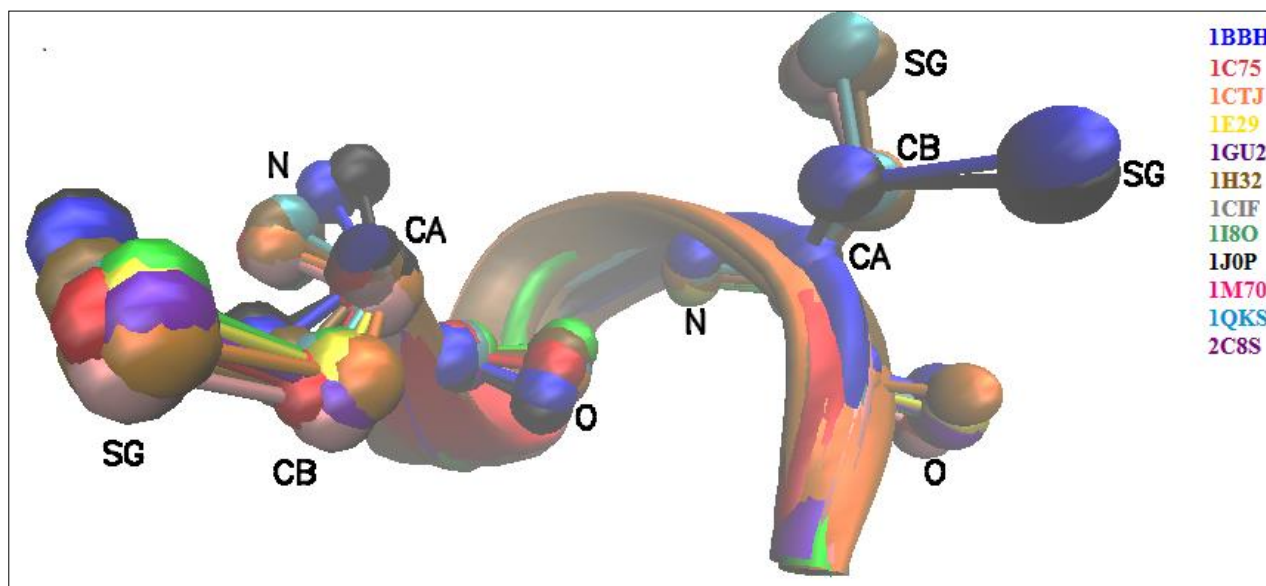


Figure 2.7: The structurally aligned C-x-x-C-H motif in the 12 cytochrome proteins is shown in different colors. The RMSD of this conserved motif is 0.54 Å. The atoms of cysteine residues are labeled and represented in ball and stick conformation. The cartoon representation of C-x-x-C-H motif. These two XX residues in the motif mostly contain alanine and serine (Figure 2.8).

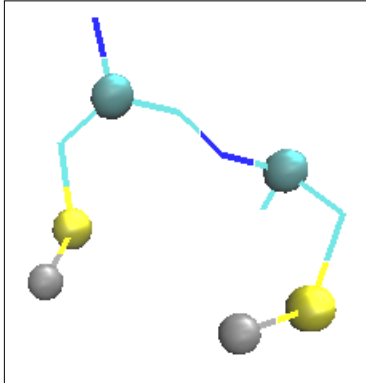
Query_10002	1	-----VDAEAVV	QQ-KL	ESCHGGDLTGAS	---APAIDKAGANYSE	[2]	-----ILDEILNGQ	47	IclQuery_10002	1C75
Query_10003	1	-----TEFKAGSAXKGATLF	KT-R	LQCHTVEKGGPH	-kvGPNLHGI	FGAHSG	QAEGYSYTDANIKKN	61	IclQuery_10003	1CIF
Query_10004	1	-----EADLALGKAVF	DG-N	LAACHGGGN	-----VIPDHTLQKAA	[4]	LDDGFNIEAIVYQIE	55	IclQuery_10004	1CTJ
Query_10005	1	[12]DEAGGTTTLTARQFTNGQKIF	VD-T	TQCHLQKTKT	hmvSLGLADLAGAEP	R	RDWVLAALVEFLKNPK	80	IclQuery_10005	1E29
Query_10006	1	[14]IAHSANPHYEAPSITDGKIFF	[9]KEA	ACASCHTNNPAHWG	-----KNI	VTGKEIP	[4]RVVTKRFTDIDKVED	90	IclQuery_10006	1GU2
Query_10007	1	[36]FRDDDTQALEHDDFENSQMVF	[14]EGKA	ADCHGAVDDGMY	-----GLRAVYPKYVE	[4]VRTVEQ	MINACTSR	118	IclQuery_10007	1H32
Query_10008	1	-----EDAKAGEAVF	KQ--	HTCHTRADKN	---mvGPALAGV	VGRKAG	TAAGFTYSPLMHN	52	IclQuery_10008	1I8O
Query_10009	1	AAPKAPADGLKMDTKQPVVF	[5]KAVK	GDCHHPVNGKED	-----LQKCATAGCHDN	[4]DKSAKGY	YHAMHDXG	74	IclQuery_10009	1J0P
Query_10010	1	-----AGDAEAGQGK-	VA-V	GACHGVDSNS	---pAPWPKLAGQGER	---	YLLKQLQDIKAG	50	IclQuery_10010	1M70
Query_10011	1	[40]GAPEGVLTALSDAQYNEANKIY	FE-R	LAGCHVLRKIGAT	---GKALTPDLTRDLG	-	FDYLQSFITYASPA	104	IclQuery_10011	1QKS
Query_10012	1	[41]TGENLYIDOK-SCLRNGESLF	AT-S	SGCHDLAEG	---kIGPGLNDN	MYMTPYS	NTTOVGLFATIFGGA	105	IclQuery_10012	2C8S
Query_10002	48	GGMPG[ 8]-EAVAAMLAEKK-----						71		
Query_10003	62	---VLWDERN--MSE-VLTHPKKYIP----					GTKMAS	108		
Query_10004	56	NGKGA MPAMDGRLDEDEIAGVAAYVY----					[2]AA---GNKM--	89		
Query_10005	81	SYDGE DQYSE---LHPNISRPDIYPEHRNY[7]VA---					GYTLIA	135		
Query_10006	91	EFTKH[ 1]NDILGAOCSPEKANFIAYLLTETK[3]-----						124		
Query_10007	119	MGAPE[10]MVALIASVSRGHPVSVADIGPAQST[7]YYtryGQLDLS[26]PSYRLKNAR[12]RDTRGVVFAVG[ 27]						261		
Query_10008	53	EAGLV WTADN--IVP-YLADPNMFLKKFLT[7]AV---					GVTXMT	114		
Query_10009	75	TKFKS[ 6]ETAGADAARKKELTGCKGSKCHS--						108		
Query_10010	51	STPGA[10]HTGHLDPISDQDLEDIAAYFSSQKG[7]ALakqGEKLF						190		
Query_10011	105	GPPNM --GTSGLSAEQVDMANYLLDPA APpeFGMKEMR						567		
Query_10012	106	NGPMG PHNEN--LTPDEMLQTIAMIRHLYT -----GPKQDA						172		

Figure 2.8: Local alignment of 11 cytochrome proteins showing conservation of –C-x-x-C-H motif region obtained from by aligning the amino acid sequences of the proteins using COBALT local alignment server (Papadopoulos and Agarwala, 2007).

### 3.3.7) Conservation of microenvironment in cysteines around specific zinc binding motifs

The metal binding cysteines found in the dataset have preferences towards metal ions, iron, and zinc. The iron binding cysteines (87 out of 170 metal binding cysteines) were mostly found as part of cytochrome proteins (12 proteins out of total 21 proteins containing iron binding cysteines), those were discussed in the previous section. In this section, the correlation between the microenvironment and the zinc binding cysteines is discussed. The total number of zinc binding cysteines are 91 (out of 170). These 91 cysteines belong to only 17 proteins.

Out of 91 zinc binding cysteines, 66 were found in 5 types of motif regions (Figure 2.9), remaining 25 cysteines were not part of any kind of motif region. The motifs identified were: (i) CC (ii) C-x-C (iii) C-x-x-C (iv) C-x-x-x-C (v) C-x-x-C-x-x-C. The highest number of cysteines were found in C-x-x-C motif region (48 out of 91).

Motif	Representing Structure
CC	



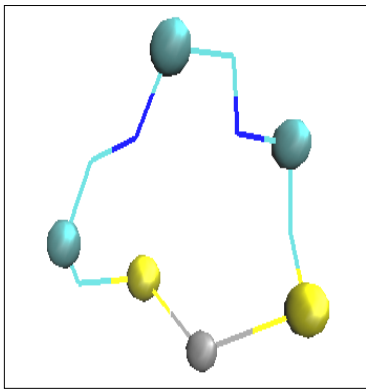
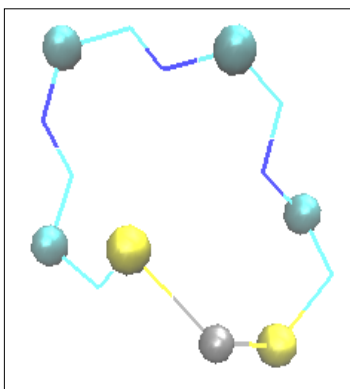
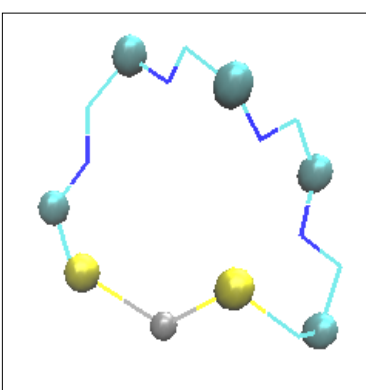
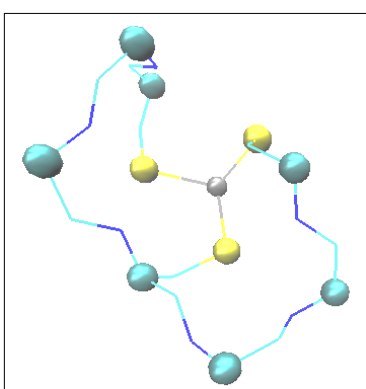
C-X-C	
C-x-x-C	
C-X-X-X-C	
C-x-x-C-X-X-C	

Figure 2.9: Zinc binding cysteine motifs. Yellow balls represent sulfur atoms, cyan balls represent C-alpha atoms and the grey balls represent zinc ion. Remaining protein parts are shown in lines.

The proteins containing the C-x-x-C motif were subjected to structural alignment using Salign server but were found to be dissimilar as full proteins. This motivated the structural alignment of only the C-x-x-C motif region from all the respective protein structures (Figure 2.10). The RMSD for the aligned C-x-x-C motif region was calculated to be 0.19 Å.

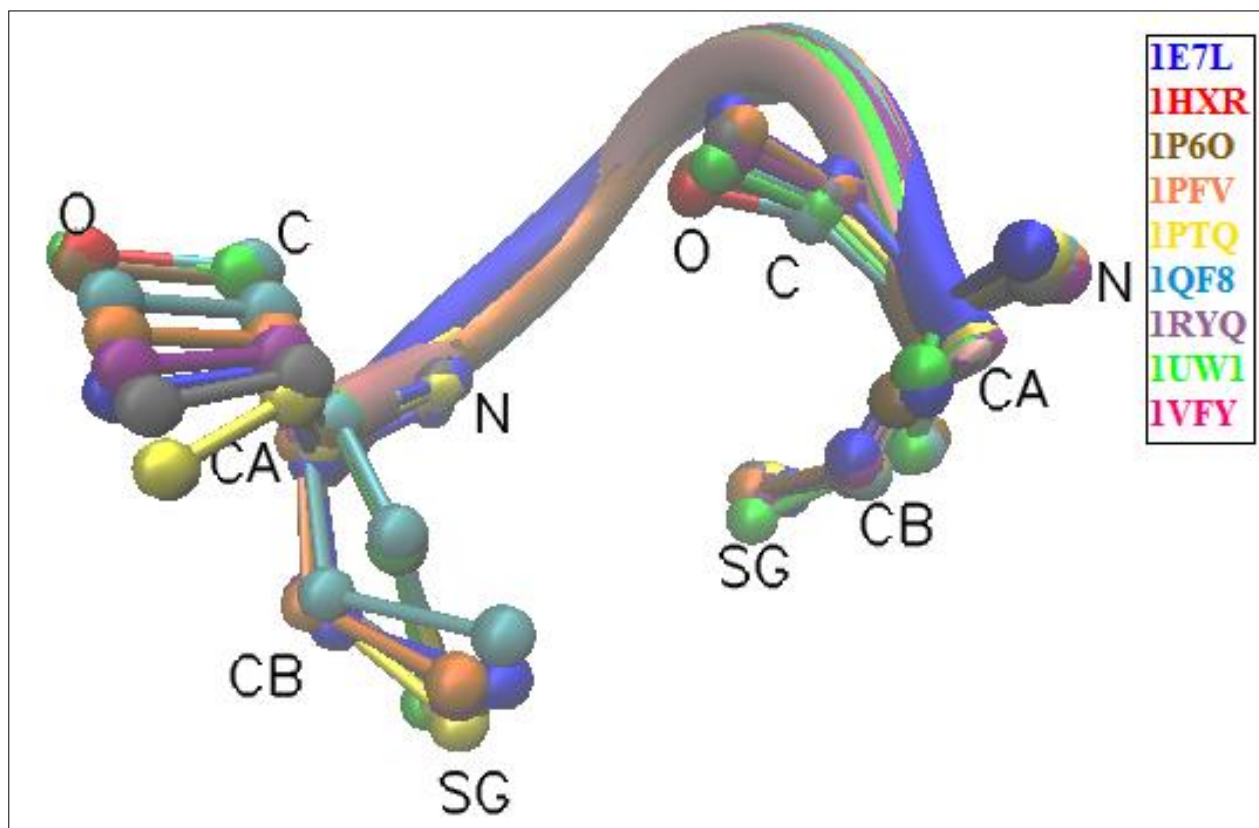


Figure 2.10: The aligned C-x-x-C motif from 9 proteins with zinc binding cysteines as part of C-x-x-C motif. For sake of clarity, only single chain from a particular protein is shown. The ball and stick model represents the two cysteines and the other two amino acids are in “New-cartoon” representation.

To further evaluate the similarity levels among the proteins with cysteines those are part of C-x-x-C motif, multiple sequence alignment was performed using t-coffee MSA tool (Li et al., 2015). The global multiple sequence alignment of these 9 proteins shows only 26% sequence similarity. Although the conservation in the microenvironment for the cysteine in the motif region was

estimated by calculating the average rHpy for the respective cysteines in the motif region (Table 2.13).

Table 2.13: Average rHpy and standard deviation values for the zinc binding cysteines found in different motifs. Out of 91 zinc binding cysteines, 66 are found as part of the different type of motifs. The total number of cysteines in a particular motif region is given in parenthesis.

Motif (66)	C-x-x-C (48)		C-X-C (4)		CC (4)		C-x-x-C-x-x-C (6)			C-x-x-x-C (4)	
	CYS1	CYS2	CYS1	CYS2	CYS1	CYS2	CYS1	CYS2	CYS3	CYS1	CYS2
Avg. rHpy	0.12	0.19	0.16	0.22	0.16	0.39	-0.04	0.07	0.21	0.45	0.20
Std. dev	0.07	0.15	0.02	0.05	0.02	0.00	0.03	0.01	0.00	0.02	0.04

The conservation of microenvironment was tested only for cysteines in C-x-x-C motif region because of a relatively higher number of data points in this motif as compared to cysteines in other motif regions. The average rHpy and the standard deviation values show that the functionally similar cysteines in the C-x-x-C motif region of the zinc binding proteins have a conserved microenvironment. This conservation of the microenvironment is due to the identical C-x-x-C motif.

The conservation of microenvironment around the cysteines in cytochrome protein and zinc binding proteins with C-x-x-C-H and C-x-x-C motif respectively is observed. The buried hydrophilic microenvironment around the cysteines in C-x-x-C-H is necessary for the action of cytochrome proteins. Similarly, these studies related to the conservation of microenvironment may aid in diagnosing the function and the catalytic mechanism of unknown protein structures.

## 2.4) Conclusions

The aim of this study was to find out the answer raised in the introduction section - “Does the specific type of thiol function prefer a certain type of protein microenvironment, irrespective of the nature and function of the individual protein?” Analysis carried out here demonstrated that active cysteines prefer buried hydrophilic cluster, metal binding cysteines prefer intermediate cluster and redox cysteines prefer exposed hydrophilic cluster. To note, these protein microenvironments were similar, in terms of their buried fraction and rHpy values, although those belong to diverse protein structures. The functions of microenvironment clusters were studied based on cysteine population in the particular environment cluster (Figure 2.11).

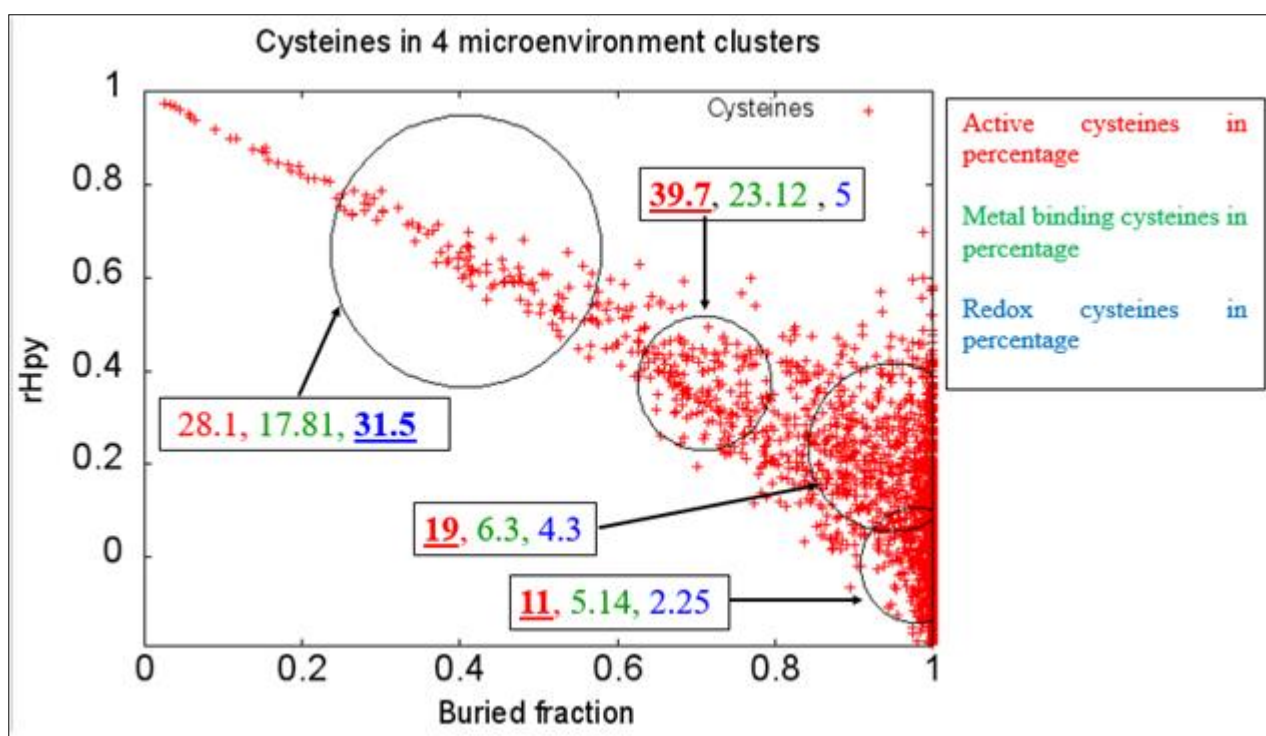


Figure 2.11: Frequency of active, metal binding, and redox cysteines in the buried-hydrophobic, buried-hydrophilic, intermediate and exposed hydrophilic clusters. The percentages of each function in a cluster based on cysteine frequency are denoted by different clusters and those in bold and underline are the cysteines of a particular function that occur mostly in a microenvironment cluster.

Though the active cysteines are mostly widespread in the buried region, with an increase in hydrophilicity the number of metal binding and redox cysteines also increases. In future, microenvironment based functional annotation of other amino acids could possibly elucidate the mechanistic details of many other biochemical processes. Moreover, the functional categorization of cysteine residues based on the secondary structure and microenvironment based free energy scoring function can be used to perform blind prediction of functions of cysteine residues in unknown proteins. Understand the amino acid function will aid in understanding the role of a particular protein in the respective biochemical reaction.

## CHAPTER 3

# Modulating Effect of microenvironment on structure and function of disulfide bridged cystine residues

### 3.1) Introduction

Cystine amino acid is formed by the oxidation reaction between two cysteine amino acid residues (Howland, 1990), (Branden and Tooze, 1991) , (R.Barnes and C.Gray, 2003). Cystines are formed as a post-translational modification of cysteine residues which are involved in intra or inter chain disulfide bonds. This disulfide linked cystine molecule is energetically stable (average energy around 60 Kcal/mol) (Hazes and Dijkstra, 1988). These cystine residues mainly impart thermodynamic and structural stability by lowering down the entropy of the unfolded state (Fass, 2012) to the respective protein structures. The disulfide bridged cystine amino acid residue provides structural stability (Lee et al., 2008) and helps in proper protein folding (Bardwell, 1994). Disulfide bonds are necessary for the immunoglobulin proteins to form a stable quaternary structure (Seegan et al., 1979). Cystine residues are also found in active site pocket in protein families like disulfide bond isomerases (Lee et al., 2008) and thiol oxidoreductases (Marino and Gladyshev, 2009). The oxidation of cysteine residues that lead to the formation of disulfide bridged cystine residue is governed by various parameters: (i) side chain distance between the two cysteine residues; (ii) pKa lower than 9.06; and (iii) significant exposure (greater than 1.3 Å) of the sulfur atom to solvent (Poole et al., 2004) . These parameters depend on the type of atoms present in the vicinity of the amino acid.

It is evident that the formation of the disulfide-bridged cystine residue depends on the local microenvironment that surrounds the two nearby cysteine residues involved in disulfide bond formation. The local microenvironment is defined as the three-dimensional arrangement of atoms around a functional group up to the first hydration shell (Bandyopadhyay and Mehler, 2008). The local environment includes the contributions from both protein medium and the surrounding solvent

medium, which is generally water. It is already known that structural and functional characteristics of amino acids vary with modulating surrounding local microenvironments (Rekker, 1977), (Eisenberg D, 1986), (Wimley et al., 1996), (Mehler and Guarnieri, 1999). Modulation in the local microenvironment alters the pKa of the amino acid's side chain. This phenomenon is mainly observed in titrable amino acids; for example; arginine in voltage-gated ion channels (Jiang et al., 2003). Aspartic acid in the photocycle of bacteriorhodopsin (Moukhametzianov et al., 2006). Aspartic acid, Histidine, Cysteine etc. in several acid-base catalyzed hydrolysis reactions (Harris and Turner, 2002). It has also been found that microenvironments around conserved residues was conserved with a protein family for example serine proteases and microenvironments around the active tryptophan residues in the proteins containing the immunoglobulin fold (Bandyopadhyay and Mehler, 2008). The above-mentioned instances clearly raise the question that specifically how microenvironment modulates the role of an amino acid in a protein and what are the biochemical intricacies that are involved in its structure and function modulation.

The disulfide bridged cystine amino acid residue is taken as a model system (due to its limited functionality of mostly structure stabilization and involvement in few catalytic reactions) to analyze the effects of modulating environment on the disulfide bond strength and the function of cystine residues in different protein microenvironments. Microenvironment was quantified using Quantitative Property Descriptor (QPD), which is known as rHpy (Bandyopadhyay and Mehler, 2008). Cystines in different microenvironments were clustered according to agglomerative hierarchical clustering.

Three protein microenvironment clusters were identified: i) buried-hydrophobic, ii) exposed-hydrophilic and iii) buried-hydrophilic. The buried-hydrophobic cluster encompasses a small group of nine redox-active cystines, all in alpha-helical conformations in a -C-x-x-C- motif from the Oxidoreductases enzyme class. All these cystines have high strain energy and near identical microenvironments. Cystines embedded in exposed-hydrophilic microenvironment cluster, mainly belong to the hydrolase enzyme class. These directly or indirectly participate in enzyme catalysis,

ligand binding or imparting stability to the active-site. Many of these have low to medium sequence conservation. Despite low sequence conservation, they share very similar microenvironments. In the buried-hydrophilic cluster, forty four cystines were identified those were completely buried yet embedded in hydrophilic microenvironments. All these were part of hydrolase enzyme class. Many of these cystines participate in stabilizing the catalytic or active sites of the enzymes, without direct involvement in enzyme function.

### **3.2) Methodology**

The microenvironment dataset contains 175 high-resolution protein crystal structures with a sequence identity of 25% that contains 1400 half-cystines (700 disulfide bridged cystines). A half-cystine corresponds to the single sulfur atom involved in the disulfide bond. The high-resolution crystal structures include the protein structures resolved with resolution less than 2.0 angstrom from Protein Data Bank database (PDB) (Berman et al., 2000). The 25% sequence identity is sufficient to generalize the analysis for the whole family of protein (Hobohm and Sander, 1994).

#### **3.2.1) Generation of Cystine microenvironment database**

The microenvironment was computed for 1400 half-cystines based on the Quantitative Property Descriptor (QPD) known as rHpy. rHpy is the ratio of the hydrophobicity of the side chain of an amino acid in a particular microenvironment with respect to the hydrophobicity in water microenvironment within the radius of the first hydration shell around the atom. This radius for carbon atom is 4.475 angstroms ( $2.275 + 0.2$ ), 2.275 is the Vander wall radius of carbon atom and the increase of 0.2 represents the contribution of polar atoms like oxygen or nitrogen. Therefore, the microenvironment range for cystine residues that allows all the contributing atoms present in the first hydration shell was defined as 4.5 and is diagrammatically shown in figure 3.1.



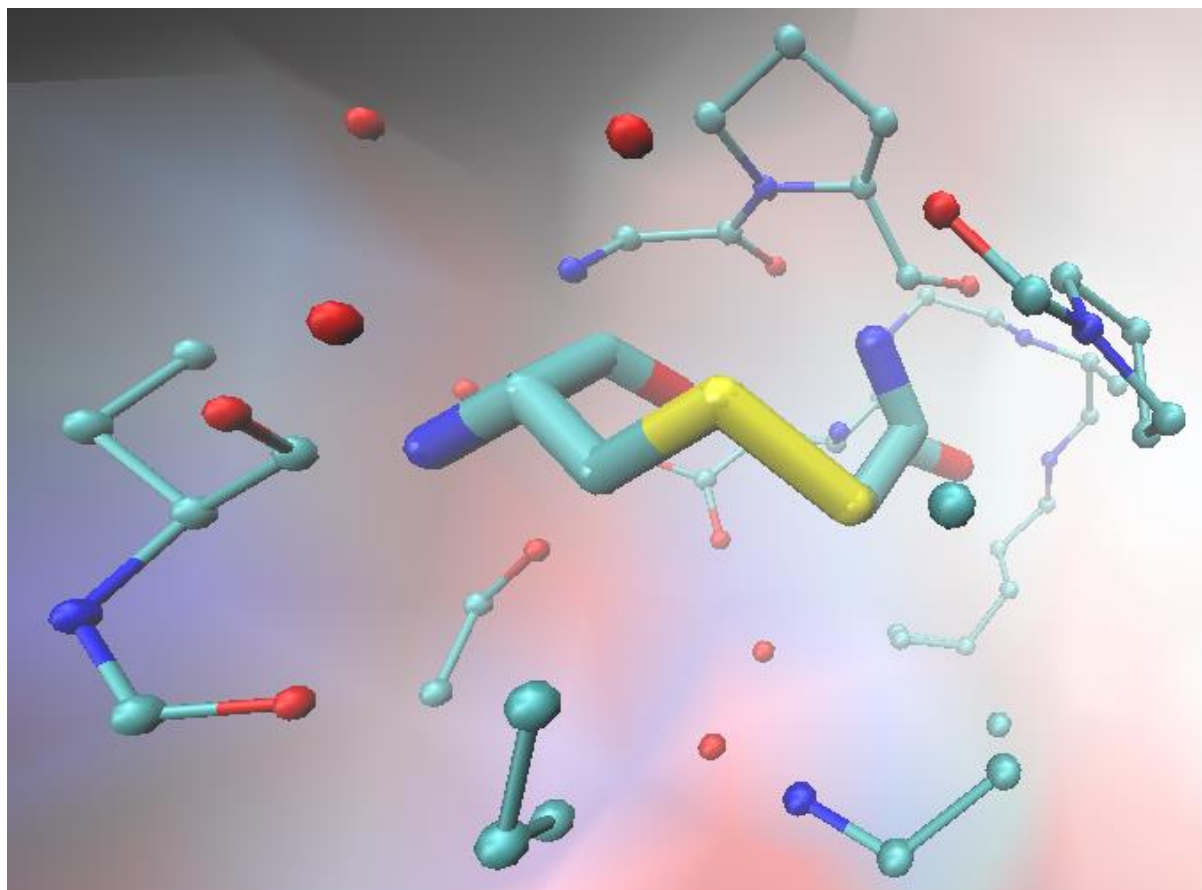


Figure 3.1: Schematic representation of microenvironment around disulfide-bridged cystine residue from a thioredoxin protein (PDB ID: 1ABA, residue no's: 14 and 17). The cystine is shown as tube representation. The microenvironment around the cystine is shown in ball and stick representation. Oxygen atoms of water molecules are shown as isolated red balls. The protein background is shown as a surface representation. The schema is created using Visual Molecular Dynamics (VMD) (Humphrey et al., 1996).

The highest limit of  $rH_{py}$  is  $\sim 1$  that corresponds to complete hydrophilic nature of the functional group, whereas the lower limit can reach any level which marks the lowering of hydrophilicity and an increase in the hydrophobicity of the functional group (Bandyopadhyay and Mehler, 2008). In case of disulfide bridged cystine residues, the lowest value of  $rH_{py}$  obtained was  $-0.4$ . Although,  $rH_{py}$  is alone capable of describing the microenvironment around an amino acid but to populate the cystine residues onto a microenvironment space, buried fraction (BF) was also calculated. Buried fraction quantifies the extent of a side chain functional group of an amino acid buried inside the protein. This parameter is computed using the program GEPOL93 (Pascual-ahuir et al., 1994). The

quantification of buried fraction is performed by taking the ratio of the solvent inaccessible surface area of an amino acid side chain functional group with the total surface area of the functional group. Buried fraction ranges from '0 to 1'. Value '1' corresponds to the fraction of the amino acid side chain that is completely buried and '0' represents the state when the fraction of the amino acid side chain is completely exposed to the solvent (Bandyopadhyay and Mehler, 2008).

The rHpy and buried fraction were computed for all the 1400 individual half-cystines (700 disulfide bridged cystine pairs) from 175 high-resolution protein crystal structures.

### **3.2.2) Curation of functional role of cystines from literature and secondary databases**

The role of a particular half-cystine in the respective protein structure was initially curated manually from the respective research articles of the 175 proteins in the dataset. The information obtained after the manual curation was further verified by the text mining software GATE (Cunningham et al., 2013). The catalytic and active site cystines in respective proteins were obtained from Catalytic Site Atlas database (Furnham et al., 2014) and PDBsum databases (de Beer et al., 2014) using a database crawling PERL script. The exact 700 disulfide connected cystines in the individual proteins were directly obtained from the respective PDB (Berman et al., 2000) header files through a customized PERL-C-SHELL Script.

The enzymatic proteins are distributed into six enzyme classes by the nomenclature committee of International Union of Biochemistry and Molecular Biology (IUBMB) (Bairoch, 2000). Such enzymatic proteins in the dataset were identified and obtained by PERL script from the individual PDB header files. The 74 enzymes obtained are reported in table 3.1.

Table 3.1: Different proteins present in all possible enzyme classes found in the current dataset. It is to note that none of the half-cystines in the dataset were found in protein part of Isomerases and Ligases enzyme class.

S.no	Enzyme class	No. of proteins	PDB ID
1	Oxidoreductases	11	1D2V, 1FI2, 1GU2, 1JR8, 1K3I,1KNG,1LYC, 1MFM,1ST9,1FVK, 1JM1,
	Electron transport*	6	1ABA,1CXY, 1DJ7,1H75,1THX, 1UMM
2	Transferases	3	1FLT,1GXY, 1VKB
3	Hydrolases	53	1BQC,1C2A,1C7K,1DTD,1DY5,1EB6,1G66,1G6I,1G6X, 1GPI,1GPQ,1H9H,1HX0,1I1W,1I71,1K55,1K5C,1K7C,1 KLI,1KNM,1KS8,1KUF,1LBU,1LLF,1LNI,1LWB,1M40, 1M4L,1OC7,1OGO,1PQ7,1PWG,1PYO,1QLW,1QNR,1Q WO,1R0R,1RTQ,1SSX,1TZP,1U4G,1UUZ,1UWC,1V0W ,2NLR,3LZT,1O5F,1HTR, 1BEA,1PGS,3SIL, 1AK0, 1QL0
4	Lyases	1	1LK9

\*The electron transport proteins are involved in redox type of reactions and perform similar functions as oxidoreductases; therefore these proteins were also included as functional proteins containing disulfide bridged cystine amino acid residues.

### 3.2.3) Categorization of the cystine microenvironment database into individual clusters

Clustering of a dataset is performed using many algorithms, designed according to the type of clustering required by the user. K-means clustering divides the dataset based on a centroid value

and describing the clusters around that central value. Hierarchical clustering works by grouping similar observations together into one cluster and thus produces dissimilar clusters. (Hartigan, 1975). Agglomerative Hierarchical clustering is a subtype of Hierarchical clustering. Agglomerative considers each observation initially as a single cluster; based on the distance proximity to the nearest cluster center, the smaller clusters combine to form large dissimilar cluster pairs. The microenvironment space was also divided into small clusters with equal spacing [(BF, rHpy) = (0.1, 0.1)] representing the smallest set of clusters. Here, the clusters should represent the individual cystine that possesses similar microenvironments, hence the Agglomerative Hierarchical type of clustering was opted to categorize the different functions of individual cystine microenvironment clusters. Agglomerative clustering method incorporates small microenvironment bins into larger dissimilar clusters where the Euclidean distance of each bin from cluster center is minimum with respect to other cluster centers. Ward's method (Ward, 1963) (Murtagh and Legendre, 2014) of hierarchical agglomerative clustering method was employed with XLSTAT software (Addinsoft, 2014). The clustering was based on the dissimilarity in the euclidean distance among the observation in the dataset. The truncation was kept automatic that enables the algorithm itself to decide the number of clusters to be generated. The missing data observations were not allowed. Properties like within-class variance, centroid distance, mean and standard deviation values were recorded.

#### **3.2.4) Identification and Distribution of protein containing cystine residues into the Structural Classification of Proteins (SCOP) class and calculation of secondary structure for 175 proteins**

The Structural Classification of Proteins Class (Murzin et al., 1995), (Hubbard et al., 1997) is a database that distributes the proteins based on the type of secondary structure predominantly found in the proteins. This classification was also performed for the 175 proteins present in the cystine microenvironment dataset.

The secondary structure distribution was verified by secondary structure calculation by using Define Secondary Structure of Proteins (DSSP) algorithm (Kabsch and Sander, 1983) , (Joosten et al., 2011) for all the 175 proteins and extracting the secondary structure where all the 1400 half-cystines are present in the individual protein using a PERL-C-SHELL script. The secondary structure of individual half-cystines was also clustered into the individual cystine microenvironment clusters.

### 3.2.5) Calculation of cystine disulfide strain energy and dihedral angles for the 700 cystines from 175 high-resolution protein crystal structures in the dataset

To understand the role of microenvironment in the disulfide bond strength; the geometry of the disulfide bond was analysed by calculating the disulfide strain energy and the five side chain cystine dihedral angles ( $\chi_1, \chi_2, \chi_3, \chi_1', \chi_2'$ ) (figure 3.2) for all the 700 cystine molecules through the Disulfide Bond Dihedral Angle Energy Server (<http://services.mbi.ucla.edu/disulfide/>) (Schmidt et al., 2006) , (Katz and Kossiakoff, 1986).

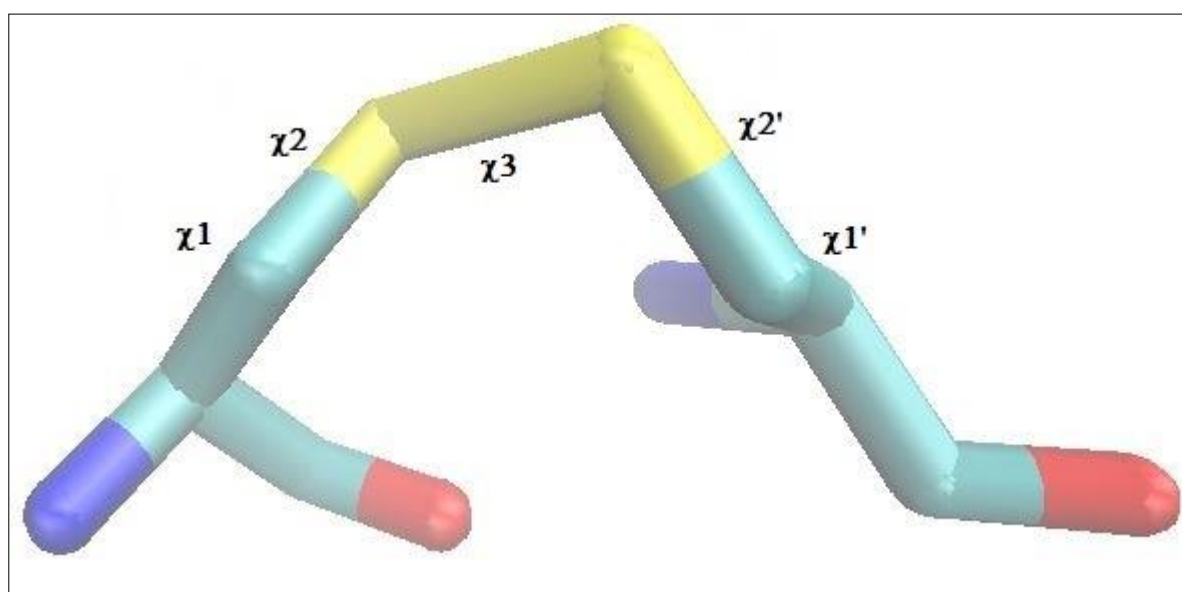


Figure 3.2: Schematic representation of different dihedral angles observed in disulfide-bridged cystine structure. This tube model representation of the cystine structure was generated using Visual Molecular Dynamics (VMD) (Humphrey et al., 1996) from the structure of oxidized bacteriophage T4 Glutaredoxin (thioredoxin) protein (PDB ID: 1ABA) (Eklund et al., 1992). The red, blue and

cyan color represents the oxygen, nitrogen, and carbon atoms respectively. Similarly, the sulfur atoms are represented in yellow color forming the disulfide bond.

### **3.2.6) Sequence and structural alignment of specific proteins in the dataset to establish microenvironment conservation around cystines within a microenvironment cluster**

The alignment studies were performed for proteins that are present in oxidoreductases and hydrolases enzyme class, as these represent the total cystine microenvironment database based on enzyme classification. The local sequence alignment refers to the alignment of sequence in terms of a particular short region within the sequences, whereas global alignment aligns the sequences by considering each and every residue in the two sequences that are mostly of similar length (Mount, 2004). The local alignment was performed using COBALT (Papadopoulos and Agarwala, 2007) and global sequence alignment was performed using ClustalW (McWilliam et al., 2013). A customized distance based PERL script is also being formulated that counts and report the type of atoms present within the 4.5 Å (first contact shell) region of an amino acid residue.

### **3.2.7) Analysis of variation of cystine disulfide bond strength with varying microenvironment**

The half cystines of a cystine molecule belong to two different microenvironments. To check and compare the effect of microenvironment on cysteine bond geometry, a single representative from each cluster pair was selected. The 4.5 Å region around any functional represents its microenvironment (Bandyopadhyay and Mehler, 2008). This region was selected around the 6 cystine molecules that belongs to six microenvironment cluster pairs (Table 3.2) using VMD (Humphrey et al., 1996).

Table 3.2: Six representative proteins from each cluster pair for quantum chemical calculations.

S.no	Cluster pair	PDB ID	-CYS-CYS-residue no (chain ID)
1	Buried-hydrophobic – Buried hydrophobic	1AK0	80 (A) – 85 (A)

2	Buried-hydrophilic – Buried-hydrophilic	1AGQ	73 (B) – 134 (B)
3	Exposed-hydrophilic – Exposed-hydrophilic	1RO2	185 (A) – 206 (A)
4	Buried-hydrophobic – Buried-hydrophilic	1ABA	14 (A) – 17 (A)
5	Buried-hydrophobic – Exposed-hydrophilic	1BX7	33 (A) – 50 (A)
6	Buried-hydrophilic – Exposed-hydrophilic	1C7K	99 (A) – 112 (A)

To maintain neutrality and a protein mimicking system, full amino acids were taken in the microenvironment region and these were neutralized by adding hydrogen and hydroxyl groups at requisite N and C terminal positions using molden (Schaftenaar and Noordik, 2000). These six systems were subjected to optimization using 6-31G (d, p) basis set. As the systems were neutral, therefore, the method used for estimation of wave function was Restricted Hartree-Fock approximation method (Cramer, 2005). The optimized structures were used as input for natural bond orbital (NBO) analysis to calculate the bond order of the disulfide bond in the 6 different systems. For reference, a cystine from glutathione protein (PDB ID: 1ABA) was simulated in water and DMSO microenvironment in a 10 Å box. The microenvironment region of 4.5 Å around these simulated cystine residues was extracted to prepare the neutral system of cystine in the respective solvent microenvironment. The simulation was performed using NAMD (Phillips et al., 2005). The charmm topology and parameter files were used to generate the respective protein structure files (psf) (Brooks et al., 2009). Before Molecular Dynamics simulation, the systems with cystine dissolved in water and DMSO solvents were energy minimized with 10000, 1000 and 100 iteration steps using both steepest descent and conjugate gradient methods. Minimization was performed at 310 kelvin. The molecular dynamics simulation was standardized by varying barostat oscillation time for Langevin piston method. An accurate barostat oscillation time constant is necessary to maintain system uniformity.

Equilibration was done for 10ns with oscillation time constant value as 1000 fs and damping time constant with a value of 500fs followed by the production run. Based on the rmsd plot over the trajectory file obtained after the production run generated using RMSD trajectory tool plugin available with VMD, the other nine equilibrations were done for 5 ns with varying values of oscillation time constant and damping time constant which ranges from 10,000,00 to 100 and 5,000,00 to 50 respectively. In the case of DMSO, the values of oscillation time constant and damping time constant varied from 10,000 to 26,000,00 and 5000 to 13,000,00 respectively.

The equilibrations were run on NPT ensemble with a constant temperature of 310 K (physiological temperature), the electrostatic interactions among the 1-4 pairs of atoms were modified according to the charmm parameter file and the other pairs were excluded from non-bonding interactions. A cutoff distance of 12 angstroms was taken at which the Vander wall potential energy becomes zero. One time step constitutes 1 fs. Temperature control was maintained by Langevin dynamics according to the X-PLOR (Schwieters et al., 2003) user manual. The periodic boundary conditions were decided as 24, 23 and 23 Å. These values were based on the length of the system in all three x, y and z directions. It is to note that the center of all the systems was maintained at the origin. The pressure control was done using Langevin piston Nose-Hoover method. The pressure was applied to the whole system. The production run was done using NVE (constant no. of atoms, constant volume, and constant energy) ensemble and therefore there was no pressure control but other parameters were kept same as that for equilibration.

The cystine which is most nearer to the origin point was selected from the obtained trajectory files after the Molecular Dynamics Simulation. This was done by using a tcl script which retrieves x, y and z coordinates of all the frames in a trajectory file. The residence time for each frame was calculated with a distance cut-off of 2 angstroms. Residence time is defined as the number of frames in the trajectory file having cystine molecule within 2 angstroms distance from the origin. The structures with cystine nearest to the center were selected for quantum chemical calculation



using GAMESS (Schmidt et al., 1993) software and the bond orders were calculated using NBO analysis (Glendening et al., 2012).

### 3.3) Results and Discussion

#### 3.3.1) Description of cystine microenvironment clusters

The agglomerative hierarchical clustering on cystine microenvironment database has resulted in formation of three clusters (figure 3.3):

- 1) Buried-hydrophobic cluster
- 2) Buried-hydrophilic cluster
- 3) Exposed-hydrophilic cluster

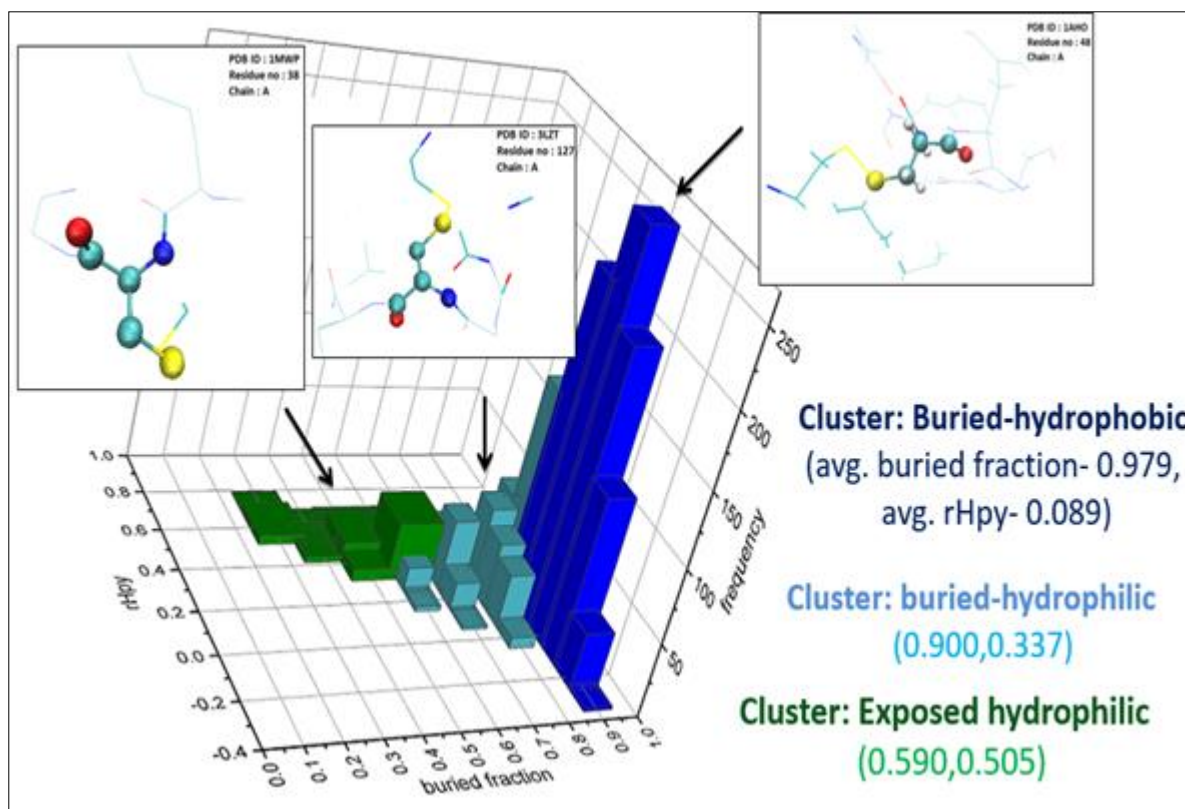


Figure 3.3: Distribution of different microenvironment clusters (obtained from hierarchical clustering) around all the half-cystines from 175 different proteins in the microenvironment dataset. Microenvironments are being clustered in the space of buried fraction and rHpy values. Frequencies in each cluster are shown along the z-axis. The plot was generated using Origin [Origin (OriginLab,

Northampton, MA)]. Insets show ball and stick representations of representative microenvironments around half-cystines from the 3 different clusters. Thicker sticks highlight bonds within the microenvironment around the sulfur atoms and thin sticks represent the extended microenvironment around the cystine residue. The average value of Buried fraction and rHpy values are also given for each cluster in the parenthesis separated by a comma (.). Buried fraction value precedes rHpy.

The number of cystine residues decreases with increase in the hydrophilicity around the disulfide-bridged cystine residues. The buried hydrophobic clusters contain the cystine residues, that are deeply buried inside the protein and possess a hydrophobic microenvironment around them, such cystines (749) are more than half of the total cystines present in the cystine dataset of 1400 half-cystines (Table 3.3). Whereas, the cystine residues found in the buried hydrophilic cluster are also buried in the protein interior but are interestingly present in relatively hydrophilic microenvironment around them (average rHpy 0.337). A very few number of cystines (90) are present in the exposed-hydrophilic cluster and those cystines are found in a relatively hydrophilic microenvironment (average rHpy 0.505) and exposed to solvent (0.590).

Table 3.3: Description of different microenvironment clusters around half-cystines,(S-S)<sup>1/2</sup>. Clusters are defined in terms of buried fraction and rHpy using agglomerative hierarchical clustering (Addinsoft, 2014). Number of other half-cystine (partner in the disulfide) is also reported along with the normalized values in parenthesis. Clusters are arranged, according to descending order of hydrophobicity, measured by cluster center values.

Cluster name	Cluster size	Center of cluster (Buried fraction, rHpy)	Average distance to centroid (Å)	Within-class variance	No. of cystines in different clusters

Buried hydrophobic	749	(0.979,0.089)	0.091	0.011	274 (0.35)
Buried hydrophilic	561	(0.900,0.337)	0.114	0.015	286 (0.50)
Exposed hydrophilic	90	(0.590,0.505)	0.100	0.014	70 (0.77)

The average buried fraction values for clusters buried-hydrophobic and buried-hydrophilic suggests that a major part (1310 half-cystines) of the cystine dataset are buried inside the protein interior but they differ in terms of the hydrophobicity of the surrounding microenvironment, as evident from the average rHpy values (figure 3.4). This shows that cystines generally prefer hydrophobic microenvironment over hydrophilic microenvironment (Bandyopadhyay and Mehler, 2008).

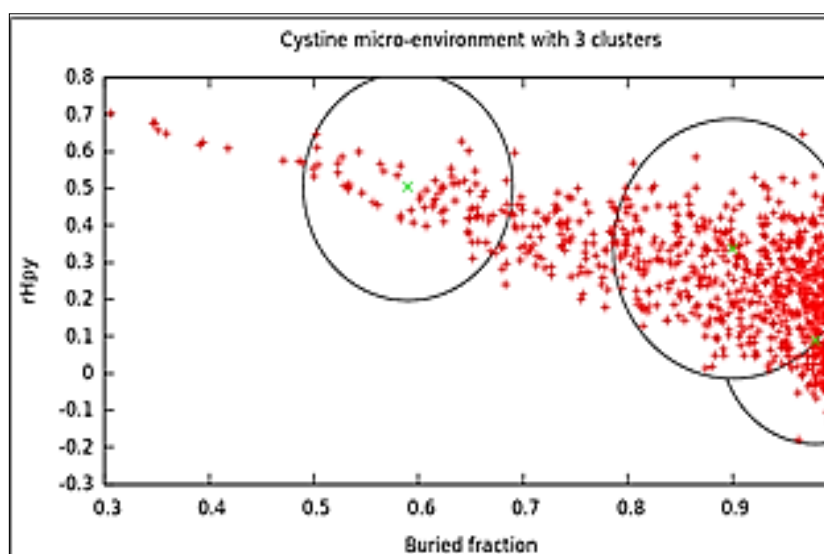


Figure 3.4: Microenvironments around half-cystines (expressed in terms of Buried fraction and rHpy values) distributed in the entire microenvironment dataset. Circles and their centroids represent the results of agglomerative clustering on different datasets. Centroid and the average distance of the clusters are obtained from Table 3.2. Red colored plus signs (+) represents the individual half-cystines and the green colored cross (x) sign is the centroid of the cluster. A very

high number of half-cystines are buried inside the protein as compared to the cystines that are in solvent-exposed region.

It has been observed that not all sulfurs from a disulfide are part of same microenvironment cluster (Table 3.3). The tendency of inter-cluster disulfide bond formation increases, when a half-cystine is exposed to solvent and is enclosed within a hydrophilic microenvironment.

### **3.3.2) Relationship between microenvironment around the half-cystine in each cluster and the secondary, super-secondary structures.**

The half-cystines that belong to a particular microenvironment cluster, whether they harness similar secondary structure is being analyzed. The relationship of cystine amino acid and the protein secondary structures is already known (Harrison and Sternberg, 1996). Local microenvironment can be considered as a local tertiary secondary structure around a particular amino acid. Therefore, there is an expected correlation between the super secondary in particular protein families and microenvironment around the disulfide-bridged cystine residues. It is already known that if the disulfide bonding patterns are similar; that would lead to similar secondary structure fold, families and super-families of proteins even if they have low sequence similarity regions (Chuang et al., 2003).

The DSSP analysis on the protein dataset provides the secondary structure information for the 1400 half-cystines categorized into three cystine microenvironment clusters shows preference of secondary structures into cystine microenvironment clusters (Table 3.4)

Table 3.4: Distribution of secondary structures in different microenvironment clusters, results obtained from DSSP analysis. The values given in parenthesis are normalized with respect to the total number of (S-S)<sup>1/2</sup> in a particular cluster.

Secondary structure	Buried-hydrophobic	Buried-hydrophilic	Exposed-hydrophilic	Total

Helix (G+H+I)	193 (0.26)	125 (0.22)	17 (0.19)	335
Beta strand (B+E)	293 (0.39)	152 (0.27)	19 (0.21)	464
Turn (T)	40 (0.05)	33 (0.06)	14 (0.16)	87
Coil (C)	156 (0.21)	206 (0.37)	34 (0.38)	396
Bend (S)	67 (0.09)	45 (0.08)	6 (0.07)	118

It is observed that helical and beta sheet geometries are favored more in the buried hydrophobic cluster, whereas; the cystine in the exposed-hydrophilic microenvironment cluster generally prefer the flexible conformation like the coil and turn regions in the protein. Similarly, the buried-hydrophilic cluster too predominates the turn secondary conformation over other secondary structures. To further verify this structural preference towards the cystine microenvironment clusters, the Structure and classification of Proteins (SCOP) class categorization into cystine microenvironment clusters (Table 3.5) was performed. Super-secondary structures or protein classes, also known as structural motifs, are defined as the specific arrangement of multiple secondary structures within a protein domain (Branden and Tooze, 1991).

Table 3.5: Number of observations of (S-S)<sup>1/2</sup> (individual sulfur atoms from disulfide bonds) found in different SCOP classes. The numbers are normalized with respect to the total number of sulfur atoms (cluster size) in each cluster given in parenthesis. To note, there are 1400 (S-S)<sup>1/2</sup> (sulfur atoms from disulfide) in 175 different proteins. Hence, same SCOP class appears multiple times in different clusters.

Cluster name	Cluster size	Small proteins	All alpha class	All beta class	Alpha + beta class	Alpha/beta class
--------------	--------------	----------------	-----------------	----------------	--------------------	------------------

Buried-hydrophobic	749 <sup>1</sup>	158 (0.21)	133 (0.18)	190 (0.25)	178 (0.23)	77 (0.10)
Buried-hydrophilic	561 <sup>2</sup>	163 (0.29)	107 (0.19)	112 (0.2)	121(0.22)	54 (0.10)
Exposed-hydrophilic	90 <sup>3</sup>	23(0.25)	8 (0.09)	26 (0.29)	20 (0.22)	10 (0.11)

1. Ten additional proteins from Multi-domain class and three from Coiled-coil class were found.
2. Two additional proteins from Multi-domain class and two from Coiled-coil class were found.
3. Three additional proteins from Coiled-coil class were found.

Half-cystines from all-alpha class of proteins least prefers the exposed-hydrophilic microenvironment cluster. This preference (in terms of normalized frequency) of all-alpha class towards buried microenvironment clusters (both hydrophobic and hydrophilic) is twice as compared to the exposed-hydrophilic cluster. Previously, it has been reported that all-alpha proteins least prefer to form disulfide cross-links, presumably, because of the geometrical constraints arising from alpha-helical structures in both the half-cystines (Thangudu et al., 2008)

Whereas, half-cystines from all-beta class proteins have slightly higher preferences (in terms of normalized frequency) towards exposed-hydrophilic microenvironment cluster. The other three protein classes (alpha+beta, alpha/beta, and small protein) are ubiquitous across all the three microenvironment clusters.

The preferences of different protein classes shown here towards three different microenvironment clusters are very general in nature. In literature, it has been reported that certain super-secondary structures prefer certain enzyme classes, for example, alpha/beta folds in hydrolase (Nardini and Dijkstra, 1999). This attempt is to identify microenvironment specific functions of cystines found with different secondary structures or folds. As the functions of cystines are mainly associated with enzymes, in the subsequent sections, analysis of cystine functions for 74 enzymes from all the enzyme classes (Bairoch, 2000) present in the current dataset is performed.

### **3.3.3) Buried-hydrophobic microenvironment cluster harness the thioredoxin fold proteins with active redox cystines in –C-x-x-C- motif**

The categorization of enzymatic proteins in the cystine microenvironment clusters shows a predominance of hydrolase enzyme class (Table 3.6) with the presence of three other enzyme classes namely: Oxidoreductases that also includes electron transport proteins, lyases and transferases (cystine residues in the current dataset are not part of isomerases and ligases).

Table 3.6: Description of different microenvironment clusters populated with  $-(S-S)_{1/2}$  - from different enzyme classes. The total numbers of  $-(S-S)_{1/2}$  -, present in an enzyme class in each cluster (E) are reported. The number of  $-(S-S)_{1/2}$  - in individual enzyme classes are also reported, along with their normalized values, given in parenthesis. The values are normalized with respect to the total number of cystines present in a particular microenvironment cluster as part of an enzymes class (E).

Cluster	E	Oxidoreductase*	Hydrolase	Lyase	Transferase
Buried hydrophobic	276	40 (0.14)	217(0.78)	4(0.01)	15(0.05)
Buried hydrophilic	213	46 (0.22)	142(0.67)	12(0.06)	13(0.06)
Exposed hydrophilic	33	2 (0.06)	29(0.88)	0(0.00)	2(0.06)

\*The half-cystines that are part of electron transport proteins are also included as they are also involved in redox reactions.

Although, the frequency of hydrolase enzyme is maximum in the buried clusters but the half-cystines from oxidoreductase and electron transport protein are majorly found in buried clusters only (out of 88 half-cystines in oxidoreductases, 86 are present in the buried clusters including both hydrophobic and hydrophilic clusters). To understand the functional role of the buried-hydrophobic microenvironment in oxidoreductase enzyme class, the secondary structure analysis (Joosten et al., 2011) was performed. It has been observed that out of the 88 half-cystine in the oxidoreductases enzyme class 22 are found in the alpha-helical geometry (Table 3.7).

Table 3.7: Number of -(S-S)<sup>1/2</sup> - with different main chain conformations found in enzyme classes, hydrolases, and oxidoreductases. In this table, the secondary structure information is reported only for the major enzyme classes where most of the cystine residues are being found.

Enzyme class	Total number of cystines	Alpha helix	Beta sheet	Loop
Hydrolases	388	72	146	170
Oxidoreductases and electron transport proteins	88	22	5	61

Out of these 22 half-cystines found in the alpha-helical secondary structure, 9 half-cystines (present in the alpha helix) were found as part of a redox active -C-x-x-C- motif region. Moreover, these half-cystines that are part of the -C-x-x-C- motif were present in thioredoxin fold containing proteins. These half-cystines have higher strain energies (Table 3.8) as compared to the average strain energy of this cluster (average strain energy of buried hydrophobic cluster is 10.54 KJ/mol)

Table 3.8: Buried hydrophobic microenvironment cluster: Secondary structures and strain energies of redox-active Cystines (part of C-x-x-C motif), present in different oxidoreductases enzymes and electron transport proteins. Respective cluster numbers are given in parenthesis. Folds of individual residues are mentioned.

PDB ID#	Protein Fold (SCOP classification)	Cystine residue number (chain)	Half-cystine in alpha helix (chain) [Cluster number]*	Half-cystine in coil, bend or turn (chain) [Cluster number]*	Strain energy (KJ/mol)
1KNG	Thioredoxin	92(A)-95(A)	A95(A) [2]	92 (A) [2] beta turn	19.3



1ST9	Thioredoxin	73(A)-76(A)	A76(A)[2] 76(B) [2]	73(A) [2] beta turn	13.0
		73(B)-76(B)		73(B) [2] beta turn	14.7
1THX	Thioredoxin	32(A)-35(A)	35(A) [2]	32(A) [2] beta turn	15.4
1ABA	Thioredoxin	17(A)- 14(A)	17 (A) [2]	14(A) [1] beta turn	12.4
1H75	Thioredoxin	11(A)-14(A)	14 (A) [2]	11(A) [1] beta turn	13.5
1FVK	Thioredoxin	30(A)-33(A),	33(A) [2] 33(B) [2]	30(A)[2] beta turn	19.2
		30(B)-33(B)		30(B) [1] beta turn	14.3
1JR8	Four-helical-up- and-down bundle	57(A) [1]		54(A)[2] gamma turn	19.0

\*[2] Represents Buried-hydrophobic cluster

\*[1] Represents Buried-hydrophilic cluster

#Respective cystines from these proteins are reported as catalytic or active in PDBsum except otherwise mentioned

The redox half-cystines with alpha-helical conformations in the –C-x-x-C- motif were present in the active site region of their respective protein structures with thioredoxin fold and they are enclosed in a small area (represented by ellipsoid) within buried-hydrophobic microenvironment cluster

(Figure. 3.5), with one exception (PDB ID: 1JR8). It has been observed in all instances that the second sulfur from the disulfide belongs to beta turn secondary structures (Table 3.8).

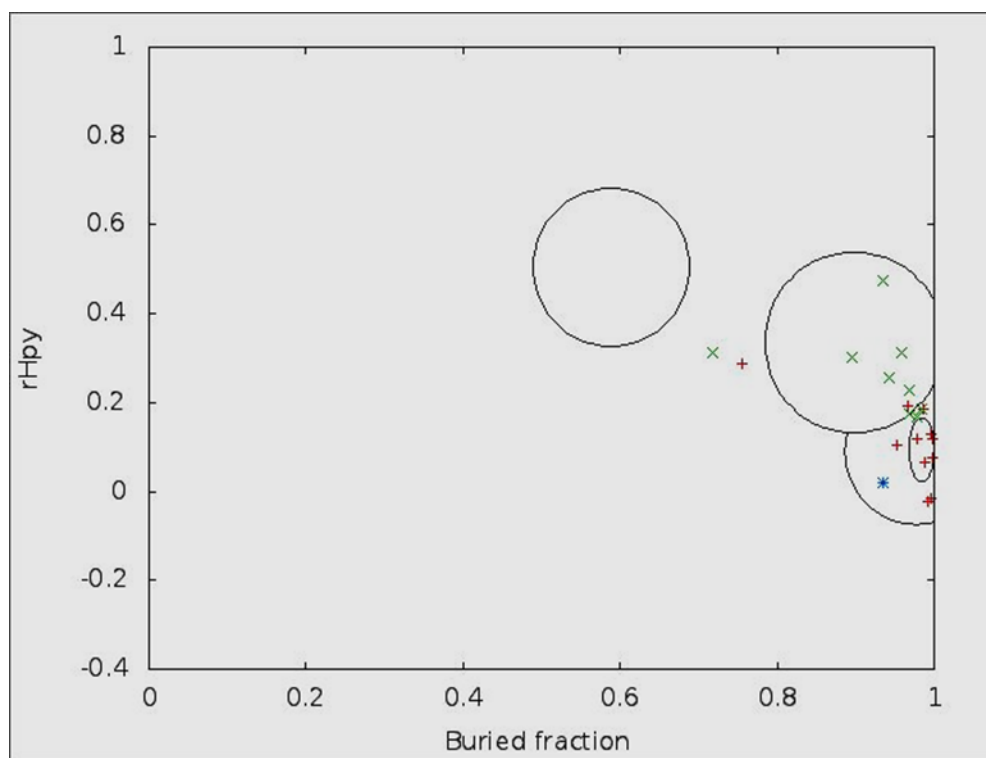


Figure 3.5: Microenvironment of redox-active half-cystines with different secondary structures [alpha-helix (red diamonds), coil (green cross) or turn (blue stars)] depicted in microenvironment (buried fraction, rHpy) space. Three different microenvironment clusters are depicted by three different circles. Centroid positions and average distances of these respective clusters are obtained from Table 3.2. Microenvironments around active half-Cystines with alpha helical geometry are depicted by an ellipsoid. The center of the ellipse consists of average buried fraction and rHpy values (Table 3.9). Minor and the major axes of this ellipse are obtained from two standard deviations ( $2\sigma$ ) values of buried fraction and rHpy.

Table 3.9: Mean and standard deviation values (given in parenthesis) for buried fraction and rHpy of functional half-cystines from different enzyme classes

Enzyme class	Buried fraction	rHpy
Oxidoreductases and electron transport	0.984 (0.016)	0.093 (0.072)

Hydrolase	0.63 (0.044)	0.48 (0.034)
-----------	--------------	--------------

This is to note that the active half-cystine which was not a part of the buried-hydrophobic microenvironment has the second sulfur (other half-cystine) in gamma turn secondary structure (figure 3.6).

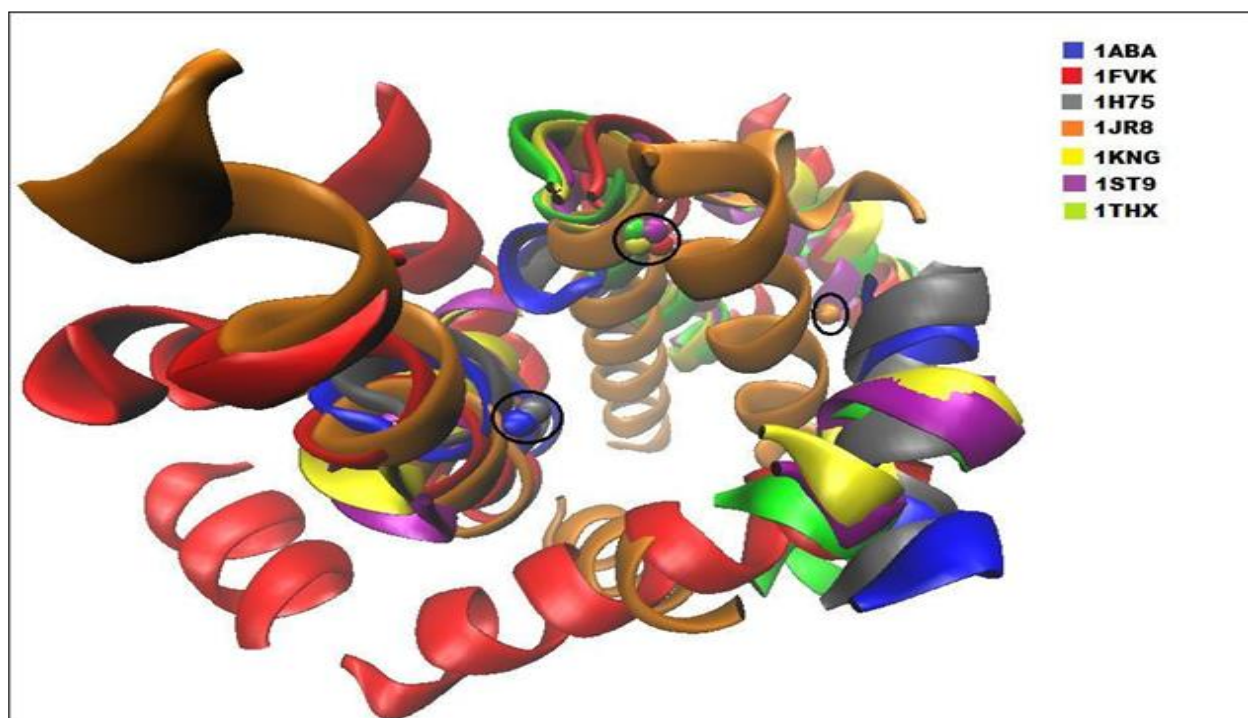


Figure 3.6: Structurally aligned regions of six oxidoreductases enzymes and one electron transport protein, all containing redox active Cystines within C-x-x-C motif. Half-cystines, that is, the single sulfur atom from the disulfide bridges are depicted as spheres from proteins with different color codes.

This outlier cystine belongs to the FAD-dependent sulfhydryl oxidase protein (PDB ID: 1JR8) with four-helical-up-and-down bundle fold, in contrast to thioredoxin fold, in common. Moreover, the remaining cystines that are found in oxidoreductases were not found in the -C-x-x-C- motif. (Table 3.10)

Table 3.10: Cystines present in oxidoreductases, part of the buried-hydrophobic cluster and not found in -C-x-x-C- motif region.

PDB ID	Fold of the protein (SCOP classification)	Cystine residue number (chain)	Half-cystine in Alpha helix (chain) [Cluster no.]*	Half-cystine in Coil/Bend/Turn [cluster no.]*	Strain energy (KJ/mol)
1JR8	Four-helical-up-and-down bundle	83(A)-100(A), 83(B)-100(B)	83 (A) [2]; 83 (B) [2]	100(A) [1] 100 (B) [1]	10.80 10.48
1JM1	ISP domain	145(A)-172(A)		145 (A) [1], 172(A) [1]	8.07
1DJ7	SH3-like barrel	87(A)-A57(A)		87 (A) [2], 57(A) [1]	14.2

\*[2] Represents Buried-hydrophobic cluster

\*[1] Represents Buried-hydrophilic cluster

The oxidoreductase enzyme class contains many folds, but the active redox cystines that are present in alpha helix secondary structure found in oxidoreductases, are present in thioredoxin fold only (Table 3.11).

Table 3.11: Folds present in 11 oxidoreductases and 6 electron transport proteins present in the dataset.

PDB ID	SCOP Class	Fold
1ABA	Alpha and beta proteins (a/b)	Thioredoxin fold
1H75	Alpha and beta proteins (a/b)	Thioredoxin fold

1THX	Alpha and beta proteins (a/b)	Thioredoxin fold
1KNG	Alpha and beta proteins (a/b)	Thioredoxin fold
1ST9	Alpha and beta proteins (a/b)	Thioredoxin fold
1FVK	Alpha and beta proteins (a/b)	Thioredoxin fold
1JR8	All alpha proteins	Four-helical up-and-down bundle
1DJ7	All beta proteins	SH3-like barrel
1GU2	All alpha proteins	Cytochrome c
1UMM	All alpha proteins	Cytochrome c
1FI2	All beta proteins	Double-stranded beta-helix
1D2V	All alpha proteins	Heme-dependent peroxidases
1LYC	All alpha proteins	Heme-dependent peroxidases
1MFM	All beta proteins	Immunoglobulin-like beta-sandwich
1K3I	All beta proteins	Immunoglobulin-like beta-sandwich
1JM1	All beta proteins	ISP domain

Previous work has shown that cystines in the alpha-helical geometry are prone to easy cleavage (De Simone et al., 2006) and participate in enzyme-substrate reaction; from our dataset, we have seen that active redox cystines in the buried-hydrophobic cluster are present in the alpha-helical secondary structure. To further investigate the microenvironment conservation and analyze the similarity of these active redox cystines found in the -C-x-x-C motif clustered into the buried-hydrophobic cluster; the sequence and structure alignment was performed (for redox active cystines in table 3.8, excluding the exception of protein with PDB ID 1JR8).

The global sequence alignment of these six thioredoxin fold containing proteins has shown only 9.5% similarity performed using ClustalW (McWilliam et al., 2013). But the local alignment performed using COBALT online tool (Papadopoulos and Agarwala, 2007) shows the conservation of the –C-x-x-C- motif (figure 3.7) and presumably suggest a conservation of the microenvironment around these cystines.

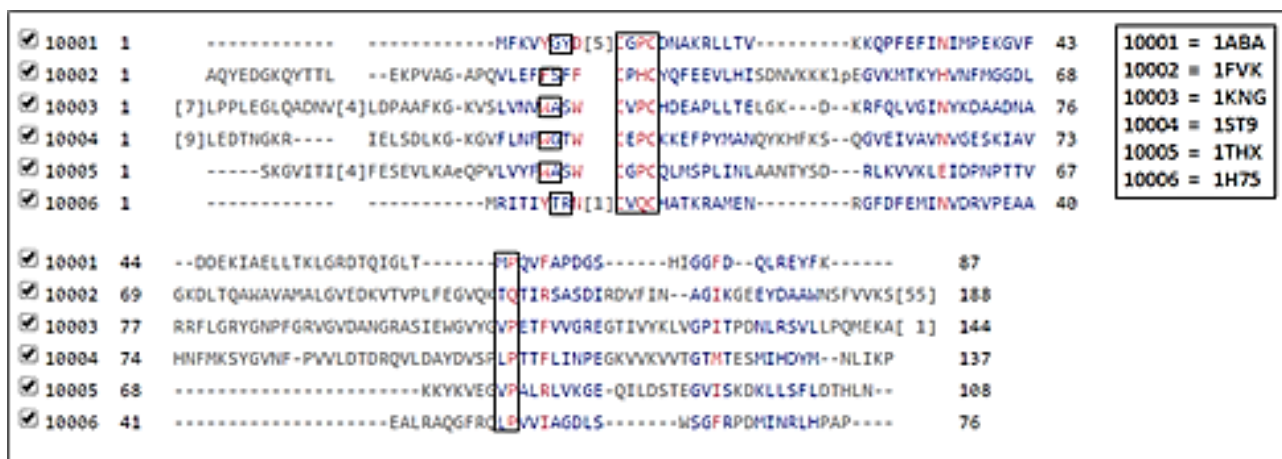


Figure 3.7: Local alignment of six oxidoreductases enzymes (except PDB ID 1JR8) that consists of half-cystines in alpha helix structure. The boxed regions show i) –C-x-x-C- motif and ii) the residues from the microenvironments around those half-Cystines. Local alignment was done using COBALT online tool (Papadopoulos and Agarwala, 2007).

This conservation of microenvironment is further verified by calculating and identifying the neighboring atoms from the respective amino acid residue that lies within the 4.5 region of the cystine disulfide (Table 3.12). It is observed that similar type of neighboring atoms from the neighboring residues are present in the microenvironment region of these cystines.

The local sequence alignment performed (figure 3.7) has also shown conservation of two other regions. These conserved regions are similar to the common active site (WCGH/PCK) in thioredoxin protein family (Holmgren and Björnstedt, 1995), (Martin, 1995), (Peter T. Chivers, Martha C. A. Laboissiere, 1998) , (Mobbs C V, Kaplitt M, 1998).

Table 3.12: Patterns of atomic distribution in all the microenvironments around active half-cystines (all are part of -C-xx-C- motif) with alpha-helical conformations in oxidoreductases and electron transport proteins. Except for 1JR8, all other half-cystines in all these motifs belong to beta turn, whereas, the other half-cystine from -CxxC- motif belongs to gamma turn (shown in italics).

PDB ID	1ABA	1FVK	1H75	1KNG	1ST9	1THX	<i>1JR8</i> (57-54)
Menv	-	-	-	-	68 PHE O	26 TYR OH, HH	<i>11 LYS</i> <i>CE, NZ</i>
	6 GLY CA, C	26 PHE CA, CB	7 THR CA, C, CB, OG	88 TRP CA, C, CB	69 TRP C, CA, CB	28 TRP C, CA, CB	<i>52 TYR</i> <i>CE2</i>
	7 TYR N, O	27 SER N, O	8 ARG N, O	89 ALA N, O	70 GLY N, O	29 ALA N, O, H	<i>56 GLU</i> <i>C, CG</i>
	-	32 HIS C	13 GLN C	94 PRO C	75 PRO C	34 PRO C	
	65 MET C, O, CB	150 VAL C, CB, CG2	51 LEU C, CB	156 VAL C, CB, CG2	139 LEU C, CB	75 VAL C, CB, CG2	
	66 PRO N, CA, CB, CD	151 PRO N, CA, CB, CG, CD	52 PRO N, CA, CB, CG, CD	157 PRO N, CA, CB, CG, CD	140 PRO N, CA, CB, CG, CD	76 PRO N, CA, CB, CG, CD	

To further verify the microenvironment conservation around these redox cystines, structural alignment of the six thioredoxin fold containing proteins was performed using SALIGN (Braberg et al., 2012). The root mean square deviation (RMSD) was measured using SuperPose server (Maiti et al., 2004). The C-alpha RMSD for the aligned six proteins was 6.69 angstroms.

However, the RMSD of the beta-alpha-beta fold of these proteins that contain the redox active cystines is only 1.58 angstroms (figure 3.8).

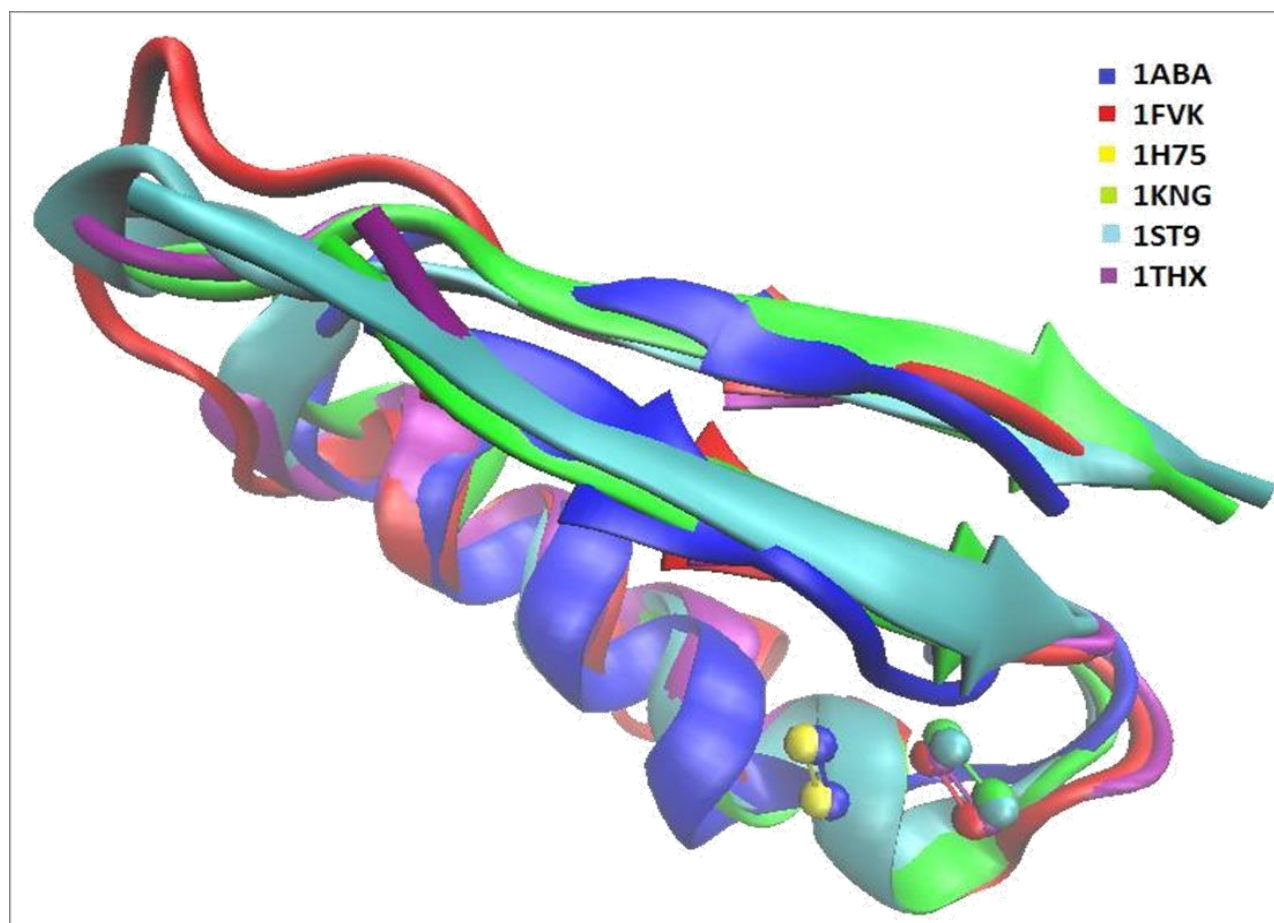


Figure 3.8: Superimposed beta-alpha-beta region from the above six proteins, excluding the enzyme with PDB ID: 1JR8. Half-cystine from different proteins in this aligned regions are depicted as spheres with different color codes.

The superimposition of the cystines shows the RMSD of 0.83 angstroms and the conservation of the -C-x-x-C- motif (figure 3.9).



```

1ABA: -FKVYGYDSNIHK-C-GPCD-----NAKRLLLTVKKQ-----PFEFIN-*
1H75: -ITIIYTRN----D-C-VQCH-----ATKRAMENRGF-----DFEMIN-*
1THX: -VLVYFWA-SWCGPCQLM-S---PL-INLAANTYSDR--LKVVKLE-*
1ST9: GVFLNFWG-TWCEPCKKE-F---PY-MANQYKHFKSQ-GVEIVAVNV*
1KNG: -SLVNVWA-SWCVPCHDE-A---PL-LTELGKD--KR--FQLVGINY*
1FVK: --VLEFFS-FFCPHCYQF-EEVLHI-SDNVKKKL--PEGVKMTKYH-*

```

Figure 3.9: Structural alignment of six oxidoreductase enzymes (except PDB ID 1JR8) containing the redox active cystines in the C-x-x-C motif of their beta-alpha-beta fold. The Conservation of the -C-x-x-C- motif is marked by blue and red boxes.

Although the conservation of this fold region is already known (Branden and Tooze, 1991), here it has been shown that the microenvironment around these redox active cystines is also conserved due to a common fold (figure 3.10). The microenvironment (4.5 Å region) around these redox cystines in the beta-alpha-beta fold was aligned using SALIGN (Braberg et al., 2012). RMSD of the aligned microenvironment around these redox active cystines was 1.10 angstroms.

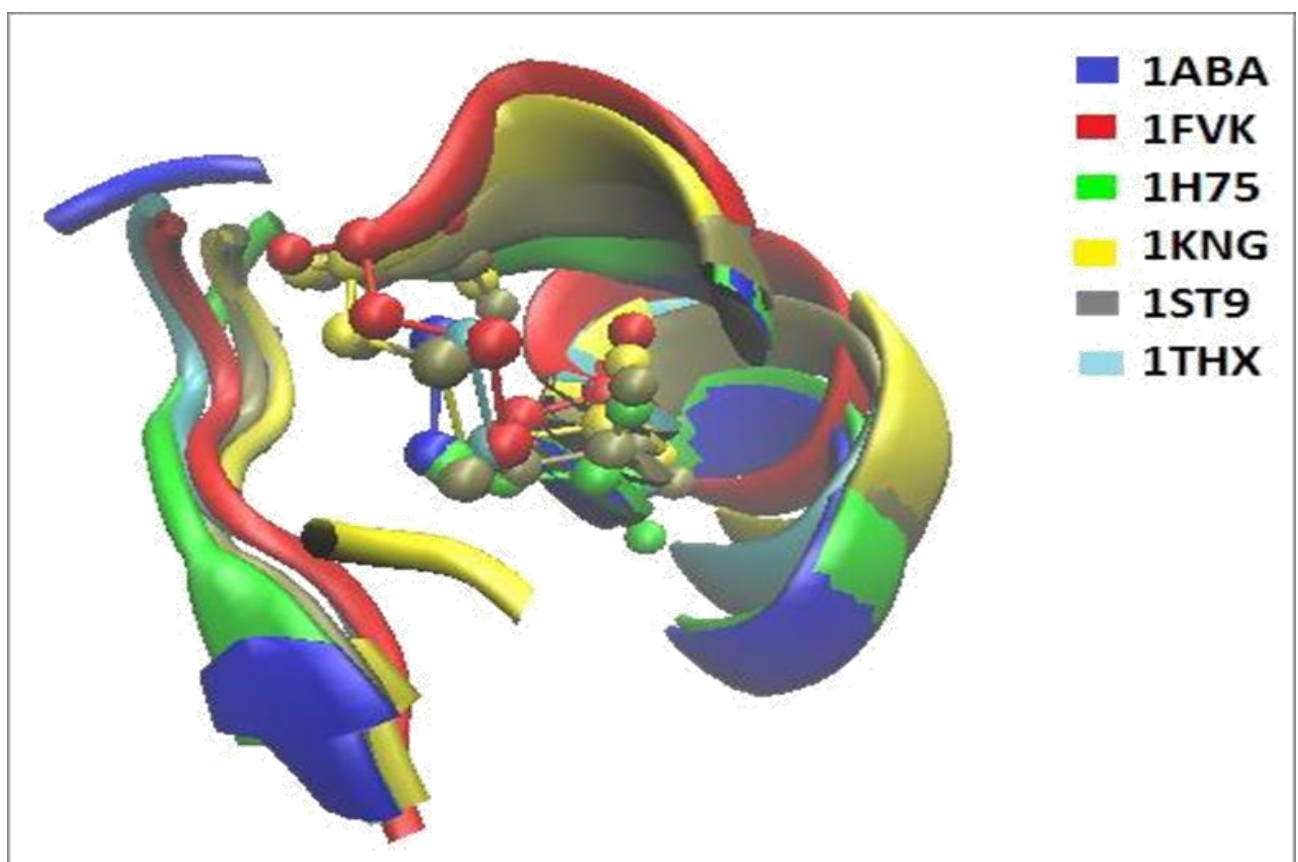


Figure 3.10: Conservation of microenvironment around C-x-x-C motif from 6 proteins having redox-active disulfides in alpha-helical conformations. Microenvironment in each case has alpha-

helical conformation on one side and beta turn on the other side. C-x-x-C motif in all these proteins is encompassed within the cleft created by the alpha helix and beta-turn region.

### 3.3.4) Buried-hydrophilic cluster hosts half-cystines within the microenvironment of enzyme catalytic or Active sites:

Buried-hydrophilic microenvironment cluster contradicts with the intrinsic hydrophobic character of disulfide-bridged cystine residues. The average rHpy value of cystine is 0.21 with one standard deviation value of 0.16 (Bandyopadhyay and Mehler, 2008). There were forty-six half-cystines detected in this cluster with buried fraction values greater than 0.93 and rHpy value greater than 0.4 (rHpy greater than 1 standard deviation value). These half-cystines were indicated here to be embedded in extremely mismatched microenvironment. These extremely mismatched microenvironments around cystines residues, presumably, have evolved to serve specific function of a protein or enzyme. Curation of the functions for all these instances has revealed that thirteen out of forty-six half-cystines belong to the microenvironment of the catalytic site of the enzyme or the active site of the enzyme (Table 3.13).

Table 3.13: Buried hydrophilic microenvironment cluster: Cystines buried within protein interior (buried fraction >0.93) yet embedded in hydrophilic microenvironment (rHpy >0.4). Some of the cystines belong to the microenvironment of catalytic or active sites of the enzyme. Residues shown in bold are present at the active site and those in italics are present at the catalytic site.

S.no	PDB ID	Cystine (chain)		Sequence conservation#		2° structure		Active residue in 4.5 Å of cystine#	Strain energy (KJ/mol)
1	1G66	52 (A)	46 (A)	medium	low	3-10 helix	Coil	<i>T13</i> <b>G47</b> , <b>Q49</b> <i>S50 Q91</i>	17.12
2	1G66	46 (A)	52 (A)	Low	Medium	Coil	3-10 helix	<i>T13</i> <b>G47</b> , <b>Q49</b> <i>S50 Q91</i>	17.12

3	1G66	179 (A)	147 (A)	low	Low	Bend	Coil	<b><i>D175, T146, D172, A173, S174, Y177</i></b>	13.56
4	1GPI	225 (A)	245 (A)	high	high	Beta bridge	Beta bridge	<i>G207</i>	16.27
5	1OC7	419 (A)	372 (A)	high	high	3-10 helix	Bend	<b><i>W371 N405</i></b>	10.48
6	1DY5	65 (A)	72 (A)	high	high	Coil	Beta strand	V63	7.21
7	1K7C	88 (A)	96 (A)	medium	medium	Beta bridge	Beta strand	N137 K140	3.89
8	1KLI	91 (L)	102 (L)	high	high	Turn	Beta strand	E94 S89	16.70
9	1KS8	372 (A)	365 (A)	medium	high	Coil	Coil	Y74 W77	2.42
10	1LLF	60 (A)	97 (A)	high	high	Coil	Coil	R361	4.55
11	1PGS	208 (A)	204 (A)	High	high	Bend	Beta strand	E206	5.81
12	1R0R	16 (I)	35 (I)	high	high	Beta strand	Coil	F37	12.36
13	3LZT	127	6	high	high	Coil	Alpha	R128	5.78

		(A)	(A)				helix		
--	--	-----	-----	--	--	--	-------	--	--

# as per PDBsum

Residues at the catalytic site are described as those which directly participate in catalytic reaction and the residues at active site indirectly participate in catalysis by imparting structural stability. Our analysis has shown that all the cystines in the microenvironment of catalytic residues have high strain energies compared to the average strain energy in this microenvironment cluster (10.52 KJ/mol). Moreover, the cystines within the microenvironment of catalytic or active site residues are structurally less conserved compared to those not part of active or catalytic site residues (Table 3.13). It has been reported earlier that there is no clear relationship between disulfide bond conservation and sequence identity between two proteins from the same family (Thangudu et al., 2008). Despite their low or medium sequence identity in respective protein families, all these cystines share very similar microenvironment. One of the reasons for low sequence identity of these cystine residues is their occurrences in irregular secondary structures (coil, turn etc.) in contrast to cystines in C-x-x-C motif of thioredoxin fold. It has already been shown in the previous section that the cystines in C-x-x-C motifs from oxidoreductase of electron transport proteins mainly occupy alpha-helical structure to facilitate easy cleavage of the disulfide bond, with high strain energy values. These observations together suggest that cystines within the microenvironment of catalytic residues of hydrolase enzymes do not require easy cleavage of the disulfide bond as in thioredoxin folds.

### **3.3.5) Exposed-hydrophilic microenvironment cluster hosts half-cystines as a part of enzyme active sites or catalytic sites**

Physico-chemical properties (in terms of buried fraction and rHpy values) of this microenvironment cluster are most dissimilar with respect to the embedded cystine residue. Average buried fraction and rHpy values for this cluster are 0.590 and 0.505 respectively (Table 3.3), in contrast to 0.92 and 0.21 for cystines (Bandyopadhyay and Mehler, 2008). Evolution of such microenvironments around cystine residues (33 in the current dataset) in many proteins pose the question; whether those

cystines in this particular cluster serve similar functions in their respective proteins. Results show that 29 out of these 33 half-cystines were found in hydrolase enzyme class. (Table 3.6). Enzyme class hydrolase is broad in terms of its function and harbor many enzyme sub-classes with specific functions. These cystines belong to diverse enzyme sub-classes of hydrolases. However, a striking observation is many of the half-cystines with exposed and hydrophilic microenvironments directly participate in enzyme catalysis or stabilize the conformation at the reaction center. There are five instances found where the cystine directly participates in the catalytic reaction of the enzyme. In seven other cases, the cystine is located within the microenvironment of the enzyme catalytic site (Table 3.14). As these half-cystines are partly exposed they are mainly found in flexible conformations, like coil, turn or bend (Table 3.14). Moreover, most of the catalytic residues are found in loop regions of an enzyme. These cystines are not always highly conserved in their respective protein class (Table 3.14). However, their microenvironments are similar. This observation points out that microenvironment is a deciding factor for amino acid function in protein structures.

Table 3.14: Exposed hydrophilic microenvironment cluster – Hydrolase enzyme: Cystines either part of the catalytic or ligand-binding site or part of the microenvironment of catalytic site.

S.no	PDB ID#	Cystine residue no (chain)		Sequence Conservation#		2 <sup>o</sup> structure		Active residue in 4.5 Å of cystine	Active cystine#	Strain energy (KJ/mol)
Cystines part of the catalytic site										
1	1G6X	14 (A)	38 (A)	High	High	Coil	Bend	G37 R39	<b>C14</b> <b>C38</b>	12.72
2	1K7C	232	214	Low	medium	Coil	alpha	K210,	<b>C214</b>	6.34

		(A)	(A)				helix	V213 T215	<b>C232</b>	
3	1LNI	7 (A)	96 (A)	high	Low	Beta strand	Coil	C6 D93	<b>C7</b>	12.02
4	1LNI	96 (B)	7 (B)	Low	High	Coil	Beta strand	T95	<b>C96</b>	10.66
5	1QNR	284 (A)	334 (A)	Low	medium	alpha helix	pi helix	P283 T285 N331	<b>C284</b>	5.50
Cystines within the microenvironment of the catalytic site										
6	1EB6	117 (A)	177 (A)	high	medium	Turn	Coil	H118		29.97
7	1G66	2 (A)	79 (A)	high	high	Coil	Bend	S1 S74		6.00
8	1I71	1 (A)	78 (A)	high	high	Beta strand	Coil	P79		7.48
9	1KNM	119 (A)	100 (A)	high	high	Coil	Beta strand	Y117		13.21
10	1LBU	81 (A)	3 (A)	Medium	medium	Bend	Bend	D80		11.12
11	2NLR	69 (A)	64 (A)	Medium	medium	Beta strand	Beta strand	H69		19.85
12	3LZT	6 (A)	127 (A)	high	high	alpha helix	Coil	R128		5.78

### 3.3.6) Hydrophilic microenvironment destabilizes the disulfide bond by widening the chi3 dihedral angle between the two half-cystines

The 700 cystine pairs were distributed into in 6 microenvironment cluster pairs (Table 3.15). It is known that surrounding microenvironment plays a crucial role in deciphering disulfide bond strength (Trivedi et al., 2009). Oxidizing environment leads to the formation of intermolecular disulfide bonds causing protein aggregation and precipitation in  $\alpha$ -crystallin protein of rabbit eye lens (Bucala et al., 1985). As a single half-cystine cannot explain the modulating effect of microenvironment on cystine structure, hence the microenvironment cluster pairs were studied.

Table 3.15: Average strain energies and average chi3 angle of 700 cystine molecules present in the 6 microenvironment cluster pairs from high-resolution protein crystal structures

Cluster pair	Population	Average Strain energy (KJ/mol)	Average Chi3 angle (degrees)
Buried hydrophobic-Buried-hydrophobic (1-1)	237	10.74	-7.89
Buried hydrophilic-Buried-hydrophilic (2-2)	137	10.41	4.55
Exposed hydrophilic-Exposed-hydrophilic (3-3)	10	10.73	-14.9
Buried hydrophilic-Buried-hydrophobic (2-1)	246	10.25	-7.34
Buried hydrophilic-Exposed-hydrophilic (2-3)	41	12.88	45.16
Buried hydrophobic-Exposed hydrophilic (1-3)	29	9.68	1.613

The average strain energies and the average chi3 dihedral angles calculated for all 700 disulfide bridged cystines shows a bimodal distribution of the Chi3 dihedral angle (Figure 3.11). It is already known that Chi3 angles range around 90-100° and -90 to -100° (Blake et al., 1967; Wyckoff et al., 1970). This support our observation of bimodal distribution of Chi3 dihedral angle to maintain stability.

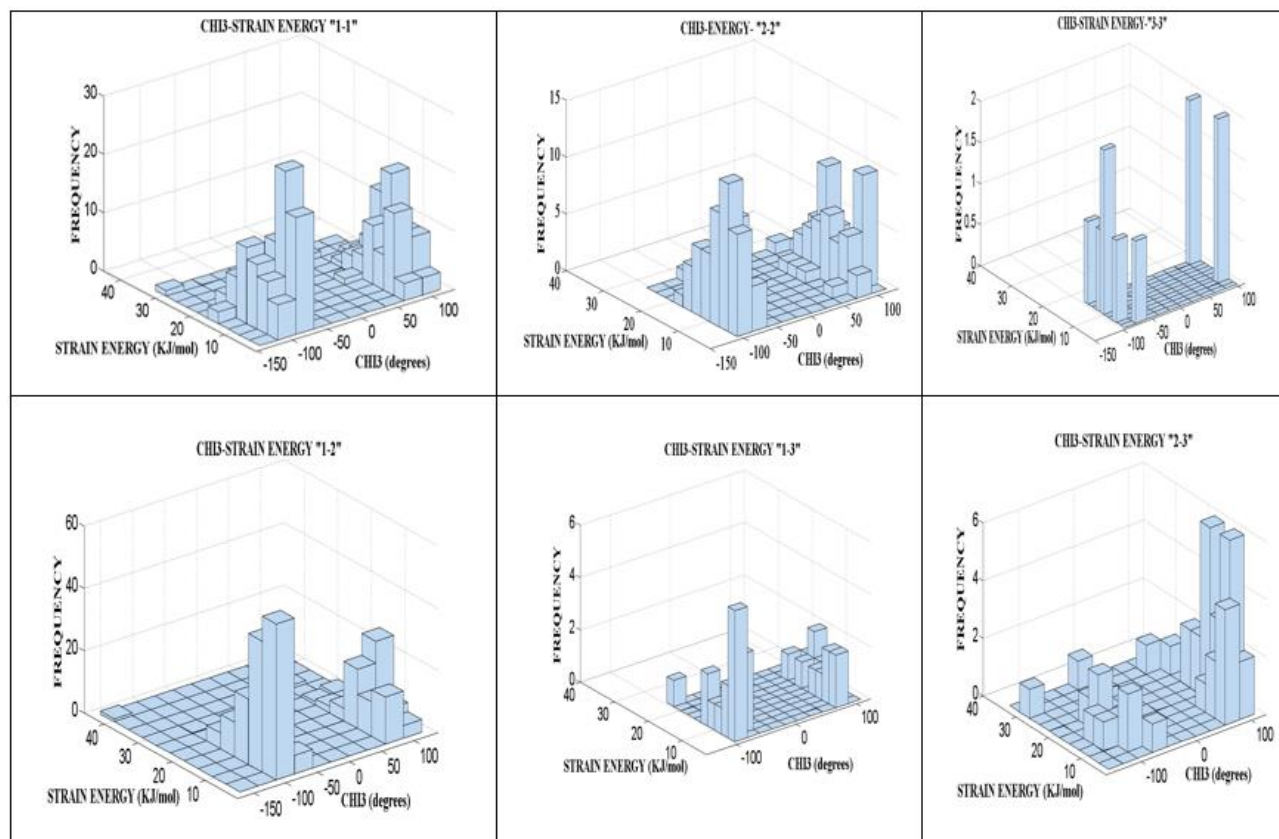


Figure 3.11: Distribution of disulfide in different microenvironment clusters with respect to disulfide structure and strain energies.

The comparative analysis of the bimodal distribution of the average strain energies and chi3 dihedral angle among microenvironment cluster pairs suggest that sulfurs of a disulfide present in different cluster pairs have higher energy than those present in same cluster pair.

To further test the effect of surrounding environment on cystine disulfide bond strength, the representative cystine from each cluster pair was subjected to quantum chemical calculations and the bond order values between the –S-S- bond of the optimized systems were noted (Table 3.16). These cystine bond order values in the 6 systems were compared with systems having cystine



molecule in solvents like water and dimethyl sulfoxide (DMSO). The Molecular Dynamics simulated cystine structure in water and DMSO were selected based on the residence time (defined in the method section of this chapter). The residence time for all the simulations shows that the simulation which used the oscillation time constant as 10000 fs has most points (132) near to the origin. Moreover, the cystine from the frame number 4578 of the same simulation is most close to the origin among all others from all the simulations with a distance of 0.25 angstroms. Similarly, the cystine nearest to the origin dissolved in DMSO solvent was in the frame number 8885 from the simulation with oscillation time constant 20,000 fs. The 4.5 Å microenvironment region around the cystine (along with the cystine molecule) in these two frames was optimized using quantum chemical calculations. The bond order for these two optimized cystines in water and DMSO solvent were used as a reference to compare the solvent properties of the protein microenvironment around the disulfide-bridged cystine molecule (Table 3.16).

Table 3.16: Estimated strengths of the donor-acceptor interactions towards the disulfide bond are shown obtained through Natural Bond Analysis of the crystal structures using 631G\*\* basis set (within parenthesis are the values obtained after optimization of the crystal structures) [BD represents bonding orbital of the disulfide and BD\* represent antibonding orbital of the disulfide].

Solvent	To BD* (Kcal/mol)	From BD (Kcal/mol)	Chi3(-C-S- S-C-) in degrees	Strain energy (KJ/mo l)	Distance (Å)	Bond order
Water#	8.56 (7.07)	9.51 (9.64)	71.65 (110.96)		2.03 (2.056)	0.932
DMSO#	10.14 (10.38)	10.54 (9.88)	102.935 (92.46)		1.999 (2.052)	0.919
Buried-hydrophobic – Buried hydrophobic	7.77 (7.88)	9.35 (9.05)	-106.19 (- 101.73)	11.94	2.050 (2.058)	0.924
Buried-hydrophobic – Buried-hydrophilic	6.80 (6.95)	10.52 (9.98)	67.84 (67.08)	12.44	2.048 (2.060)	0.918
Buried-hydrophilic – Buried-hydrophilic	8.25 (8.47)	6.4 (9.55)	-84.57 (- 84.57)	7.34	2.020 (2.048)	0.929
Exposed-hydrophilic – Exposed- hydrophilic	6.59 (7.16)	8.57 (8.81)	94.81 (96.371)	16.6	2.038 (2.052)	0.948
Buried-hydrophilic –	7.63 (8.29)	9.31 (9.08)	88.91	3.67	2.041	0.919

Exposed-hydrophilic			(89.06)		(2.050)	
Buried-hydrophobic – Exposed-hydrophilic	8.90 (8.31)	12.86 (9.41)	-73.01 (- 74.80)	6.44	2.044 (2.053)	0.925

#These values are obtained by simulating the cystine amino acid (from glutathione protein, PDB ID: 1ABA) in the respective solvent.

The Natural Bond order (NBO) analysis shows that protein microenvironment slightly effects the disulfide bond distance (Table 3.16) but it does have an impact on the chi3 dihedral angle (-C-S-S-C-). In presence of a hydrophilic microenvironment the dihedral angle increases and the disulfide bond widens that may cause dissociation of the bond, and such cystine residues can be attacked readily by the substrate. Whereas, in presence of hydrophobic microenvironment, the dihedral angle reduces and disulfide bond shrinks that may stabilize the bond. The disulfide bond stability is comparatively higher when the two sulfurs of a cystine molecule are found in similar microenvironments as compared two those present in different microenvironments. It is to be noted that, though oxidizing conditions are preferred for disulfide bond formation (Voet et al., 2008) but this study shows that the disulfide bond strength of cystine molecule decreases with increase in hydrophilicity of the surrounding environment, hence it supports the thiol disulfide reversible exchange mechanism (Nagy, 2013). This work shows a different type of donor-acceptor interactions of the cystine disulfide bond in different microenvironments regions.

### 3.4) Conclusions

It was already known that similar folds lead to similar functions in proteins (Mobbs C V, Kaplitt M, 1998) , (Chen and Bahar, 2004), (Vogel et al., 2004), (Holbourn et al., 2008). In this study, we have shown that cystine from different proteins, evolved within similar microenvironments, perform similar functions. Our underlying aim was to test the hypothesis that cystines with similar functions from different proteins will belong to similar microenvironment clusters. To this end, by using hierarchical clustering method, we have identified three different microenvironment clusters: I)

buried-hydrophobic, ii) buried-hydrophilic and exposed-hydrophilic, and correlated these with their structural and enzymatic functions (Figure 3.12).

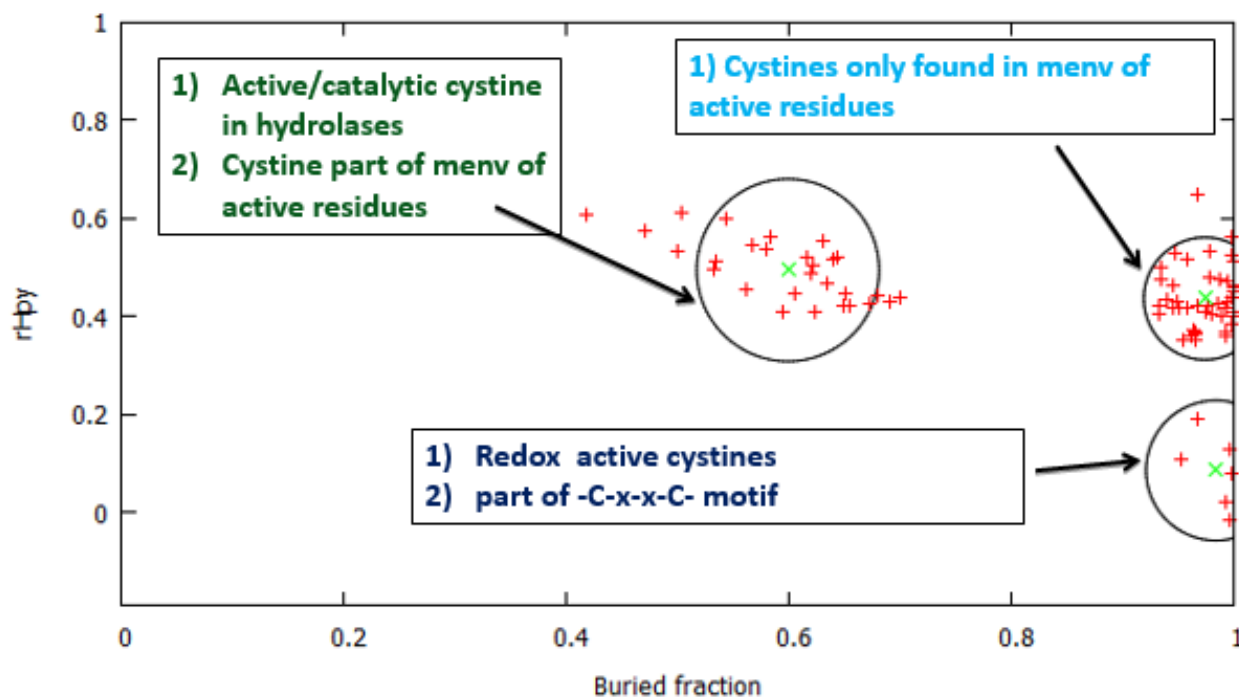


Figure 3.12: Cystines representing the function of three microenvironment clusters. Buried fraction represents the extent of cystine buried inside the protein. rHpy estimates the relative hydrophobicity of the cystine microenvironment region.

In our analysis, special emphasis has been given to enzymatically functional cystines. As demonstrated here, cystines from -C-x-x-C- motifs in Oxidoreductase enzyme class all have very similar microenvironment, that is, buried and hydrophobic. The catalytic cystines from the hydrolase enzyme class always prefer partly exposed hydrophilic microenvironment, not otherwise. Other cystines from hydrolase enzymes which participate in stabilizing catalytic or active sites were mainly found in the hydrophilic microenvironment, either buried or partly exposed. Despite low to medium sequence conservation of cystines from hydrolase enzymes they belong to similar microenvironments, according to their functions. This report illustrates that irrespective of sequence conservation or specific functions of proteins, cystine residue function in similar ways when embedded in the similar microenvironment, thus validating the working hypothesis.

We believe this conclusion should be further verified for other amino acids, particularly, titratable amino acids, like Aspartic acid, Glutamic acid, Arginine etc. Titratable amino acids are expected to be more sensitive towards change in microenvironment due to alteration in their protonation states. Hence, microenvironment modulated switching of protonation states in titratable amino acids would be of interest in major biochemical reactions, like photosynthesis. For example, carbon dioxide fixing enzyme, Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCo), involves seven charged residues in its active site. It is known that switching some of these protonation states could lower the transition state, thus assist to prevent backward reaction and trap more carbon dioxide (Collings and Critchley, 2007). This can be further verified by modifying the protein structures with altered protonation states of charged residues and test the capacity of carbon dioxide fixation.

Amino acids in heterogeneous protein microenvironments are analogous to amino acids in different solvents with variable dielectric media. As solubility, bond dissociation energy and spectral properties of cystine vary from hydrophilic to hydrophobic solvents; the same is also expected when cystine is transferred from hydrophilic part of the protein microenvironment to its hydrophobic part. This could be exploited to guide experimentation into the local dielectric medium within protein microenvironments. We have shown here that disulfide-bridged cystine molecule is a good model system to examine the effect of various dielectric medium on S-S bond dissociation energy and can be extended for experimental verifications.

## CHAPTER 4

### Effect of protein microenvironment on the protonation state of aspartic acid side chain using quantum chemical calculations

#### 4.1) Introduction

Functions of titrable amino acids depend on the protonation state of their side chain (Chen et al., 2009), (Petukh et al., 2013), (Díaz and Suárez, 2016). Changes in these protonation states of amino acids may lead to conformational changes in proteins affecting their functions (Huang et al., 2016). Protonation states alter due to alteration in the pKa of a particular functional group (Nielsen, 2007). pKa alteration largely depends on the electrostatic behavior of the surrounding environment (Mehler and Guarnieri, 1999; Mehler et al., 2002). The electrostatic factor is a crucial component in receptor-ligand complexes during cell signaling (Voet et al., 2008), protein-protein interactions (Norel et al., 2001), protein-DNA interactions (Privalov et al., 2011). For example, Binding of the ER-alpha nuclear receptor to its DNA response element known as the estrogen response element is facilitated by the alteration in the protonation states of the histidine 196 and glutamic acid 203 residues (Deegan et al., 2010). The conservation of such polar residues in the nuclear transport family proteins indicates the switching role of protonation state in maintaining protein-DNA interactions (Deegan et al., 2010). Exchange of protons marks the basis of enzyme catalysis chemistry (Greene et al., 2015).

Protonation state of any amino acid depends on the interaction energy between the two respective ionizing forms. According to coulombs law, this interaction energy is inversely proportional to the dielectric value of the medium (equation 5)

$$E = (q_1q_2)/4\pi\epsilon_0\epsilon_r d^2 \dots\dots\dots\text{Equation 5 (Coulombs law)}$$

Where, E is the energy, q1 and q2 are the charges on the two atoms (in units of the electronic charge), d is the distance between the two atoms (in angstroms), e is the dielectric constant (which

accounts for the effects of the intervening medium), and the value of  $\epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2/\text{Nm}^2$  (Berg et al., 2002). Dielectric constant is related to solvent polarity. The Higher dielectric value represents high polarity like water ( $\epsilon=80$ ), whereas lower dielectric values relate to the weak polarization of the ions like vacuum ( $\epsilon=1$ ). Dielectric constant is an intrinsic property of a solvent. It is the ability of the solvent to keep two ionic species apart from each other. Protein structure contains a combination of amino acid side chains. This leads to a very heterogeneous nature of the protein microenvironment in terms of hydrophobicity. This chapter aims to understand the effect of this heterogeneous protein microenvironment on the protonation state of aspartic acid by comparing the protein microenvironment with solvents having different dielectric constants ranging from 0 to 80.

It is known that the dielectric constant value is indirectly related to the pKa of a titrable amino acid in proteins (Takashima and Schwan, 1965). In 1938 Kirkwood and the group also gave the relation between solvent dielectric constant and pKa of an acid (equation 6) (Kirkwood and Westheimer, 1938; Westheimer and Shookhoff, 1939).

$$\log (K_1 / \sigma K_2) = e^2 / 2.303kTRD \dots\dots\dots \text{(Equation 6)}$$

In equation 6,  $K_1$  and  $K_2$  are the first and second dissociation constants of the acid, “e” is the electronic charge, “k” is the Boltzmann constant, “T” is the absolute temperature, “R” is the interprotonic distance. “ $\sigma$ ” the statistical factor. “D” is the dielectric constant of the solvent. Although, this was a preliminary development by the group, but it indicates towards the hidden relationship between pKa of amino acids and protein dielectric constant.

Protein interior is highly heterogeneous in terms of electrostatic behavior, therefore a gross range of dielectric constant value was empirically decided (4 to 8) (Tanford and Roxby, 1972; Haque et al., 2000; Harris and Turner, 2002). However, local dielectric medium inside the protein is extremely challenging to determine. Although perturbations in amino acid side chain pKa are due to electrostatic effects, but the correlation between the pKa and dielectric medium is less understood (Tanford and Roxby, 1972). Dielectric medium plays a major role in the stability of protein

structures by formation and dissociation of different types of bonds, mainly hydrogen bonds (Berg et al., 2002), (Landsteiner, 2013). Hence it is of immense academic interest to understand the diversity of dielectric medium inside the protein.

There are few reports that suggests that the protonation state of an amino acid side chain is dependent on the heterogeneous protein dielectric medium. For example, the relationship of the protonation states of the titrable residues of the lumen of *Escherichia coli* with the surrounding pH showed that the protonation state of D127 residue of the native OmpF protein varies with change in the dielectric constant values (Varma and Jakobsson, 2004). It has been shown earlier that electrostatic forces overpower the impact of the non-electrostatic interaction in deciding the protonation states of the amino acid (Rao et al., 2011). The protonation constants of L-methionine were studied pH metrically using the computer program “MINIQUAD75” and the linear variation of step-wise protonation constants ( $\log K$ ) with the reciprocal of the dielectric constant of the solvent mixture suggests the dominance of the electrostatic forces (Rao et al., 2011). However, these reports do not directly compare the effect of the dielectric constant of a solvent and that of the protein heterogeneous microenvironment.

In this work, we have tried to study the effect of varying dielectric medium (using various implicit solvents) on deprotonation of the aspartic acid side chain. Aspartic acid was chosen as model system because i) it has relatively small side chain ii) pKa value of aspartic acid display a wide range from 1.0 to 8.0, with default of 3.7 (data shown below), iii) aspartic acid was observed with second highest propensity at catalytic centers, particularly in enzyme sub classes like, aspartate proteases, serines proteases, aspartate transferases etc. (Gutteridge and Thornton, 2005). Deprotonation of carboxylic group in aspartic acid was measured using quantum chemical calculations, in terms electron density in the bonding orbital of “O-H” group carboxylic acid. This parameter was compared across aspartic acid in 16 different implicit solvents (including vacuum) and two proteins. Deprotonation of carboxylic side chain was affected most significantly within the



dielectric range from 1 to 10. Effect of dielectric medium on aspartic acid as a part of protein structure was discussed in the following sections.

Aspartic acid side chain exists in the deprotonated form in some proteins, for example; D52 in Hen-egg-white lysozyme (Refaee et al., 2003), whereas in protonated form in other proteins, for example; bacteriorhodopsin (ASP 96) (Patzelt et al., 2002).

Here, we have attempted to study the role of dielectric medium on the protonation state of aspartic side chain carboxylic acid group in presence of sixteen model systems using quantum chemical calculations. The proton dissociation of carboxylic group steeply decreases in low dielectric range ( $\epsilon$  ranges from 1 to 9), that is the protein hydrophobic core and remain invariant in higher dielectric medium ( $\epsilon > 20$ ). Two aspartic acids embedded in protein local microenvironment were also studied. The bond order of aspartic acid side chain carboxylic “O-H” bond in different solvents with varying dielectric constants was compared with bond of aspartic acid side chain carboxylic “O-H” bond in different proteins to estimate the dielectric nature of these protein microenvironments.

#### **4.2) Methodology**

The side chain carboxylic group of aspartic acid was used as the model system. The effect of different dielectric constants on aspartic acid side protonation state was studied. Molden software was used to create an aspartic acid molecule. The N- and C- terminals of the aspartic acid were capped by acetylation (addition of acetyl group to the N-terminal) and amidation (addition of the amino group to C-terminal) (Figure 4.1). This capping mimics the aspartic acid main chain arrangement in protein structures.

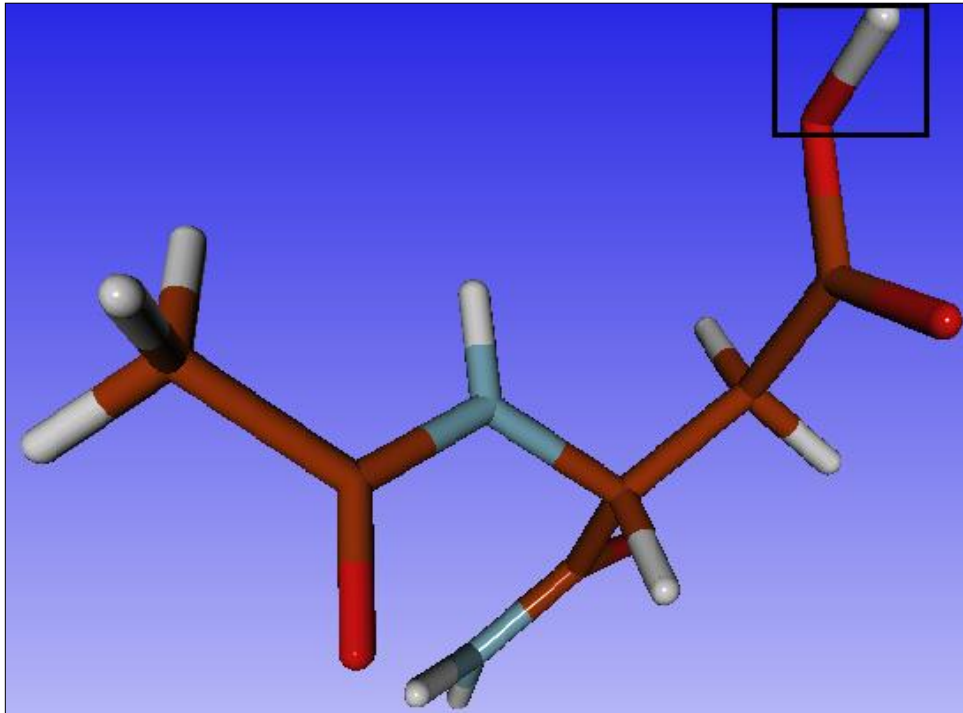


Figure 4.1: Capped aspartic acid molecule. The side chain carboxylic “O-H” bond is highlighted within a box. The C-terminal is amidated by addition of -NH<sub>2</sub> group and N-terminal acetylated by addition of -COCH<sub>3</sub> group.

The capped aspartic acid (ASPN) was then subjected to Molecular Dynamics Simulation in TIP3 (Mark and Nilsson, 2001) water box of 10 Å side lengths for 10 nanoseconds using NAMD (Phillips et al., 2005). The force field parameters were taken from CHARMM topology (top\_all27\_lipid.inp) and parameter (par\_all27\_prot\_lipid.inp) files (Brooks et al., 2009). The simulation was optimized for the water solvated capped aspartic acid molecule based on the varying oscillation period that dictates the cell volume oscillation (LangevinPistonPeriod). This value was kept 10,000 steps and 50,000 steps.

#### 4.2.1) Minimization

The minimization of the water solvated capped aspartic acid was performed for 100 steps using conjugate gradient method. The periodic boundary conditions were set at 10 angstroms along x, y and z directions. The minimization was done in NPT conditions (constant pressure control and variable volume) using Langevin Dynamics.

## **4.2.2) Molecular Dynamics simulation of a protonated aspartic acid side chain**

### **4.2.2.1) Equilibration**

The equilibration was performed for 10ns in NPT conditions. The Langevin Piston pressure was taken as 1.01325 bar (1 atm.). The time step was 1fs/step. To mimic the in-vivo conditions the temperature was decided as the normal physiological temperature, that is, 300K.

### **4.2.2.2) Production run**

The production was followed after the equilibration but in the production run, both the pressure and volume were allowed to change (NVE ensemble) during the course of simulation which ran for 10 nanoseconds. The periodic boundary conditions were again set at 10 angstroms in x, y and z directions. The trajectory file was punched after every 1 picosecond (1000 steps).

The trajectory after the production run of 10 nanoseconds for the oscillation period of 10,000 fs and 50,000 fs was analyzed using VMD to find out the aspartic acid at the center of TIP3 water box. Oscillation period is the time in femtoseconds that represents the barostat oscillation time scale for Langevin piston method. It is the angular period with which the cell volume oscillates in case the decay time is infinite (that is; there is no friction).

This analysis was automated using a combination of TCL and PERL scripts. The Tcl script calculates the coordinates of all the frames in the trajectory file and prints them on to a new text file. The PERL script calculates the distances of all the frames between the origin (0, 0, and 0) and the center of capped aspartic acid. The frame with aspartic acid nearest to the center was used for quantum chemical calculations.

### **4.2.3) Quantum chemical calculations on aspartic acid in presence of 16 different implicit solvents with varying dielectric constants.**

The capped aspartic acid obtained from the previous step was subjected to quantum chemical calculations using Polarizable Continuum Model (PCM) (Mennucci et al., 2002), an implicit solvent model. 16 such models were used having different dielectric constants (Table 4.1) varying from 1.0

(vacuum) to 80.4 (water). GAMESS software was employed for the calculation (Schmidt et al., 1993). Quantum chemical calculations were performed using Restricted Hartree-Fock, Density Functional Theory (B3LYP) and MP2 methods. The basis set used was 6-31G (d, p) for the entire system except for the side chain hydrogen bonded oxygen atom of the aspartic acid residue; additional diffusion function was also employed to this oxygen atom. The implicit solvent model used here resembles the explicit solvent system in terms of average properties of polarity, force constants etc., however, differs in terms of few properties, like hydrogen bonding capability etc. Calculations with explicit solvents are expected to be more accurate but are beyond the scope of this thesis.

Table 4.1: Sixteen different types of implicit solvents used in quantum chemical calculations of capped neutral aspartic acid molecules. The dielectric constant values are also shown (Lide, 2004).

S.no	Solvent	Dielectric constant
1	Water	80.1
2	DMSO	46.7
3	Nitromethane	35.87
4	Methanol	32.7
5	Ethanol	24.5
6	Acetone	20.7
7	1,2-dichloroethane	8.93
8	THF	7.58
9	Aniline	6.89
10	Cl-benzene	5.62
11	Chloroform	4.81
12	Toluene	2.38
13	Benzene	2.27

14	CCl <sub>4</sub>	2.24
15	Cyclohexane	2.02
16	Vacuum	1.00

#### **4.2.4) Setting up a system for weak acid (aspartic acid) and weak base (water) interaction.**

A system of capped aspartic acid interacting with single water molecule was created using molden software (Schaftenaar and Noordik, 2000). The distance between the oxygen atom of the water molecule and the hydrogen atom of the carboxylic group was maintained at 2 Å distance so that there could be no apparent bond formation between these two moieties. This system was subjected to optimization with Hartree-Fock, DFT (B3LYP) and MP2 methods using basis set of 6-31G (d, p). Additional diffuse functions were added to both the oxygen atoms of carboxylic group and water oxygen atom. The optimization was performed in vacuum.

#### **4.2.5) Setting up a system for weak acid (aspartic acid) and strong base (sodium hydroxide) interaction.**

A system of capped aspartic acid interacting with single sodium hydroxide molecule was prepared. Basis set and optimization methods were identical to the aspartic acid – single water system.

#### **4.2.6) Integrated Quantum mechanics and Molecular Mechanics (QM/MM) approach to study deprotonation of aspartic acid in proteins**

QM/MM method was applied on two protein NMR structures, bacteriorhodopsin protein (PDB ID: 1R2N) and toxin protein (PDB ID: 1ORL), downloaded from Protein Databank (PDB) (Berman et al., 2000). Aspartic acid 96 was found in protonated form in the bacteriorhodopsin protein. To study the electronic distribution, the side chain part of the aspartic acid 96 was subjected as QM region and the remaining part of the protein was subjected as MM region. QM/MM calculation was performed using tinker QM/MM executable (Ponder and Richards, 1987; Kundrot et al., 1991) linked with GAMESS (Schmidt et al., 1993). Molden software (Schaftenaar and Noordik, 2000)

was used to generate the QM/MM tinker format files for bacteriorhodopsin protein. This tinker format file was used to perform QM/MM calculations with the QM atoms in the \$DATA group and MM atoms in \$TINXYZ group using Avogadro software (Hanwell et al., 2012). Due to computational memory constraints, the MM region consists of the residues within in 4.5 angstrom microenvironment region of the protonated aspartic acid at 96 position. Similar type of system was generated around D43 residue in toxin protein (PDB ID: 1ORL) and QM/MM calculations were performed.

#### **4.2.6.1) Selection of dummy atoms marking the boundary of the QM/MM region**

It has been reported earlier that for an efficient QM/MM calculation, the bond cleaved to mark the boundary atoms of the QM/MM interface should not be a polar bond (Thellamurege et al., 2013). Therefore; the C $\alpha$ -C $\beta$  bond was used to create and define the boundary of the QM and MM region. Hence the QM part contains the side chain atoms (-CH<sub>2</sub>COOH) where the dummy atom replaced by a hydrogen atom is having the coordinates of the C $\alpha$  atom. Rest of the protein part is subjected to MM calculations. The atoms in the QM region are defined by providing the QM atom numbers within the \$LINK group in the input file. Morokuma's MOMM scheme was used (Maseras and Morokuma, 1995). The protein structures subjected to QM/MM calculation were optimized using 6-31G(d, p) basis set and to maintain consistency all three methods; Hartree-Fock, Density Functional Theory and Møller Plesset (MP2) calculations were performed for both the protein structures. The bond order for both the proteins was noted and compared against that noted for other 16 implicit solvents to understand the dielectric behavior of protein solvents.

#### **4.2.7) Quantum chemical calculations on the protonated aspartic acid residues in bacteriorhodopsin and toxin proteins to calculate the bond order of side chain carboxylic group in the two protein microenvironments**

The 4.5 microenvironment region around the ASP96 and ASP43 was extracted from the bacteriorhodopsin (PDB ID: 1R2N) and toxin protein (PDB ID: 1ORL) respectively. The microenvironment region in bacteriorhodopsin protein predominantly contains the atoms from

residues 92 to 100, therefore to maintain an ordered structure the stretch from residues 92 to 100 was selected to perform quantum chemical calculations. Similarly, in the case of toxin protein the residues ranging from 41 to 45 were used for quantum chemical calculations. The generated systems were capped at the terminal positions with requisite acetylation at N-terminus and amidation at C-terminus using molden software (Schaftenaar and Noordik, 2000). After capping, the two protein solvated systems were subjected to Hartree-Fock calculations using 6-31G (d, p+) basis set. The method used for optimization of the system was Restricted Hartree-Fock method.

### **4.3) Results and Discussion**

#### **4.3.1) Effects of implicit solvents on side chain deprotonation of aspartic acid**

##### **4.3.1.1) Measurement of side chain carboxylic “O-H” bond properties when immersed in 16 different solvents.**

The capped aspartic acid molecule was studied in 16 implicit solvent models as listed in Table 4.1. These 16 systems were subjected to quantum chemical calculations using GAMESS software (Schmidt et al., 1993) to monitor the change in the side chain protonation state of the carboxylic “O-H” bond with varying dielectric constant values. The properties studied for this side chain carboxylic “O-H” bond were: 1) bond distance 2) bond order 3) the electronic distribution over the respective oxygen and hydrogen atoms 4) energy of the system

Quantum chemical optimization calculations attempt to determine a geometric arrangement of atoms of the system, so as to attain minimum energy state and hence the maximum stable conformation. The number of steps to achieve the optimization directly relates to the number of intermediate structures to reach an energy-minimized state. This process of optimization to locate the equilibrium geometry was done using Newton-Raphson algorithm (Filippone et al., 2001). Based on these optimized structures the bond order and bond distance of the aspartic acid side chain carboxylic “O-H” bond were calculated. These bond order, bond distance, and the optimized energy values are dependent on the electron density distributed over the respective orbitals. This charge

distribution analysis is performed by “Mulliken and Lowdin population analysis” (Reed et al., 1985). To measure these properties (bond order, bond distance and optimized energy of the system), different types of electron correlation methods were applied: (1) Hartree-Fock (2) Density Functional Theory (B3LYP) (3) Møller-Plesset (MP2).

#### 4.3.1.2) Analysis of side chain carboxylic “O-H” bond using Hartree-Fock method

The capped aspartic acid in 16 implicit solvent models, when optimized using Hartree-Fock method, shows that the optimization of the system in the range of 56 (chloro-benzene) to 71 (water) steps for all the solvent systems (Table 4.2). A similar trend was observed with total optimized energies of the system (Table 4.2).

Table 4.2: The number of steps and total energy for each optimization of capped aspartic acid molecule in 16 different solvents using Hartree-Fock method.

Solvent	Dielectric constant value	Number of steps of optimization	Energy (Hartree)
Water	80.1	71	-641.504
DMSO	46.7	57	-641.503
Nitromethane	35.87	58	-641.503
Methanol	32.7	58	-641.503
Ethanol	24.5	57	-641.503
Acetone	20.7	57	-641.502
1,2-dichloroethane	8.93	57	-641.501
THF	7.58	58	-641.500
Aniline	6.89	57	-641.500
Cl-benzene	5.62	56	-641.499



Chloroform	4.81	57	-641.498
Toluene	2.38	59	-641.492
Benzene	2.27	57	-641.491
CCl <sub>4</sub>	2.24	58	-641.491
Cyclohexane	2.02	59	-641.490
Vacuum	1.00	69	-641.477

To check the variation in the aspartic acid side chain carboxylic “O-H” bond, the bond order, bond distance, and electronic distribution on the respective oxygen and hydrogen atoms for Hartree-Fock method were calculated based on the optimized structures from the Bond order and valence bond population analysis (Table 4.3).

Table 4.3: Variation in bond lengths, bond order, bond distance and valence electron distribution on oxygen and hydrogen atoms of carboxylic “O-H” in aspartic acid side chain in various implicit solvents using Hartree-Fock method. The Initial bond length was 0.95886 angstroms. The valence electron (O) and valence electron (H) shows the distribution of electrons in the valence shell of the oxygen and hydrogen atoms respectively.

Solvent	Dielectric constant	Final Bond distance (side chain carboxylic “O-H” bond) (Å)	Bond order	Valence electron (O)	Valence electron (H)
Water	80.1	0.951	0.793	1.771	0.807
DMSO	46.7	0.951	0.794	1.771	0.808
Nitromethane	35.87	0.951	0.794	1.771	0.808

Methanol	32.7	0.951	0.794	1.771	0.808
Ethanol	24.5	0.951	0.794	1.771	0.808
Acetone	20.7	0.951	0.794	1.771	0.808
1,2-dichloroethane	8.93	0.951	0.795	1.771	0.810
THF	7.58	0.950	0.796	1.770	0.811
Aniline	6.89	0.950	0.796	1.770	0.811
Cl-benzene	5.62	0.950	0.797	1.770	0.812
Chloroform	4.81	0.950	0.798	1.770	0.812
Toluene	2.38	0.950	0.803	1.768	0.818
Benzene	2.27	0.950	0.803	1.768	0.819
CCl <sub>4</sub>	2.24	0.950	0.803	1.768	0.819
Cyclohexane	2.02	0.949	0.804	1.769	0.820
Vacuum	1.00	0.948	0.815	1.768	0.832

The valence electrons (O) and valence electrons (H) is the electron distribution in the valence shell of respective oxygen and hydrogen atoms. These values indicate the strength of the “O-H” bond in a given implicit solvent. The bond order and the valence electron (O) and valence electron (H) values in different solvents were comparable to each other (Table 4.3). The slightest of variation occurs when there is a change in the range of dielectric constant values. Based on this observation, we performed density functional theory analysis and MP2 calculations on the Hartree-Fock optimized structures.

#### 4.3.1.3) Analysis of side chain carboxylic “O-H” bond using Density Functional Theory

Density functional theory provides an improvement in Hartree-Fock method by involving the electron-electron correlation terms (Cramer, 2005). The hybrid functional B3LYP (Becke, 3-parameter, Lee-Yang-Parr) was applied to optimized the 16 solvent systems with capped aspartic acid molecule (Kim and Jordan, 1994). As the structures were already optimized with Hartree-Fock

method, therefore, these optimizations finished with a lesser number of iterations and more stable structures (Table 4.4).

Table 4.4: The number of steps and total energy for each optimization of capped aspartic acid molecule in 16 different solvents using Density Functional theory (B3LYP).

Solvent	Dielectric constant value	Number of steps	Energy (Hartree)
Water	80.1	28	-644.817
DMSO	46.7	30	-644.817
Nitromethane	35.87	29	-644.817
Methanol	32.7	29	-644.817
Ethanol	24.5	30	-644.817
Acetone	20.7	29	-644.817
1,2-dichloroethane	8.93	30	-644.815
THF	7.58	30	-644.814
Aniline	6.89	31	-644.814
Cl-benzene	5.62	31	-644.813
Chloroform	4.81	29	-644.813
Toluene	2.38	31	-644.808
Benzene	2.27	31	-644.807
CCl <sub>4</sub>	2.24	29	-644.807
Cyclohexane	2.02	31	-644.806
Vacuum	1.00	46	-644.796

Similar to the Hartree-Fock method, the optimization of the capped aspartic acid in vacuum has taken comparatively more steps than other solvent systems. The optimized structures were used for calculating the bond order and electron distribution on the oxygen and hydrogen atoms of the aspartic acid side chain carboxylic “O-H” bond (Table 4.5).

Table 4.5: Variation in bond lengths, bond order, bond distance and valence electron distribution on oxygen and hydrogen atoms of carboxylic “O-H” in aspartic acid side chain in various implicit solvents using Density Functional Theory (B3LYP). The Initial bond length was 0.959 angstroms.

Solvent	Dielectric constant	Final Bond distance (side chain carboxylic “O-H” bond) (Å)	Bond order	Valence electron (O)	Valence electron (H)
Water	80.1	0.974	0.792	1.904	0.821
DMSO	46.7	0.974	0.792	1.903	0.821
Nitromethane	35.87	0.974	0.792	1.903	0.821
Methanol	32.7	0.974	0.792	1.903	0.822
Ethanol	24.5	0.974	0.793	1.903	0.822
Acetone	20.7	0.974	0.793	1.903	0.822
1,2-dichloroethane	8.93	0.974	0.794	1.903	0.823
THF	7.58	0.974	0.795	1.903	0.824
Aniline	6.89	0.974	0.795	1.903	0.825
Cl-benzene	5.62	0.974	0.796	1.903	0.826
Chloroform	4.81	0.974	0.796	1.902	0.827
Toluene	2.38	0.973	0.801	1.902	0.832
Benzene	2.27	0.973	0.802	1.902	0.833
CCl4	2.24	0.973	0.802	1.902	0.833

Cyclohexane	2.02	0.973	0.803	1.902	0.834
Vacuum	1.00	0.972	0.813	1.906	0.847

The bond order and bond distance of the aspartic acid side chain carboxylic “O-H” bond in different solvent systems shows that with a decrease in dielectric constant values the bond order increases and bond distance decreases. Although, it is important to note that this trend is observed when different dielectric constant values are analyzed together, to define a typical dielectric region. For example; from table 4.5, the bond order value of 0.792 is same in water, DMSO, nitromethane and methanol. Whereas, only in ethanol and in acetone bond order is 0.793. In Tetrahydrofuran (THF) and aniline, bond order value is 0.794. Bond order in benzene and CCl<sub>4</sub> is 0.802 (Table 4.5). This slight variation in the bond order of aspartic acid side carboxylic group indicates that aspartic acid side chain protonation state is affected when there is a larger change in the dielectric constant value of the surroundings. The Density Functional Theory calculation have improved from the analysis from the Hartree-Fock method. This is due to the addition of electron-electron correlation terms. Another method which also implements the electron-electron correlation terms is the Møller-Plesset (MP2) method.

#### **4.3.1.4) Analysis of side chain carboxylic “O-H” bond using Møller-Plesset (MP2) method**

The second order Møller-Plesset (MP2) method was applied on the Hartree-Fock optimized systems (Møller and Plesset, 1934). As compared to density functional theory, MP2 calculations are much slower. Although, to perform a full comparison of the dielectric values and the corresponding bond order, these calculations were performed for all the solvent systems. Similar to DFT analysis, the number of steps required for each optimization was decreased (Table 4.6), although an increase in the energy values was observed.

Table 4.6: The number of steps and total energy for each optimization of capped aspartic acid molecule in 16 different solvents using Møller-Plesset (MP2) method.

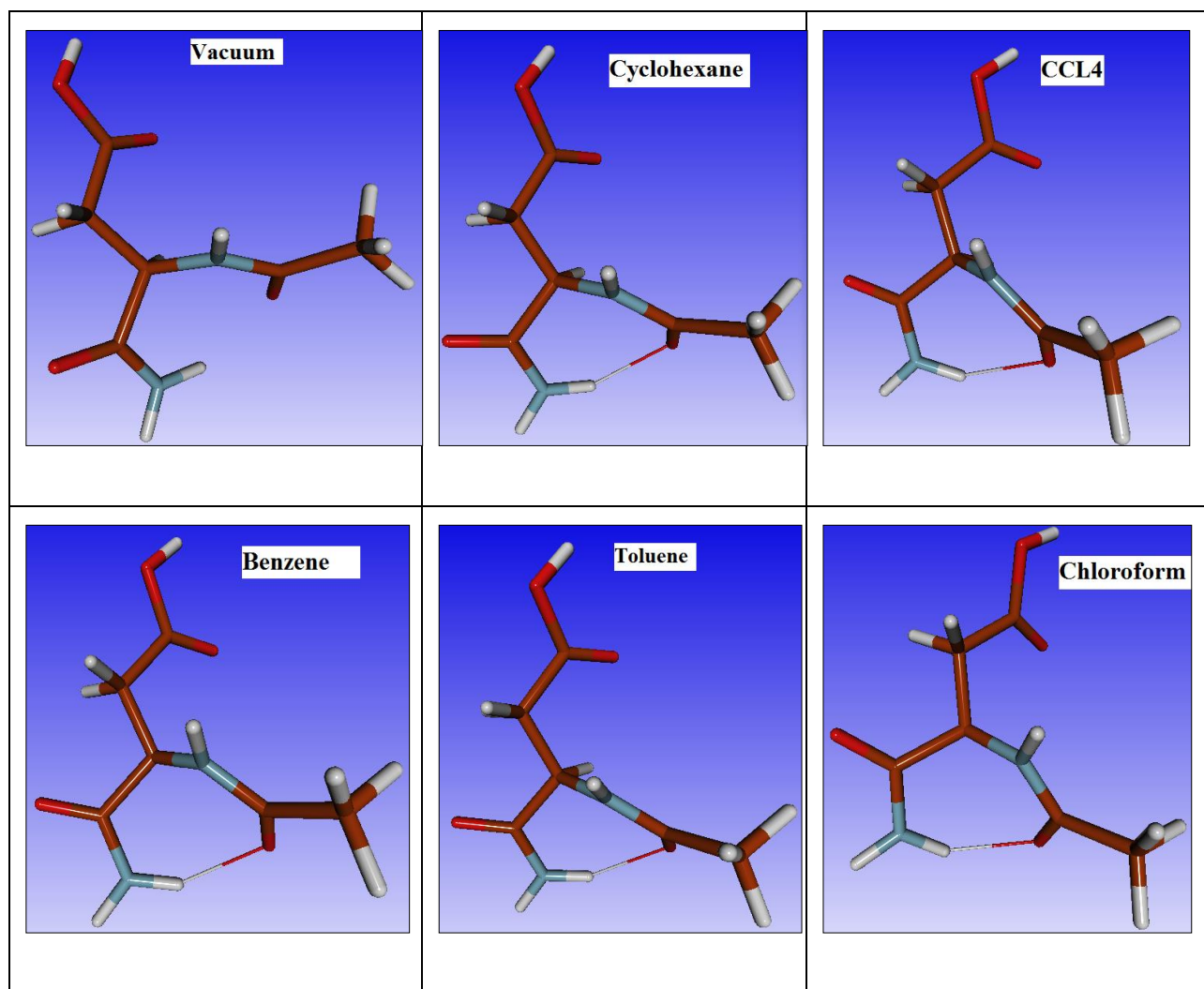
Solvent	Dielectric constant value	Number of steps	Energy (Hartree)
Water	80.1	24	-643.352
DMSO	46.7	22	-643.352
Nitromethane	35.87	27	-643.352
Methanol	32.7	25	-643.352
Ethanol	24.5	26	-643.352
Acetone	20.7	25	-643.352
1,2-dichloroethane	8.93	24	-644.350
THF	7.58	22	-643.350
Aniline	6.89	22	-643.349
Cl-benzene	5.62	23	-643.349
Chloroform	4.81	22	-643.348
Toluene	2.38	22	-643.343
Benzene	2.27	22	-643.373
CCl <sub>4</sub>	2.24	24	-643.343
Cyclohexane	2.02	25	-643.342
Vacuum	1.00	66	-643.343

Similarly, the variation in bond length, bond order and electronic distribution of the oxygen and hydrogen atoms was calculated (Table 4.7)

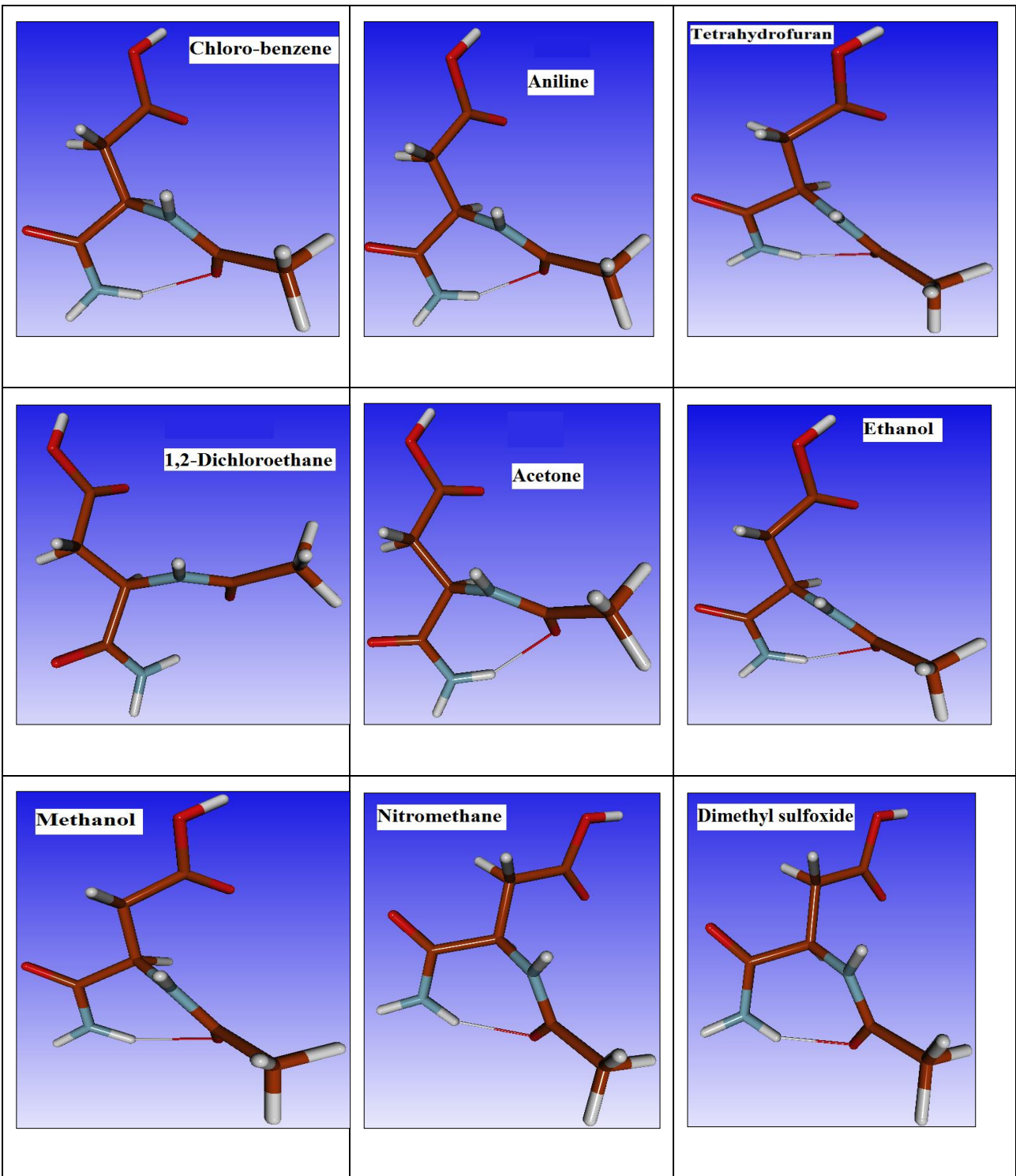
Table 4.7: Variation in bond lengths, bond order, bond distance and valence electron distribution on oxygen and hydrogen atoms of carboxylic “O-H” in aspartic acid side chain in various implicit solvents using Møller-Plesset (MP2) method. The Initial bond length was 0.95886 angstroms.

Solvent	Dielectric constant	Final Bond distance (side chain carboxylic “O-H” bond) (Å)	Bond order	Valence electron (O)	Valence electron (H)
Water	80.1	0.975	0.761	2.132	0.824
DMSO	46.7	0.975	0.762	2.132	0.824
Nitromethane	35.87	0.975	0.762	2.132	0.825
Methanol	32.7	0.975	0.762	2.132	0.825
Ethanol	24.5	0.975	0.762	2.132	0.825
Acetone	20.7	0.974	0.762	2.132	0.825
1,2-dichloroethane	8.93	0.974	0.763	2.132	0.827
THF	7.58	0.974	0.764	2.132	0.828
Aniline	6.89	0.974	0.764	2.132	0.828
Cl-benzene	5.62	0.974	0.765	2.132	0.829
Chloroform	4.81	0.974	0.766	2.132	0.830
Toluene	2.38	0.974	0.770	2.133	0.836
Benzene	2.27	0.973	0.771	2.133	0.836
CCl <sub>4</sub>	2.24	0.973	0.771	2.133	0.836
Cyclohexane	2.02	0.973	0.772	2.134	0.838
Vacuum	1.00	0.933	0.781	2.151	0.848

The variation in bond order and bond distance in aspartic acid side chain carboxylic “O-H” bond using MP2 is similar to the density functional theory analysis. Hence all three methods of optimization suggest a similar trend of increase in bond order and decrease in the bond distance with decreasing surrounding dielectric medium. The structures after optimization were slightly different from each other with RMSD of 0.088 angstroms calculated using visual molecular dynamics (Humphrey et al., 1996) (Figure 4.2).







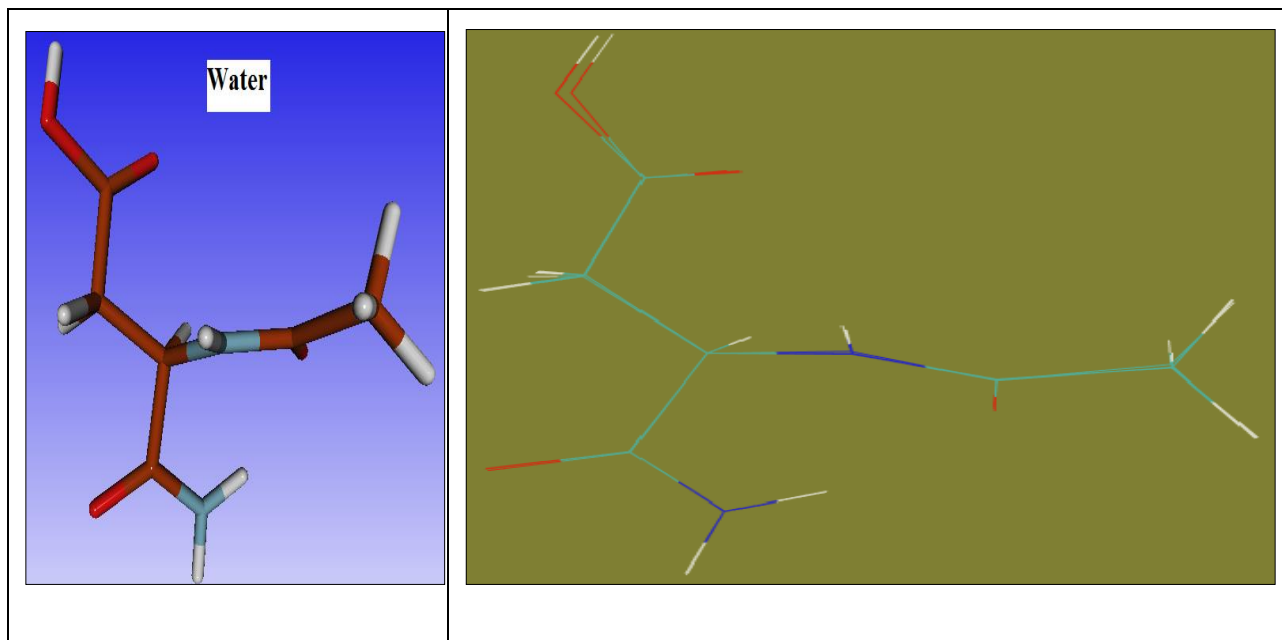


Figure 4.2: The DFT optimized capped aspartic acid in 16 different implicit models are shown. The last image is the superimposed structure of all the 16 optimized structures.

It has been observed that according to the superimposed structures, the two structures simulated into water and vacuum were visible clearly, other were superimposed but not visible as they were in the intermediate state of two structures (in vacuum and in water).

As per our data, beyond the dielectric value of 20, no further change in proton dissociation (in terms of bond order) in aspartic acid side chain carboxylic acid "O-H" bond was observed. This dielectric dependent proton dissociation in different implicit solvents can be compared to transfer of aspartic acid from protein interior to the protein surface. Protein interior is believed to be hydrophobic with an average dielectric constant of 2 to 4 (Haque et al., 2000a), whereas, protein surface exposed to water should have dielectric constant, close to 80.

To understand the effect of surrounding groups on proton dissociation, the capped aspartic acid was optimized in presence of one water and one sodium hydroxide molecule that is discussed in following sections. The one water and one sodium hydroxide molecules represent a weak and strong base respectively. The comparison of bond order change in these two explicit molecules will show the impact of the type surrounding atoms on the protonation state of a molecule.

#### 4.3.2) Effect of weak base (water) and strong base (NaOH) on the proton dissociation of aspartic acid side chain, from quantum chemical study

The above study of aspartic acid side chain dissociation in different implicit solvents showed a trend in decreasing bond order with increasing solvent polarity. Although, complete dissociation of the proton was not observed in this study, however, experimental pKa calculations have shown that complete dissociation of aspartic acid is possible within protein structures (Dugas and Penney, 2013).

Here we have studied two bi-molecular systems using quantum chemical methods; i) bi-molecular interaction between an aspartic acid and a water molecule. It is an example of weak acid and weak base interaction studied with Hartree-Fock, DFT and MP2 methods in vacuum (Table 4.8)

Table 4.8: Optimization methods and basis sets used for optimizing the system with capped aspartic acid in presence of single water molecule. The number of steps required and the corresponding final energy values are also reported for all the optimization methods.

Method employed	Basis set	Number of steps for optimization	Energy (Hartree)
Hartree-Fock	6-31G (d, p+)	111	-717.525
Density Functional Theory (B3LYP)	6-31G (d, p+)	52	-721.212
Møller-Plesset	6-31G (d, p+)	43	-719.586

The capped aspartic acid when optimized using Hartree-Fock method in vacuum in presence of one water molecule shows a slight increment of bond length (Table 4.9) and a decrease in bond order as compared to the bond order of the “O-H” bond when aspartic acid was optimized in vacuum only (table 4.3).

Table 4.9: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one water molecule using Hartree-Fock method. Initial bond distance was 0.958 angstroms.

Final bond length	Valence electron (O)	Valence electron (H)	Bond order
0.961	1.736	0.774	0.744

The optimized system of capped aspartic acid in presence of one water molecule was used for DFT and MP2 calculations. In case of DFT calculations, the bond length has increased significantly (Table 4.10) and simultaneously the bond order has also decreased as compared to the bond order value of the same aspartic acid side chain carboxylic “O-H” bond in vacuum (Table 4.5)

Table 4.10: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one water molecule using Density Functional Theory (B3LYP). Initial bond distance was 0.958 angstroms.

Final bond length	Valence electron (O)	Valence electron (H)	Bond order
0.996	1.847	0.805	0.708

The MP2 calculations on the aspartic acid side chain carboxylic “O-H” bond show a similar trend of bond length and bond order (Table 4.11) as compared to DFT calculations. The bond length has increased and the bond order has also decreased significantly as compared to the aspartic acid side chain carboxylic “O-H” bond in vacuum (Table 4.7).

Table 4.11: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one water molecule using Møller Plesset method. Initial bond distance was 0.958 angstroms.

Final bond length	Valence electron (O)	Valence electron (H)	Bond order
0.992	2.089	0.796	0.692

The variation in bond order and bond distance of the aspartic acid side chain “O-H” bond in presence of one water molecule shows that there is an increase in the bond distance and a decrement of bond order as compared to the bond order in vacuum (table 4.3, table 4.5, and table 4.6). Based on the bond order values the proton dissociation was not possible (figure 4.1). The presence of bond order shows that due to the one water molecule the aspartic acid side chain carboxylic “O-H” bond has become labile but still there is no dissociation as it acts as a weak base. Therefore, the single water molecule was replaced by strong base to test the dissociation.

ii) Bi-molecular interaction between an aspartic acid and a sodium hydroxide molecule; it is an example of weak acid and strong base interaction studied with Hartree-Fock, DFT and MP2 methods in vacuum (Table 4.12).

Table 4.12: Optimization methods and basis sets used for optimizing the system with capped aspartic acid in presence of single NaOH molecule. The number of steps required and the corresponding final energy values are also reported for all the optimization methods.

Method employed	Basis set	Number of steps for optimization	Energy (Hartree)
Hartree-Fock	6-31G (d, p+)	110	-878.861
Density Functional Theory (B3LYP)	6-31G (d, p+)	36	-882.948
Møller-Plesset	6-31G (d, p+)	59	-880.934

The capped aspartic acid in presence of a single NaOH molecule was optimized using Hartree-Fock method. Based on the optimized structures, the bond distance, and the electronic distribution was calculated (Table 4.13). As expected, there was no reported bond order which suggested the dissociation of the aspartic acid side chain.

Table 4.13: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one NaOH molecule using Hartree-Fock method. Initial bond distance was 0.958 angstroms.

Final bond length	Valence electron (O)	Valence electron (H)	Bond order
3.789	1.409	0.788	-

To further test the optimization process, the system of capped aspartic acid in presence of single NaOH molecule was subjected to DFT and MP2 calculations. The DFT calculation showed that the distance between the Oxygen and the hydrogen atom of the aspartic acid side chain carboxylic “O-H” bond has increased significantly and again there was no bond order reported (Table 4.14)

Table 4.14: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one NaOH molecule using Density Functional Theory (B3LYP). Initial bond distance was 0.958 angstroms.

Final bond length	Valence electron (O)	Valence electron (H)	Bond order
3.590	1.598	0.818	-

The MP2 calculations on the aspartic acid molecule in presence of one NaOH molecule showed that the optimized structure does not have bonded aspartic acid side chain carboxylic “O-H” bond (Table 4.15) and similar to DFT calculations, there is an increase in the distance between the aspartic acid side chain carboxylic “O-H” bond atoms (not bonded now).

Table 4.15: Variation in aspartic acid side chain carboxylic “O-H” bond when interacting with one NaOH molecule using Møller Plesset method. Initial bond distance was 0.958 angstroms.

Final bond length	Valence electron (O)	Valence electron (H)	Bond order
3.590	1.890	0.808	-

The presence of weak base does not dissociate the proton from the side chain carboxylic group, whereas, in presence of NaOH, which is a strong base, deprotonation is easily observed due to subsequent acid base neutralization reaction and leading to the formation of a water molecule (Figure 4.3).

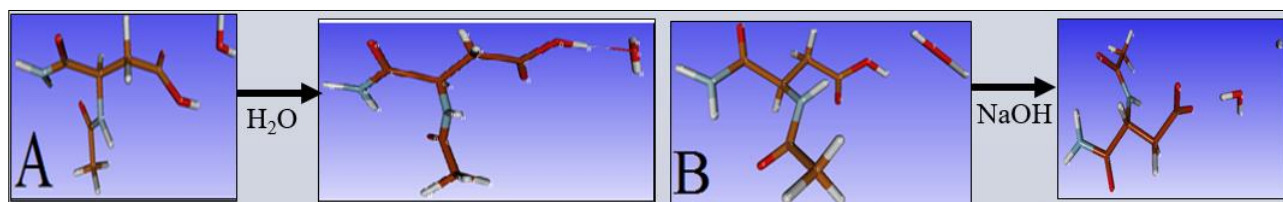


Figure 4.3: Aspartic acid side chain proton A) not dissociated in presence of water and B) dissociated in presence of NaOH, in vacuum ( $\epsilon=1$ ).

However, proton dissociation of aspartic acid is often observed within protein structure, in absence of any strong base. The only available base around the proteins is water molecule which is weak and is unable to dissociate proton, as observed from the above experiments. How the carboxylic proton dissociates within the protein structure is an interesting question. It can be explained, if the local dielectric medium inside protein interior is sufficiently high to favor the deprotonation. Based on this postulate, we attempted to study deprotonation of two aspartic acids in two different protein structures.

### 4.3.3) Quantum chemical calculation to study proton dissociation of aspartic acid side chains within bacteriorhodopsin and toxin proteins using Hartree-Fock method.

The aspartic acids at 96<sup>th</sup> position of bacteriorhodopsin protein (PDB ID: 1R2N) and at 43<sup>rd</sup> position of toxin protein (PDB ID: 1ORL) were chosen for our current analyses. According to experimental pK<sub>a</sub> measurement by NMR spectroscopy (Patzelt et al., 2002), the aspartic acid 96 in 1R2N was expected to be protonated (pK<sub>a</sub> value 8.12). Toxin protein was chosen as it is a small protein. NMR structures of both proteins were used so that aspartic acid coordinates were available in protonated form. The 4.5 Å region around these aspartic acids in individual proteins was considered to be the protein microenvironment that influences its protonation state (Bandyopadhyay and Mehler, 2008) (Figure 4.4(a), (b)).

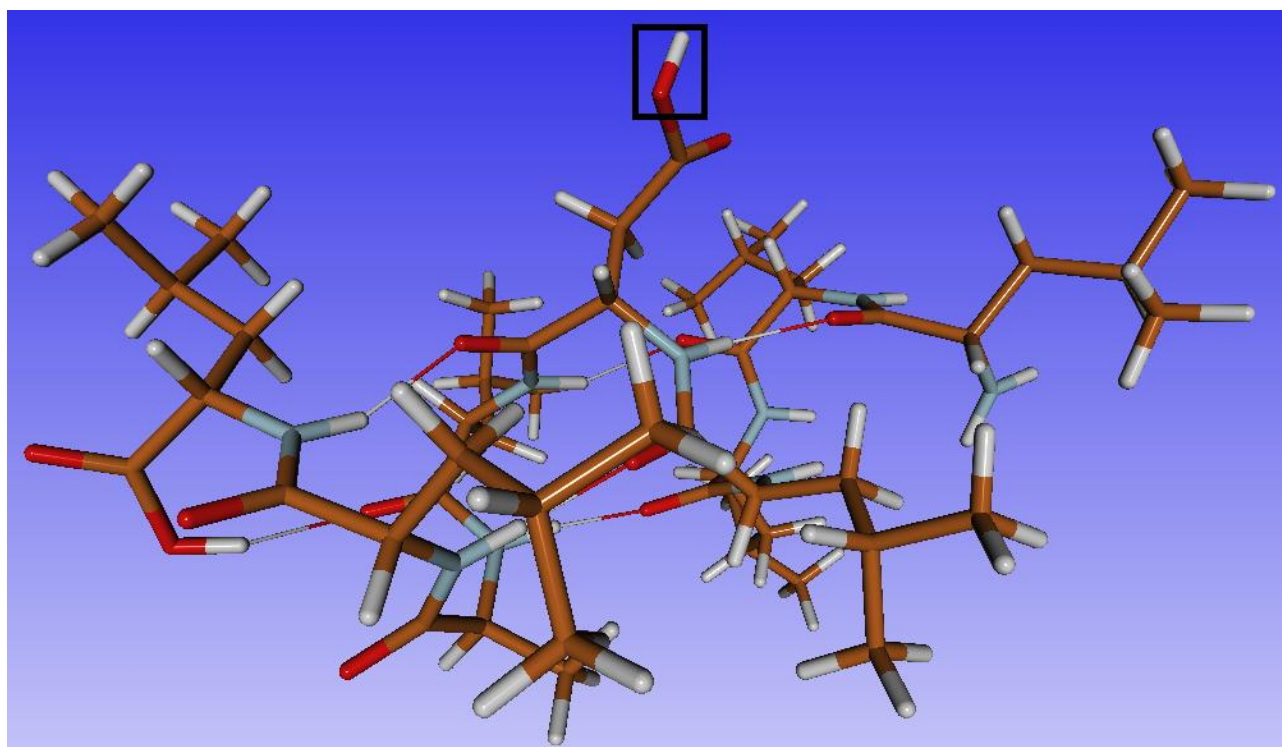


Figure 4.4(a): Microenvironment region of 4.5 angstroms around the Asp96 residue of bacteriorhodopsin protein (PDB ID: 1R2N). The aspartic acid side chain carboxylic “O-H” bond is highlighted within a box.



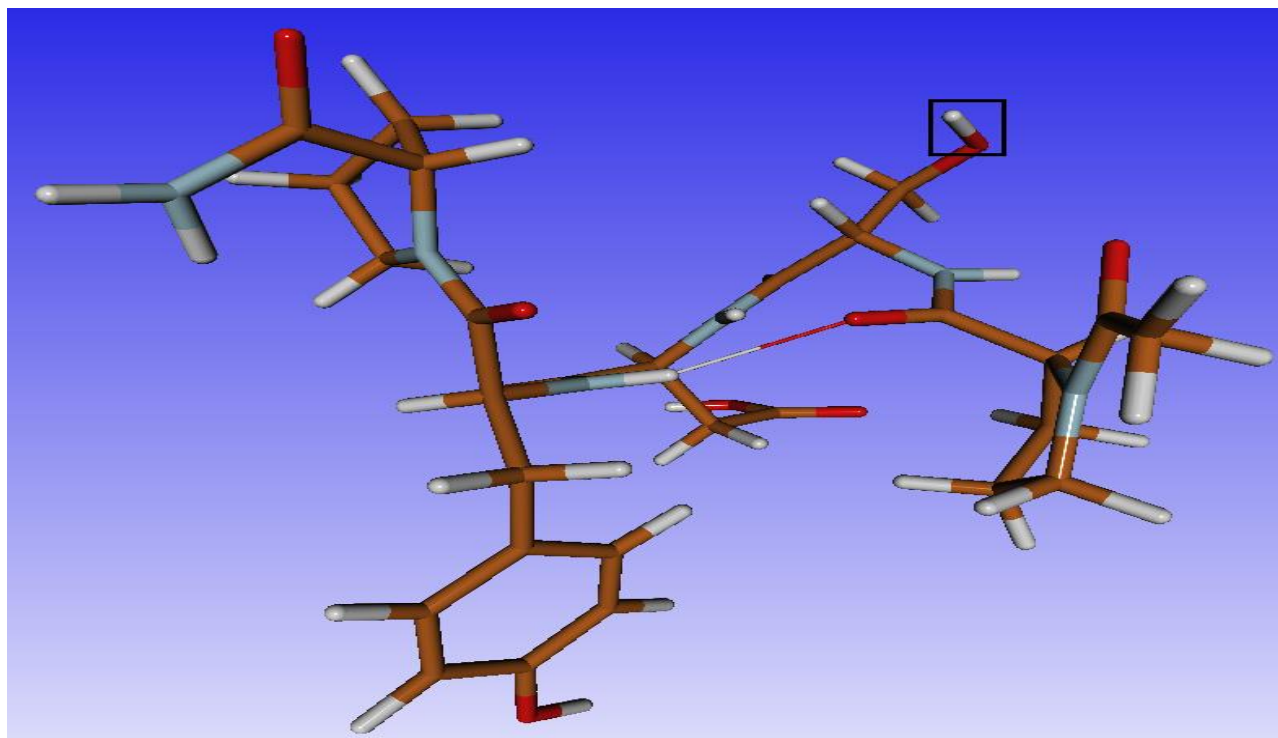


Figure 4.4(b): Microenvironment region of 4.5 angstroms around the Asp43 residue of toxin protein (PDB ID: 1ORL). The aspartic acid side chain carboxylic “O-H” bond is highlighted within a box.

The two protein systems were optimized using GAMESS (Schmidt et al., 1993). The optimization performed leads to the formation of energy minimized structures and based on the optimized structures the bond order, bond length, bond angle, dihedral angle were calculated (Table 4.16).

Table 4.16: Quantum chemical calculations on the aspartic acid side chain carboxylic “O-H” bond in bacteriorhodopsin (PDB ID: 1R2N, D96) and toxin protein (PDB ID: 1ORL, D43). Initial bond length, bond angle, and dihedral angle values are shown in parenthesis.

Protein Name	No. of steps for optimization	Energy (Hartree)	Bond order	Final bond distance (Å)	Final bond angle (C $\gamma$ -“O-H”)	Final dihedral angle (C $\beta$ -C $\gamma$ -“O-H”)
Bacteriorhodopsin	260	-	0.812	0.947 (0.973)	109.17 <sup>0</sup>	-178.458

		3296.208			(101.398 <sup>0</sup> )	(-138.59 <sup>0</sup> )
Toxin	114	- 2157.988	0.823	0.944 (0.961)	113.053 <sup>0</sup> (113.106 <sup>0</sup> )	-2.422 <sup>0</sup> (0.216 <sup>0</sup> )

The quantum chemical calculations on these two aspartic acids within their embedded protein microenvironment solvent shows a decrement of bond distance and bond angle from the initial bond distance and bond angle respectively. As the systems were truncated from the full proteins, therefore the optimization took very a high number of optimization steps. Although, the bond order values are equivalent to the values in presence of vacuum. It is suspected that due to the truncation the “O-H” bond of interest was exposed to vacuum itself and hence protein microenvironment effect was not clearly observed.

To analyze the dielectric medium of these two protein microenvironments, the bond orders of the two optimized protein systems were compared with the bond order values of the optimized implicit solvent model systems (Figure 4.5).

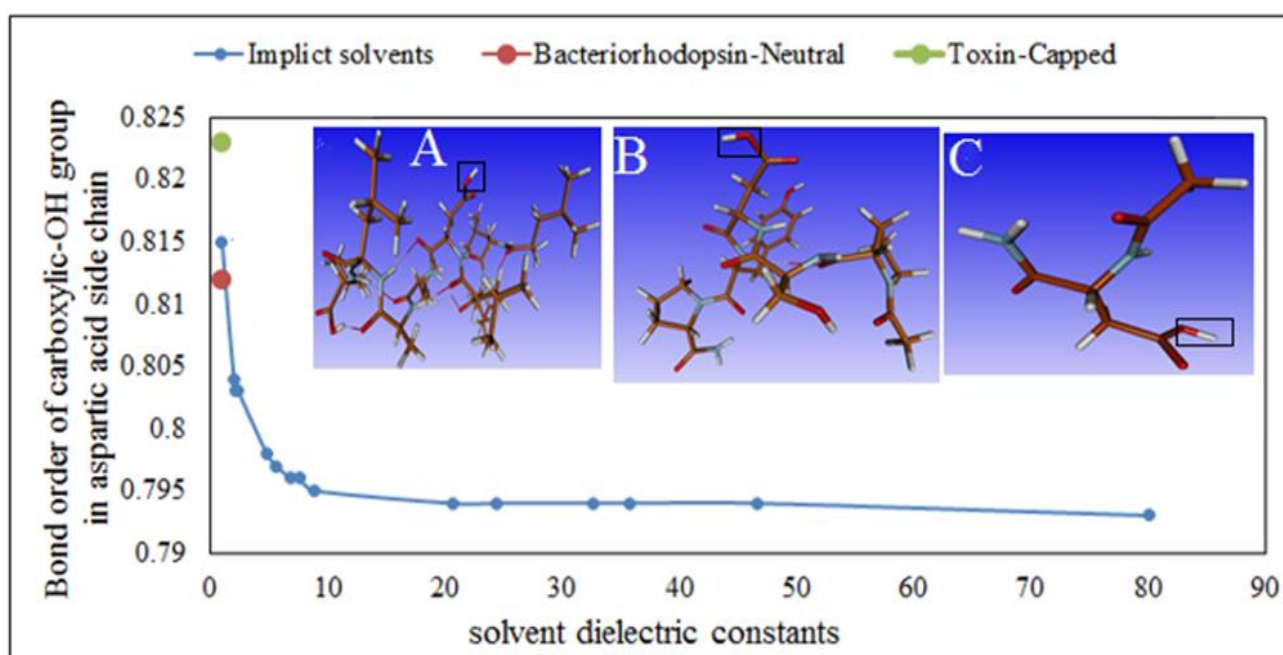


Figure 4.5: Variation in bond order of carboxylic-OH in aspartic acid side chain as a function of solvent dielectric constants. Inset pictures highlight the protons of aspartic acid in A) bacteriorhodopsin, B) toxin protein and C) capped aspartic acid in vacuum.

Proton dissociation in different implicit solvents exhibits steep decrease in oxygen-proton bond order within the low dielectric media ( $\epsilon$  from 1 to 9), representing the protein hydrophobic core. Bond order is invariant in higher dielectric media ( $\epsilon > 20$ ) representing the solvent exposed-hydrophilic region of protein structures. Bond order in higher dielectric media is comparable to that of aspartic acid in presence of weak base water, in vacuum (Figure 4.5). Aspartic acids in both the proteins were exposed to vacuum, hence, their bond orders closely resemble that of vacuum. The chosen protein microenvironment around the aspartic acid is similar to the vacuum conditions. Therefore, it was necessary to perform the quantum chemical calculation on the full protein which is computationally not feasible. Therefore, we tried QM/MM calculations on the protein structures that is explained in the next section.

#### **4.3.3) Proton dissociation of aspartic acid side chains within bacteriorhodopsin and toxin proteins using Hartree-Fock calculation using QM/MM calculations**

The QM/MM approach was used to estimate the bond order of carboxylic “O-H” bond in protein solvent. Although due to the high number (3655) of atoms in bacteriorhodopsin protein and in toxin proteins (679 atoms), the QM/MM optimization procedure did not complete. Therefore, the QM/MM calculations were performed on the microenvironment region (4.5 Å) of the ASP 96 and D43 residues of bacteriorhodopsin and toxin proteins respectively. The side chain carboxylic part was kept for QM calculations while the remaining microenvironment region was subjected to MM calculations. The variation in the bond order, bond distance and other bond properties were calculated from the optimized structures (Table 4.17).

Table 4.17: Quantum mechanical/Molecular Mechanics calculations on the aspartic acid side chain carboxylic “O-H” bond in bacteriorhodopsin (PDB ID: 1R2N, D96) and toxin protein (PDB ID: 1ORL, D43). Initial bond length, bond angle and dihedral angle values are shown in parenthesis.

Protein Name	No. of steps for optimization	Energy (Hartree)	Bond order	Final bond distance (Å)	Final bond angle (“C $\gamma$ -O-H”)	Final dihedral angle (“C $\beta$ -C $\gamma$ -O-H”)
Bacteriorhodopsin	23	-229.354	0.847	0.948 (0.973)	108.242 <sup>0</sup> (101.398 <sup>0</sup> )	-178.319 <sup>0</sup> (-138.59 <sup>0</sup> )
Toxin	21	-227.939	0.859	0.943 (0.961)	112.294 <sup>0</sup> (113.106 <sup>0</sup> )	-3.886 <sup>0</sup> (0.216 <sup>0</sup> )

The QM/MM optimization of the truncated microenvironment regions finished in a comparatively lesser number of optimization steps (Table 4.17). The bond distance decrement was observed similar to the QM calculations (Table 4.16). Although, the bond angle has increased from the initial value suggesting favoring the dissociation in bacteriorhodopsin protein, but the bond orders were again found comparable to the values in QM optimized structures (Table 4.16). This is because of the truncated region used for the QM/MM analysis has the side chain aspartic acid carboxylic “O-H” bond exposed to vacuum region, due to which the protein solvent effect on the bond orders was not observed. Therefore, QM/MM analysis must be performed on full proteins to study the protein effect on the side chain protonation state of aspartic acid.

#### **4.4) Conclusions**

In this work we have studied the influence of different i) implicit solvents, ii) weak and strong bases and iii) protein structures towards aspartic acid side chain deprotonation. A considerable amount of variation has been observed in proton dissociation of the aspartic acid side chain, under different chemical conditions. Our specific interest was to understand deprotonation of aspartic acid side chains in proteins, those were compared in presence of the other conditions. This comparison helped to identify dielectric medium of the local protein structure around the aspartic acid side chain. Solvents with low dielectric media (presumably resembling protein hydrophobic core) favor protonated form of the aspartic acid side chain (those supported by experimentally observed large upward pKa shift).

## CHAPTER 5

### Conclusions and Future perspective

#### 6.1) Functional role of thiol group containing cysteines in different microenvironment regions

Cysteine residues perform multiple functions in protein structures. It is known that cysteine residues have a reactive side chain thiol group. A slight modification in the cysteine microenvironment modulates the function of cysteine residues. In this thesis, it has been shown that variation in the surrounding microenvironment around cysteine residues defines their function as active, metal binding or redox cysteines. Moreover, the role of the microenvironment in maintaining the functional form of cysteines in different functional motifs like C-x-x-C-H motif in cytochrome proteins and C-x-x-C motif in zinc binding proteins has been explained.

It is important to note that understanding the redox modification of cysteine residues will enhance the knowledge in the development of cysteine based redox agents that produces a redox environment and help in different types of disease treatments (Giles et al., 2003). Such agents are known as redox drugs (Giles et al., 2003).

#### 6.2) Functions of different microenvironments based on the population of disulfide bridged cystine residues

Cystine residues are the oxidized form of cysteine amino acids. This is an important form of cysteine residues because it maintains the protein stability by forming strong disulfide bonds. Cystine residues were present in three types of microenvironment regions: Buried-hydrophobic, Buried-hydrophilic and Exposed-hydrophilic. It has been shown that cystines in different microenvironment regions perform different types of functions. Moreover, the disulfide bond strength of cystine residues varies with varying microenvironment of the involved half cystines. Therefore, this study helps in understanding the conditions required for maintaining disulfide bond strength in cystine residues. This analysis will be helpful in designing and engineering proteins containing cystine residues by maintaining a stable protein structure through disulfide bonds.

Similarly, understanding the microenvironment around other residues will improve our knowledge of protein structures and subsequent novel designing of proteins and important enzymes.

Microenvironment based modulation of the amino acid function will also help in disease treatment by modulating the function of important residues of the protein involved in diseases. For example; anti-malarial drugs alter the microenvironment of a leucine residue in spermidine synthase protein in *Thermotoga Maritima* (Bandyopadhyay and Mehler, 2008).

### **6.3) Effect of protein solvent medium on protonation state of aspartic acid side chain**

Amino acids in heterogeneous protein microenvironments are analogous to amino acids in different solvents with variable dielectric constants. The analysis on the aspartic acid side chain is an attempt to correlate the dielectric properties of the protein medium with the established solvents. This study shows that protein medium does not act as a single solvent but a range of dielectric medium varying from hydrophobic core to a relatively hydrophilic protein exterior. It was observed that though water acts as a standard in-vivo solvent system, yet protein microenvironment around the aspartic acid side chain may have a crucial role in determining the side chain protonation state. It was also observed that lower dielectric media resembles the protein hydrophobic core, whereas the higher dielectric constant values represent the hydrophilic microenvironment.

### **6.4) Specific contribution of the work**

This work shows why a particular amino acid can perform different the functions in different proteins. This work shows (both qualitatively and quantitatively) the significance of hydrophobicity of the surrounding region in the modulation of structure and function of the amino acids like cysteine and aspartic acid.

This work gives an approach to understand the dielectric nature of protein microenvironments by comparing protein microenvironments with other solvents.

### **6.5) Future plans based on the analysis and knowledge obtained from this thesis work**

### **6.5.1) Blind prediction of cysteine functions in unknown proteins based on the surrounding microenvironments**

The statistical data obtained by clustering the microenvironment functions based on cysteine population can be used to perform a blind prediction of functions of cysteine residues in unknown proteins. The quantified microenvironment around the cysteine residues shows the type of microenvironment cluster it belongs to. This helps to predict the function of that particular cysteine residue in the respective protein structure. The proposed results should be verified against experiments.

### **6.5.2) Analysis of the role of microenvironment in modulating an enzymatic reaction**

Microenvironment calculations of the enzyme before and after substrate binding may help in understanding the conformational changes in enzyme and its role in a particular biochemical reaction. This study will help in modulating various important enzymatic reactions that are involved in disease caused due to malfunctioning of respective enzymes. Moreover, modulating an enzymatic reaction may help in increasing the product formation that will be very helpful for downstream processing industries.

Calculation of microenvironment around similar enzymatic proteins will show conservation of microenvironments around catalytic center in the same protein family.

### **6.5.3) Extension of understanding the heterogeneous protein dielectric nature**

In this study, we have used only two proteins to elucidate the alternation in aspartic acid side chain protonation state with varying dielectric medium. Due to computational constraints, QM/MM calculations were not performed on more protein structures with side chain subjected to QM optimization and remaining protein structure subjected to MM optimization. Further investigation will help in the generalization of the dielectric nature of proteins around aspartic acid side chains.

Similar work should be extended to other titrable residues, like Lysine, Glutamic acid, Arginine etc., to identify different types of dielectric medium around those titrable amino acids.



## References

1. Addinsoft (2014). XLSTAT 2014, Data analysis and statistics software for Microsoft Excel.
2. Aguzzi A, O'Connor T (2010). Protein aggregation diseases: pathogenicity and therapeutic perspectives. *Nat Rev Drug Discov* 9, 237–48.
3. Akabas MH (2015). Cysteine Modification: Probing Channel Structure, Function and Conformational Change. *Adv Exp Med Biol* 869, 25–54.
4. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002). *Molecular Biology of the Cell*.
5. Angelini G, Gardella S, Ardy M, Ciriolo MR, Filomeni G, Di Trapani G, Clarke F, Sitia R, Rubartelli A (2002). Antigen-presenting dendritic cells provide the reducing extracellular microenvironment required for T lymphocyte activation. *Proc Natl Acad Sci U S A* 99, 1491–6.
6. Bagley SC, Altman RB (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci* 4, 622–35.
7. Bairoch A (2000). The ENZYME database in 2000. *Nucleic Acids Res* 28, 304–305.
8. Bandyopadhyay D, Mehler EL (2008). Quantitative expression of protein heterogeneity: Response of amino acid side chains to their local environment. *Proteins* 72, 646–659.
9. Bardwell JCA (1994). Building bridges: disulfide bond formation in the cell. *Mol Microbiol* 14, 199–205.
10. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002). Analysis of Catalytic Residues in Enzyme Active Sites. *J Mol Biol* 324, 105–121.
11. de Beer TAP, Berka K, Thornton JM, Laskowski RA (2014). PDBsum additions. *Nucleic Acids Res* 42, D292-6.

12. Berg JM, Tymoczko JL, Stryer L (2002). *Biochemistry*, 5th ed. W H Freeman., New York, USA
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. *Nucleic Acids Res* 28, 235–42.
14. Bhandary B, Marahatta A, Kim H-R, Chae H-J (2012). An involvement of oxidative stress in endoplasmic reticulum stress and its associated diseases. *Int J Mol Sci* 14, 434–56.
15. Blake CC, Mair GA, North AC, Phillips DC, Sarma VR (1967). On the conformation of the hen egg-white lysozyme molecule. *Proc R Soc London Ser B, Biol Sci* 167, 365–77.
16. Braberg H, Webb BM, Tjioe E, Pieper U, Sali A, Madhusudhan MS (2012). SALIGN: a web server for alignment of multiple protein sequences and structures. *Bioinformatics* 28, 2072–3.
17. Branden C, Tooze J (1991). *Introduction to protein structure*, second edi. Garland Publishing Taylor and Francis group., New York
18. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al (2009). CHARMM: The biomolecular simulation program. *J Comput Chem* 30, 1545–1614.
19. Cederbaum AI (2015). Molecular mechanisms of the microsomal mixed function oxidases and biological and pathological implications. *Redox Biol* 4, 60–73.
20. Chen J, Yang M, Hu G, Shi S, Yi C, Zhang Q (2009). Insights into the functional role of protonation states in the HIV-1 protease-BEA369 complex: molecular dynamics simulations and free energy calculations. *J Mol Model* 15, 1245–52.
21. Chen S, Bahar I (2004). Mining frequent patterns in protein structures: a study of protease families. *Bioinformatics* 20, 1–18.
22. Chothia C (1976). The nature of the accessible and buried surfaces in proteins. *J Mol Biol* 105, 1–12.

23. Chou PY, Fasman GD (1974). Prediction of protein conformation. *Biochemistry* 13, 222–245.
24. Chuang C-C, Chen C-Y, Yang J-M, Lyu P-C, Hwang J-K (2003). Relationship between protein structures and disulfide-bonding patterns. *Proteins* 53, 1–5.
25. Collings AF, Critchley C (2007). Artificial photosynthesis: From Basic Biology to Industrial application. *In* Anthony F. Collings et. al, ed, John Wiley & Sons, pp 280–282.
26. Cooper GM (2000). The Central Role of Enzymes as Biological Catalysts.
27. Cornish-Bowden A (2014). Understanding allosteric and cooperative interactions in enzymes. *FEBS J* 281, 621–32.
28. Cuesta SM, Rahman SA, Furnham N, Thornton JM (2015). The Classification and Evolution of Enzyme Function. *Biophys J* 109, 1082–6.
29. Cunningham H, Tablan V, Roberts A, Bontcheva K (2013). Getting more out of biomedical documents with GATE’s full lifecycle open source text analytics. *PLoS Comput Biol* 9, e1002854.
30. Das A, Chakrabarti J, Ghosh M (2014). Thermodynamics of interfacial changes in a protein-protein complex. *Mol Biosyst* 10, 437–45.
31. Daubner SC, Le T, Wang S (2011). Tyrosine hydroxylase and regulation of dopamine synthesis. *Arch Biochem Biophys* 508, 1–12.
32. DeBerardinis RJ, Thompson CB (2012). Cellular metabolism and disease: what do metabolic outliers teach us? *Cell* 148, 1132–44.
33. Deegan BJ, Seldeen KL, McDonald CB, Bhat V, Farooq A (2010). Binding of the ERalpha nuclear receptor to DNA is coupled to proton uptake. *Biochemistry* 49, 5978–88.
34. Denisov DA, Drozdov-Tichomirov LN, Grigoryeva DN (1973). Pattern recognition methods for prediction of helical region in proteins. *J Theor Biol* 41, 431–439.

35. Depuydt M, Messens J, Collet J-F (2011). How proteins form disulfide bonds. *Antioxid Redox Signal* 15, 49–66.
36. Díaz N, Suárez D (2016). Role of the Protonation State on the Structure and Dynamics of Albumin. *J Chem Theory Comput* 17, 1972–1988.
37. Dinner AR (2000). Local deformations of polymers with nonplanar rigid main-chain internal coordinates. *J Comput Chem* 21, 1132–1144.
38. Dugas H, Penney C (2013). *Bioorganic Chemistry: A Chemical Approach to Enzyme Action*. Springer Science & Business Media.
39. Ebert JC, Altman RB (2008). Robust recognition of zinc binding sites in proteins. *Protein Sci* 17, 54–65.
40. Eisenberg D, McLachlan AD (1986). Solvation energy in protein folding and binding. *Nature* 319, 199–203.
41. Eisenberg D, Wilcox W, McLachlan AD (1986). Hydrophobicity and amphiphilicity in protein structure. *J Cell Biochem* 31, 11–7.
42. Eisenberg D, McLachlan AD (1986). Solvation energy in protein folding and binding. *Nature* 319, 199–203.
43. Eklund H, Ingelman M, Söderberg BO, Uhlin T, Nordlund P, Nikkola M, Sonnerstam U, Joelson T, Petratos K (1992). Structure of oxidized bacteriophage T4 glutaredoxin (thioredoxin). Refinement of native and mutant proteins. *J Mol Biol* 228, 596–618.
44. Fass D (2012a). Disulfide bonding in protein biophysics. *Annu Rev Biophys* 41, 63–79.
45. Fass D (2012b). Disulfide bonding in protein biophysics. *Annu Rev Biophys* 41, 63–79.
46. Feige MJ, Hendershot LM (2011). Disulfide bonds in ER protein folding and homeostasis. *Curr Opin Cell Biol* 23, 167–75.
47. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R (2005). Optimal docking area: a

- new method for predicting protein-protein interaction sites. *Proteins* 58, 134–43.
48. Filippone F, Meloni S, Parrinello M (2001). A novel implicit Newton–Raphson geometry optimization method for density functional theory calculations. *J Chem Phys* 115, 636.
  49. Finkel T (2003). Oxidant signals and oxidative stress. *Curr Opin Cell Biol* 15, 247–254.
  50. Frausto da Silva JJ., Williams RJ. (2001). *The Biological Chemistry of the Elements*, Second. Oxford Press.
  51. Furnham N, Holliday GL, de Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM (2014). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res* 42, D485-9.
  52. van Geest M, Lolkema JS (2000). Membrane Topology and Insertion of Membrane Proteins: Search for Topogenic Signals. *Microbiol Mol Biol Rev* 64, 13–33.
  53. Giles NM, Watts AB, Giles GI, Fry FH, Littlechild JA, Jacob C (2003). Metal and Redox Modulation of Cysteine Protein Function. *Chem Biol* 10, 677–693.
  54. Gilson MK, Honig BH (1986). The dielectric constant of a folded protein. *Biopolymers* 25, 2097–119.
  55. Glusker J., Amy.K K, Bock C. (1999). METAL IONS IN BIOLOGICAL SYSTEMS. *Rigaku J* 16, 8–17.
  56. Grabarczyk DB, Chappell PE, Eisel B, Johnson S, Lea SM, Berks BC (2015). Mechanism of thiosulfate oxidation in the SoxA family of cysteine-ligated cytochromes. *J Biol Chem* 290, 9209–21.
  57. Greene BL, Wu C-H, McTernan PM, Adams MWW, Dyer RB (2015). Proton-coupled electron transfer dynamics in the catalytic mechanism of a [NiFe]-hydrogenase. *J Am Chem Soc* 137, 4558–66.
  58. Gutteridge A, Thornton JM (2005). Understanding nature’s catalytic toolkit. *Trends*

59. Haigi A. (2013). Methodologies and Applications for Chemoinformatics and Chemical Engineering. IGI Global., Hershey PA
60. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR (2012). Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J Cheminform* 4, 17.
61. Haque E, Ray S, Chakrabarti A (2000a). Polarity Estimate of the Hydrophobic Binding Sites in Erythroid Spectrin : A Study by Pyrene Fluorescence. 10:
62. Haque ME, Ray S, Chakrabarti A (2000b). Polarity Estimate of the Hydrophobic Binding Sites in Erythroid Spectrin: A Study by Pyrene Fluorescence. *J Fluoresc* 10, 1–6.
63. Harris TK, Turner GJ (2002). Structural Basis of Perturbed pK<sub>a</sub> Values of Catalytic Groups in Enzyme Active Sites. *IUBMB Life* 53, 85–98.
64. Harrison PM, Sternberg MJ (1996). The disulfide beta-cross: from cystine geometry and clustering to classification of small disulfide-rich protein folds. *J Mol Biol* 264, 603–23.
65. Hartigan JA (1975). Clustering Algorithms. John Wiley & Sons, Inc., New York, USA
66. Hartmann T, Bieger SC, Brühl B, Tienari PJ, Ida N, Allsop D, Roberts GW, Masters CL, Dotti CG, Unsicker K, et al (1997). Distinct sites of intracellular production for Alzheimer's disease A $\beta$ 40/42 amyloid peptides. *Nat. Med.*
67. Hazes B, Dijkstra BW (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng* 2, 119–25.
68. Hobohm U, Sander C (1994). Enlarged representative set of protein structures. *Protein Sci* 3, 522–4.
69. Holbourn KP, Acharya KR, Perbal B (2008). The CCN family of proteins: structure-function relationships. *Trends Biochem Sci* 33, 461–73.

70. Holmgren A, Björnstedt M (1995). Thioredoxin and thioredoxin reductase. *Methods Enzymol* 252, 199–208.
71. Hooke RC (1665). *Micrographia: or Some Physiological Descriptions of Miniature Bodies Made by Magnifying Glasses*. London, England
72. Houk J, Singh R, Whitesides GM (1987). Measurement of thiol-disulfide interchange reactions and thiol pKa values. *Methods Enzymol* 143, 129–40.
73. Howland J (1990). *Biochemistry*. *Biochem Educ* 18, 212.
74. Huang Y-MM, You W, Caulkins BG, Dunn MF, Mueller LJ, Chang C-EA (2016). Protonation states and catalysis: Molecular dynamics studies of intermediates in tryptophan synthase. *Protein Sci* 25, 166–83.
75. Hubbard TJ, Murzin AG, Brenner SE, Chothia C (1997). SCOP: a structural classification of proteins database. *Nucleic Acids Res* 25, 236–9.
76. Humphrey W, Dalke A, Schulten K (1996). VMD: visual molecular dynamics. *J Mol Graph* 14, 33–8, 27–8.
77. Hwang C, Sinskey AJ, Lodish HF (1992). Oxidized redox state of glutathione in the endoplasmic reticulum. *Science* 257, 1496–502.
78. Jacob C, Giles GI, Giles NM, Sies H (2003). Sulfur and Selenium: The Role of Oxidation State in Protein Structure and Function. *Angew Chemie - Int Ed* 42, 4742–4758.
79. Jacob C, Knight I, Winyard PG (2006). Aspects of the biological redox chemistry of cysteine: from simple redox responses to sophisticated signalling pathways. *Biol Chem* 387, 1385–97.
80. Jha AN, Vishveshwara S, Banavar JR (2010). Amino acid interaction preferences in proteins. *Protein Sci* 19, 603–16.
81. Jiang Y, Ruta V, Chen J, Lee A, Mackinnon R (2003). The principle of gating charge

- movement in a voltage-dependent K<sup>+</sup> channel. *Nature* 423, 42–48.
82. Jocelyn PC (1967). The Standard Redox Potential of Cysteine-Cystine from the Thiol-Disulfide Exchange Reaction with Glutathione and Lipoic Acid. *Eur J Biochem* 2, 327–331.
  83. Jones DD (1975). Amino acid properties and side-chain orientation in proteins: A cross correlation approach. *J Theor Biol* 50, 167–183.
  84. Jones JG, Otieno S, Barnard EA, Bhargava AK (1975). Essential and nonessential thiols of yeast hexokinase. Reactions with iodoacetate and iodoacetamide. *Biochemistry* 14, 2396–403.
  85. Jones W (1949). *Inorganic Chemistry*. Philadelphia, Blakiston
  86. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Res* 39, D411-9.
  87. Kabsch W, Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–637.
  88. Katz BA, Kossiakoff A (1986). The crystallographically determined structures of atypical strained disulfides engineered into subtilisin. *J Biol Chem* 261, 15480–5.
  89. Kiley PJ, Storz G (2004). Exploiting thiol modifications. *PLoS Biol* 2, e400.
  90. Kim K, Jordan KD (1994). Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J Phys Chem* 98, 10089–10094.
  91. Kirkwood JG, Westheimer FH (1938). The Electrostatic Influence of Substituents on the Dissociation Constants of Organic Acids. I. *J Chem Phys* 6, 506.
  92. Kordysh M, Kornelyuk A (2006). Conformational flexibility of cytokine-like C-module of tyrosyl-tRNA synthetase monitored by Trp144 intrinsic fluorescence. *J Fluoresc* 16, 705–



- 11.
93. Koshland DE (1995). The Key–Lock Theory and the Induced Fit Theory. *Angew Chemie Int Ed English* 33, 2375–2378.
94. Kukic P, Farrell D, McIntosh LP, García-Moreno E B, Jensen KS, Toleikis Z, Teilum K, Nielsen JE (2013). Protein dielectric constants determined from NMR chemical shift perturbations. *J Am Chem Soc* 135, 16968–76.
95. Kundrot CE, Ponder JW, Richards FM (1991). Algorithms for calculating excluded volume and its derivatives as a function of molecular conformation and their use in energy minimization. *J Comput Chem* 12, 402–409.
96. Lakowicz J (1983). *Principles of fluorescence spectroscopy*. Plenum Press., New York, USA
97. Landsteiner K (2013). *The Specificity of Serological Reactions*. Courier Corporation.
98. Laskowski RA (2001). PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res* 29, 221–2.
99. Lee B, Richards FM (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 55, 379–400.
100. Lee Y, Kim Y, Yeom S, Kim S, Park S, Jeon CO, Park W (2008). The role of disulfide bond isomerase A (DsbA) of *Escherichia coli* O157:H7 in biofilm formation and virulence. *FEMS Microbiol Lett* 278, 213–22.
101. Lehninger A, Cox M, Nelson D (2008). *Lehninger principle of biochemistry*, 5th ed. W.H.Freeman., New York, USA
102. Lei K, Townsend DM, Tew KD (2008). Protein cysteine sulfinic acid reductase (sulfiredoxin) as a regulator of cell proliferation and drug response. *Oncogene* 27, 4877–87.

103. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* gkv279-.
104. Liao DI, Kapadia G, Reddy P, Saier MH, Reizer J, Herzberg O (1991). Structure of the IIA domain of the glucose permease of *Bacillus subtilis* at 2.2-Å resolution. *Biochemistry* 30, 9583–9594.
105. Lide DR (2004). *CRC Handbook of Chemistry and Physics*, 85th ed. CRC Press., Boca Raton, FL
106. Ma B, Elkayam T, Wolfson H, Nussinov R (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A* 100, 5772–7.
107. Maiorov V, Abagyan R (1998). Energy strain in three-dimensional protein structures. *Fold Des* 3, 259–269.
108. Maiti R, Van Domselaar GH, Zhang H, Wishart DS (2004). SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res* 32, W590-4.
109. Manavalan P, Ponnuswamy PK (1978). Hydrophobic character of amino acid residues in globular proteins. *Nature* 275, 673–674.
110. Manavalan P, Ponnuswamy PK (1977). A study of the preferred environment of amino acid residues in globular proteins. *Arch Biochem Biophys* 184, 476–87.
111. Marino SM, Gladyshev VN (2009). A Structure-Based Approach for Detection of Thiol Oxidoreductases and Their Catalytic Redox-Active Cysteine Residues. *PLoS Comput Biol* 5, e1000383.
112. Mark P, Nilsson L (2001). Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J Phys Chem A* 105, 9954–9960.
113. Martin JL (1995). Thioredoxin —a fold for all reasons. *Structure* 3, 245–250.

114. Martínez-Ruiz A, Lamas S (2007). Signalling by NO-induced protein S-nitrosylation and S-glutathionylation: convergences and divergences. *Cardiovasc Res* 75, 220–8.
115. Maseras F, Morokuma K (1995). IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J Comput Chem* 16, 1170–1179.
116. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R (2013). Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* 41, W597-600.
117. Mehler EL, Fuxreiter M, Simon I, Garcia-Moreno EB (2002). The role of hydrophobic microenvironments in modulating pKa shifts in proteins. *Proteins* 48, 283–92.
118. Mehler EL, Guarnieri F (1999). A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophys J* 77, 3–22.
119. Meng EC, Pettersen EF, Couch GS, Huang CC, Ferrin TE (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* 7, 1.
120. Mennucci B, Tomasi J, Cammi R, Cheeseman JR, Frisch MJ, Devlin FJ, Gabriel S, Stephens PJ (2002). Polarizable Continuum Model (PCM) Calculations of Solvent Effects on Optical Rotations of Chiral Molecules. *J Phys Chem A* 106, 6102–6113.
121. Messens J, Collet J-F (2013). Thiol-disulfide exchange in signaling: disulfide bonds as a switch. *Antioxid Redox Signal* 18, 1594–6.
122. Meunier B, de Visser SP, Shaik S (2004). Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes. *Chem Rev* 104, 3947–80.
123. Miseta A, Csutora P (2000). Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms. *Mol Biol Evol* 17, 1232–1239.
124. Mobbs C V, Kaplitt M PD (1998). In Prolyl Hydroxylase, Protein Disulfide Isomerase, and

125. Møller C, Plesset MS (1934). Note on an Approximation Treatment for Many-Electron Systems. *Phys Rev* 46, 618–622.
126. Moriarty-Craige SE, Jones DP (2004). Extracellular thiols and thiol/disulfide redox in metabolism. *Annu Rev Nutr* 24, 481–509.
127. Moukhametzianov R, Klare JP, Efremov R, Baeken C, Göppner A, Labahn J, Engelhard M, Büldt G, Gordeliy VI (2006). Development of the signal in sensory rhodopsin and its transfer to the cognate transducer. *Nature* 440, 115–119.
128. Mount D (2004). *Bioinformatics: Sequence and Genome Analysis*, Second Edi. CSH Press., New York, USA
129. Murtagh F, Legendre P (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J Classif* 31, 274–295.
130. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536–40.
131. Nagahara N (2014). Catalytic Site Cysteines of Thiol Enzyme : Sulfurtransferases. 2011, 6–11.
132. Nagano K (1973). Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. *J Mol Biol* 75, 401–20.
133. Nagy P (2013). Kinetics and mechanisms of thiol-disulfide exchange covering direct substitution and thiol oxidation-mediated pathways. *Antioxid Redox Signal* 18, 1623–41.
134. Nair S, Li W, Kong A-NT (2007). Natural dietary anti-cancer chemopreventive compounds: redox-mediated differential signaling mechanisms in cytoprotection of normal cells versus cytotoxicity in tumor cells. *Acta Pharmacol Sin* 28, 459–472.

135. Nardini M, Dijkstra BW (1999). Alpha/beta hydrolase fold enzymes: the family keeps growing. *Curr Opin Struct Biol* 9, 732–7.
136. Nathani RI, Moody P, Chudasama V, Smith MEB, Fitzmaurice RJ, Caddick S (2013). A novel approach to the site-selective dual labelling of a protein via chemoselective cysteine modification. *Chem Sci* 4, 3455.
137. Nielsen JE (2007). Analysing the pH-dependent properties of proteins using pKa calculations. *J Mol Graph Model* 25, 691–9.
138. Norel R, Sheinerman F, Petrey D, Honig B (2001). Electrostatic contributions to protein-protein interactions: fast energetic filters for docking and their physical basis. *Protein Sci* 10, 2147–61.
139. Pandey T, Singh SK, Chhetri G, Tripathi T, Singh AK (2015). Characterization of a Highly pH Stable Chi-Class Glutathione S-Transferase from *Synechocystis* PCC 6803. *PLoS One* 10, e0126811.
140. Papadopoulos JS, Agarwala R (2007). COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* 23, 1073–9.
141. Pascual-ahuir JL, Silla E, Tunon I (1994). GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface. *J Comput Chem* 15, 1127–1138.
142. Patzelt H, Simon B, terLaak A, Kessler B, Kühne R, Schmieder P, Oesterhelt D, Oschkinat H (2002). The structures of the active center in dark-adapted bacteriorhodopsin by solution-state NMR spectroscopy. *Proc Natl Acad Sci U S A* 99, 9765–70.
143. Pauling L (1988). *General Chemistry*. New York
144. Pérez-Cañadillas JM, Campos-Olivas R, Lacadena J, Martínez Del Pozo a, Gavilanes JG, Santoro J, Rico M, Bruix M (1998). Characterization of pKa values and titration shifts in the cytotoxic ribonuclease alpha-sarcin by NMR. Relationship between electrostatic

- interactions, structure, and catalytic function. *Biochemistry* 37, 15865–15876.
145. Peter T. Chivers, Martha C. A. Laboissiere and RTR (1998). In *Prolyl Hydroxylase, Protein Disulfide Isomerase, and Other Structurally Related Proteins*. Marcel Dekker., New York
  146. Petukh M, Stefl S, Alexov E (2013). The role of protonation states in ligand-receptor recognition and binding. *Curr Pharm Des* 19, 4182–90.
  147. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005). Scalable molecular dynamics with NAMD. *J Comput Chem* 26, 1781–802.
  148. Pinitglang S, Watts AB, Patel M, Reid JD, Noble MA, Gul S, Bokth A, Naeem A, Patel H, Thomas EW, et al (1997). A classical enzyme active center motif lacks catalytic competence until modulated electrostatically. *Biochemistry* 36, 9968–82.
  149. Pitera JW, Falta M, van Gunsteren WF (2001). Dielectric properties of proteins from simulation: the effects of solvent, ligands, pH, and temperature. *Biophys J* 80, 2546–55.
  150. Pokkuluri PR, Londer YY, Duke NEC, Long WC, Schiffer M (2004). Family of cytochrome *c7*-type proteins from *Geobacter sulfurreducens*: structure of one cytochrome *c7* at 1.45 Å resolution. *Biochemistry* 43, 849–59.
  151. Ponder JW, Richards FM (1987). An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J Comput Chem* 8, 1016–1024.
  152. Ponnuswamy PK, Prabhakaran M, Manavalan P (1980). Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim Biophys Acta* 623, 301–16.
  153. Poole LB, Karplus PA, Claiborne A (2004). PROTEIN SULFENIC ACIDS IN REDOX SIGNALING. *Annu Rev Pharmacol Toxicol* 44, 325–347.
  154. Porter CT, Bartlett GJ, Thornton JM (2004). The Catalytic Site Atlas: a resource of

- catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32, D129-33.
155. Privalov PL, Dragan AI, Crane-Robinson C (2011). Interpreting protein/DNA interactions: distinguishing specific from non-specific and electrostatic from non-electrostatic components. *Nucleic Acids Res* 39, 2483–91.
156. R.Barnes M, C.Gray I (2003). *Bioinformatics for Geneticists (Hierarchical Exotoxicology Mini Series)*, 1st ed. Wiley.
157. Rao KB, Simhadri NVV, Rao TS, Rao GN (2011). Influence of dielectric constant on protonation equilibria of L-methionine in aqueous solutions of propylene glycol and acetonitrile. *Der Pharma Chem* 3, 87–93.
158. Reed AE, Weinstock RB, Weinhold F (1985). Natural population analysis. *J Chem Phys* 83, 735.
159. Refaee M, Tezuka T, Akasaka K, Williamson MP (2003). Pressure-dependent changes in the solution structure of hen egg-white lysozyme. *J Mol Biol* 327, 857–65.
160. Rekker (1977). *The hydrophobic fragmental constant*. Elsevier., Amsterdam
161. RF R, Nauta W (1977). *The hydrophobic fragmental constant*. Elsevier., Amsterdam. The Netherlands
162. Rhee SG, Kang SW, Jeong W, Chang T-S, Yang K-S, Woo HA (2005). Intracellular messenger function of hydrogen peroxide and its regulation by peroxiredoxins. *Curr Opin Cell Biol* 17, 183–9.
163. Robson B, Pain RH (1972). Directional Information Transfer in Protein Helices. *Nature* 238, 107–108.
164. Samsonov S, Teyra J, Pisabarro MT (2008). A molecular dynamics approach to study the importance of solvent in protein interactions. *Proteins* 73, 515–25.

165. Schaftenaar G, Noordik JH (2000). Molden: a pre-and post-processing program for molecular and electronic structures\*. *J Comput Aided Mol Des* 14, 123–134.
166. Schmidt B, Ho L, Hogg PJ (2006). Allosteric disulfide bonds. *Biochemistry* 45, 7429–33.
167. Schmidt MW, Baldrige KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su S, et al (1993). General atomic and molecular electronic structure system. *J Comput Chem* 14, 1347–1363.
168. Seegan GW, Smith CA, Schumaker VN (1979). Changes in quaternary structure of IgG upon reduction of the interheavy-chain disulfide bond. *Proc Natl Acad Sci U S A* 76, 907–11.
169. Shan J, Mehler EL (2011). Calculation of pK(a) in proteins with the microenvironment modulated-screened coulomb potential. *Proteins* 79, 3346–55.
170. Shaw A, Bott R, Vornrhein C, Bricogne G, Power S, Day AG (2002). A novel combination of two classic catalytic schemes. *J Mol Biol* 320, 303–9.
171. Shirabe K, Yubisui T, Nishino T, Takeshita M (1991). Role of cysteine residues in human NADH-cytochrome b5 reductase studied by site-directed mutagenesis. Cys-273 and Cys-283 are located close to the NADH-binding site but are not catalytically essential. *J Biol Chem* 266, 7531–6.
172. De Simone A, Berisio R, Zagari A, Vitagliano L (2006). Limited tendency of alpha-helical residues to form disulfide bridges: a structural explanation. *J Pept Sci* 12, 740–7.
173. Simone ADE, Berisio R, Zagari A, Vitagliano L (2006). Limited tendency of  $\alpha$  -helical residues to form disulfide bridges : a structural explanation. 740–747.
174. Suzuki T, Kudo Y (1990). Automatic log P estimation based on combined additive modeling methods. *J Comput Aided Mol Des* 4, 155–98.
175. Takashima S, Schwan HP (1965). Dielectric Dispersion of Crystalline Powders of Amino Acids, Peptides, and Proteins 1. *J Phys Chem* 69, 4176–4182.



176. Tanaka S, Scheraga HA (1976). Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules* 9, 945–950.
177. Tanford C (1962). Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. *J Am Chem Soc* 84, 4240–4247.
178. Tanford C, Roxby R (1972). Interpretation of protein titration curves. Application to lysozyme. *Biochemistry* 11, 2192–8.
179. Tew KD (2007). Redox in redux: Emergent roles for glutathione S-transferase P (GSTP) in regulation of cell signaling and S-glutathionylation. *Biochem Pharmacol* 73, 1257–69.
180. Teyra J, Doms A, Schroeder M, Pisabarro MT, Phizicky E, Fields S, Bader G, Betel D, Hogue C, Zanzoni A, et al (2006). SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics* 7, 104.
181. Thangudu RR, Manoharan M, Srinivasan N, Cadet F, Sowdhamini R, Offmann B (2008). Analysis on conservation of disulfide bonds and their structural features in homologous protein domain families. *BMC Struct Biol* 8, 55.
182. Thellamurege NM, Si D, Cui F, Zhu H, Lai R, Li H (2013). QuanPol: a full spectrum and seamless QM/MM program. *J Comput Chem* 34, 2816–33.
183. Tolosa EA, Chepurnova NK, Khomutov RM, Severin ES (1969). Reactions catalysed by cysteine lyase from the yolk sac of chicken embryo. *Biochim Biophys Acta - Enzymol* 171, 369–371.
184. Townsend DM (2007). S-glutathionylation: indicator of cell stress and regulator of the unfolded protein response. *Mol Interv* 7, 313–24.
185. Trowell HC (1982). Ants distinguish diabetes mellitus from diabetes insipidus. *Br Med J (Clin Res Ed)* 285, 217.
186. Tryon, R. C. and Bailey DE (1973). *Cluster Analysis*. McGraw-Hill., New York, USA

187. Varlamova EG, Goltyaev M V., Novoselov S V., Novoselov VI, Fesenko EE (2013). Characterization of several members of the thiol oxidoreductase family. *Mol Biol* 47, 496–508.
188. Varma S, Jakobsson E (2004). Ionization states of residues in OmpF and mutants: effects of dielectric constant and interactions between residues. *Biophys J* 86, 690–704.
189. Voet D, Voet JG, Pratt CW (2008). *Principles of Biochemistry*. Wiley.
190. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004). Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14, 208–16.
191. Ward J (1963). Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* 58, 236–244.
192. Waring AJ, Faull KF, Leung C, Chang-Chien A, Mercado P, Taeusch HW, Gordon LM (1996). Synthesis, secondary structure and folding of the bend region of lung surfactant protein B. *Pept Res* 9, 28–39.
193. Watanabe A, Yoshimura T, Mikami B, Hayashi H, Kagamiyama H, Esaki N (2002). Reaction mechanism of alanine racemase from *Bacillus stearothermophilus*: x-ray crystallographic studies of the enzyme bound with N-(5'-phosphopyridoxyl)alanine. *J Biol Chem* 277, 19166–72.
194. Westheimer FH, Shookhoff MW (1939). The Electrostatic Influence of Substituents on the Dissociation Constants of Organic Acids. III. *J Am Chem Soc* 61, 555–560.
195. Wilkinson B, Gilbert HF (2004). Protein disulfide isomerase. *Biochim Biophys Acta* 1699, 35–44.
196. Wilson A, Tobin D (2010). *Aging Hair*, 1st ed. Springer., New York, USA
197. Wimley WC, Creamer TP, White SH (1996). Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* 35, 5109–24.

198. Winther JR, Thorpe C (2014). Quantification of thiols and disulfides. *Biochim Biophys Acta* 1840, 838–46.
199. Woo HA, Jeong W, Chang T-S, Park KJ, Park SJ, Yang JS, Rhee SG (2005). Reduction of cysteine sulfinic acid by sulfiredoxin is specific to 2-cys peroxiredoxins. *J Biol Chem* 280, 3125–8.
200. Wood EJ (1996). Harper's biochemistry 24th edition. *Biochem Educ* 24, 237.
201. Wu C, Belenda C, Leroux J-C, Gauthier MA (2011). Interplay of chemical microenvironment and redox environment on thiol-disulfide exchange kinetics. *Chemistry* 17, 10064–70.
202. Wu S, Liu T, Altman RB (2010). Identification of recurring protein structure microenvironments and discovery of novel functional sites around CYS residues. *BMC Struct Biol* 10, 4.
203. Wyckoff HW, Tsernoglou D, Hanson AW, Knox JR, Lee B, Richards FM (1970). The three-dimensional structure of ribonuclease-S. Interpretation of an electron density map at a nominal resolution of 2 Å. *J Biol Chem* 245, 305–28.
204. Ye L, Kuang Q, Jiang L, Luo J, Jiang Y, Ding Z, Li Y, Li M (2014). Prediction of hot spots residues in protein–protein interface using network feature and microenvironment feature. *Chemom Intell Lab Syst* 131, 16–21.
205. Zhang C, Kim SH (2000). Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci U S A* 97, 2550–5.
206. Zheng L, White RH, Cash VL, Jack RF, Dean DR (1993). Cysteine desulfurase activity indicates a role for NIFS in metallocluster biosynthesis. *Proc Natl Acad Sci U S A* 90, 2754–8.
207. Zhou HX, Shan Y (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* 44, 336–43.

208. Zimmerman JM, Eliezer N, Simha R (1968). The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* 21, 170–201.

## List of publications and conferences

- 1) Bhatnagar A, Apostol MI, Bandyopadhyay D (2016). Amino acid function relates to its embedded protein microenvironment: A study on disulfide-bridged cystine. *Proteins*. doi: 10.1002/prot.25101.
- 2) Bhatnagar A, Varanasi S, Pramod Ghiya D, Gali Sai Ganesh C, Bandyopadhyay D (2016). Protonation and Deprotonation Reaction of Aspartic Acid Side Chain Modulated by the Surrounding Dielectric Medium - AB Initio Quantum Chemical Studies on Aspartic Acid in Sixteen Different Solvents and Two Protein Structures. *Biophys J* 110, 380a–381a.
- 3) Akshay Bhatnagar and Debashree Bandyopadhyay, “Microenvironment based functional preferences of cystine residues in high resolution protein crystal structures”, Proceedings of 39th National Symposium on Biophysics and Golden Jubilee Meeting of Indian Biophysical Society 2015, *Journal of Proteins and Proteomics* ISSN : 0975-8151; 6(1), 2015.
- 4) Role of Microenvironment in modulating structure and function of Cystine residues in high-resolution protein crystal structures, Akshay Bhatnagar and Debashree Bandyopadhyay, Annual Meeting of the Indian Biophysical Society, February, 7-10, 2014, Kolkata, India
- 5) Akshay Bhatnagar and Debashree Bandyopadhyay, “Role of Microenvironment in modulating structure and function of Cystine residues in high-resolution protein crystal structures”, Proceedings of 4th Annual International Conference on Advances in Biotechnology 2014, pages 35-40.
- 6) Characterization of cysteine functions based on local secondary structures and embedded protein microenvironments (Manuscript to be submitted)
- 7) Role of microenvironment in modulating cysteine thiol functions in cytochromes and zinc finger proteins (Manuscript to be submitted)

## **Biography of the supervisor**

Dr. Debashree Bandyopadhyay is a chemist by academic training and computational biophysicist by professional training. Her career interest is two-fold, advancement in computational structural biology research and training students to develop scientific cult and excellence.

She gathered a broad spectrum of experience in research in highly reputed educational institutes hosted by Singapore (Bioinformatics Institute, A-STAR), USA (Weill Medical College of Cornell University, New York) and India (Saha Institute of Nuclear Physics) within the period of 1997 to 2010. Her research interest is understanding structure-function relationship in biological molecules and macro-molecular assemblies and possible prevention of few diseases. Her research expertise cover, molecular modelling, molecular dynamics simulation, quantum chemical calculations and development of softwares to address specific biological problems.

Contributions are made in fields of chemistry, biophysics, bioinformatics and structural biology. Accomplishments can be judged in terms of twelve international publications with major contributions and in twenty-two conference proceedings at national and international scientific meetings. Publications are worthy enough to be referred by other scientific articles and reviews (Accounts of Chemical Research, 2002, 35, 350-357), including "Science" (Science, 2007, 315, 1549-1553).

She has served as a Lecturer at Department of Chemistry, Raja Peary Mohan College, Uttarpara, Hooghly, West Bengal, India from 2002 to 2005 and as an Assistant Professor at Department of Biological Sciences, BITS PILANI, from 2012 onwards. Chemistry courses offered by her are, Atomic structure, Periodic Table, Group Chemistry, Quantum Chemistry, Radioactivity and Bioinorganic Chemistry. At BITS PILANI courses offered by her are, Bioinformatics, Biophysics, Biomolecular Modeling, General Biology and Biology Laboratory.

In the last four years at BITS PILANI, she has guided total 23 M.Sc (Biological Sciences) project students, 9 M.Sc (Biological Sciences) thesis student and guided one Ph.D student.

### **Biography of the candidate**

Mr. Akshay Bhatnagar is a full time Ph.D. student at BITS Pilani Hyderabad Campus under the supervision of Dr. Debashree Bandyopadhyay at Biological Sciences department. He has completed his B.Tech from Amity University Rajasthan in Bioinformatics. His research interests are proteome analysis and understand the mechanism of enzymatic reaction and their application in medical biology. He has presented his Ph.D. work in two international and two national conferences. He has one published research article in the journal “Proteins: Structure, Function, and Bioinformatics”. He is a member of Indian biophysical Society and Biophysical Society.

# Amino acid function relates to its embedded protein microenvironment: A study on disulfide-bridged cystine

Akshay Bhatnagar,<sup>1</sup> Marcin I. Apostol,<sup>2</sup> and Debashree Bandyopadhyay<sup>1\*</sup>

<sup>1</sup> Department of Biological Sciences, Birla Institute of Technology and Science, Hyderabad 500078, India

<sup>2</sup> ADRx. Inc. 515 Marin St., Suite 314, Thousand Oaks, California 91360

## ABSTRACT

In our previous study, we have shown that the microenvironments around conserved amino acids are also conserved in protein families (Bandyopadhyay and Mehler, *Proteins* 2008; 72:646–659). In this study, we have hypothesized that amino acids perform similar functions when embedded in a certain type of protein microenvironment. We have tested this hypothesis on the microenvironments around disulfide-bridged cysteines from high-resolution protein crystal structures. Although such cystines mainly play structural role in proteins, in certain enzymes they participate in catalysis and redox reactions. We have performed and report a functional annotation of enzymatically active cystines to their respective microenvironments. Three protein microenvironment clusters were identified: (i) buried-hydrophobic, (ii) exposed-hydrophilic, and (iii) buried-hydrophilic. The buried-hydrophobic cluster encompasses a small group of 22 redox-active cystines, mostly in alpha-helical conformations in a –C-x-x-C- motif from the Oxido-reductase enzyme class. All these cystines have high strain energy and near identical microenvironments. Most of the active cystines in hydrolase enzyme class belong to buried hydrophilic microenvironment cluster. In total there are 34 half-cystines detected in buried hydrophilic cluster from hydrolases, as a part of enzyme active site. Even within the buried hydrophilic cluster, there is clear separation of active half-cystines between surface exposed part of the protein and protein interior. Half-cystines toward the surface exposed region are higher in number compared to those in protein interior. Apart from cystines at the active sites of the enzymes, many more half-cystines were detected in buried hydrophilic cluster those are part of the microenvironment of enzyme active sites. However, no active half-cystines were detected in extremely hydrophilic microenvironment cluster, that is, exposed hydrophilic cluster, indicating that total exposure of cystine toward the solvent is not favored for enzymatic reactions. Although half-cystines in exposed-hydrophilic clusters occasionally stabilize enzyme active sites, as a part of their microenvironments. Analysis performed in this work revealed that cystines as a part of active sites in specific enzyme families or folds share very similar protein microenvironment regions, despite of their dissimilarity in protein sequences and position specific sequence conservations.

*Proteins* 2016; 84:1576–1589.  
© 2016 Wiley Periodicals, Inc.

**Key words:** amino acid function; embedded protein microenvironment; disulphide bridged cystine; half-cystine; redox active cystine; enzyme class; enzyme active site; protein microenvironment cluster; protein dielectric medium; sequence conservation.

## INTRODUCTION

Properties of amino acids are influenced by the surrounding dielectric medium, like water/octanol,<sup>1</sup> protein interior<sup>2–4</sup> and membrane bilayers.<sup>5–8</sup> The protein interior is highly heterogeneous in terms of its dielectric medium; the environment created by the complex mosaic of 20 different amino acids. This heterogeneous protein dielectric medium can be viewed as a collection of smaller dielectric media, what we refer to here as the

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: University Grants Commission; Basic Research Start-up grant, India; BITS-PILANI, research initiation Grant.

\*Correspondence to: Debashree Bandyopadhyay, Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, Hyderabad campus, Hyderabad 500078, India. E-mail: banerjee\_debi@yahoo.com; banerjee.debi@hyderabad.bits-pilani.ac.in

Received 2 April 2015; Revised 30 June 2016; Accepted 3 July 2016  
Published online 13 July 2016 in Wiley Online Library (wileyonlinelibrary.com).  
DOI: 10.1002/prot.25101



local microenvironment. Microenvironment describes the three dimensional arrangement of atoms around any given amino acid (or its functional group) within its first contact shell.<sup>9</sup> It has been shown earlier that certain properties of amino acids, for example  $pK_a$  in titrable amino acids, can be altered due to transition between different microenvironments.<sup>4</sup> It has also been shown that protein microenvironments are crucial in modulating the protonation states of titratable amino acids such as arginine in voltage-gated ion channels,<sup>10</sup> aspartic acid in photo cycle of bacteriorhodopsin,<sup>11</sup> glutamic acid in hen egg white lysozyme, and others in several acid-base catalyzed hydrolysis reactions.<sup>12</sup> Despite of all these facts, there is a gap in knowledge of how specifically microenvironment influences the particular role of an amino acid in a protein structure. This is due to insufficient biochemical (experimental) information on individual amino acid microenvironment inside the protein. Microenvironment measurement is experimentally limited only to certain surface exposed amino acid residues which are intrinsic fluorophores or those can be tagged with extrinsic fluorophores.<sup>13</sup> Recently we have generated a database of microenvironments around all the amino acid side chains in the context of high-resolution protein crystal structures.<sup>9</sup> In that work, we have demonstrated that each amino acid side-chain is embedded in a wide range of protein microenvironment. The average hydrophobicity of the microenvironment is similar to that of the embedded amino acid side chain. 20–30% of the protein microenvironments significantly differ from that of the corresponding amino acid, those termed as “mismatched microenvironment.” It has also been shown that amino acids in these mismatched microenvironments were evolved to perform specific structure or functional roles in the protein. Moreover that work has also demonstrated that the microenvironments around the conserved residues in respective protein family are also conserved. These observations triggered the question whether different sets of microenvironments are associated with different kinds of functions in an amino acid. Hypothesis based on this question has been tested here on disulfide-bridged cystine residues embedded in the protein microenvironments obtained from an updated microenvironment database. Cystine residues occur by the oxidation of two cysteine amino acids covalently linking them through a disulfide bond. This was chosen as model system because this residue has a limited and simple functional role in proteins, it either plays a structural role or in an enzymatic capacity participates in redox reactions. Out of the six broad enzyme classes (according to the nomenclature committee of International Union of Biochemistry and Molecular Biology, IUBMB)<sup>14</sup> cystines are directly involved in catalytic and redox reactions in the oxidoreductase (thioreductase fold)<sup>15</sup> and hydrolase families.<sup>16</sup>

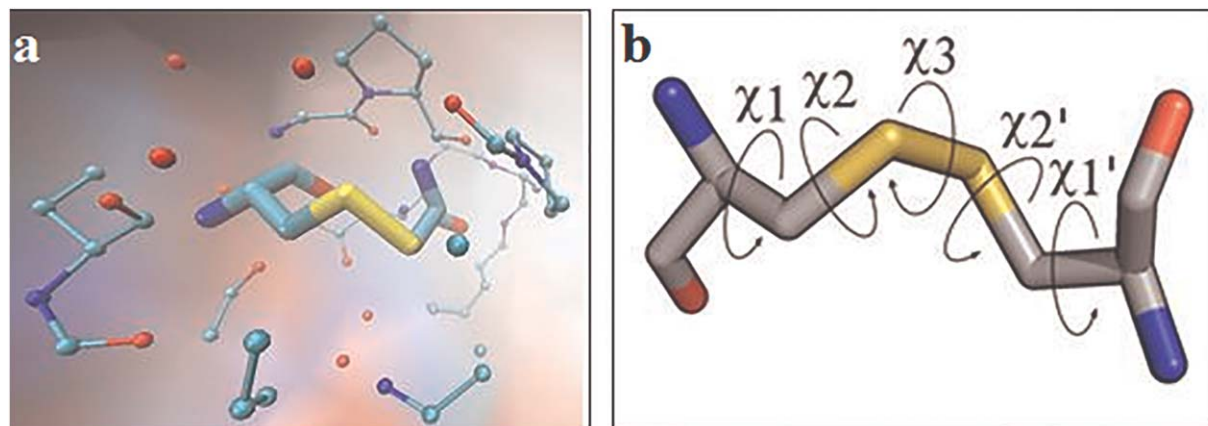
Here, we analyze the previously unexplored cystine functions (here functions imply presence of a cystine in enzyme active site or within its microenvironment) in various protein microenvironments. In this study, protein microenvironments were clustered using hierarchical clustering. Functions in entire microenvironment database (mainly out of the enzymes) were curated from literature and different websites. Grouping of cystine functions based on their microenvironment clusters are discussed in the result section. The notable observation from this study is that the cystines those are part of active sites in specific enzyme classes share similar microenvironments despite of their dissimilarity in overall protein sequences.

## METHODOLOGY

### Description of microenvironment dataset around disulfide-bridged cystine residues

The current microenvironment dataset around disulfide-bridged cystine residues contains total 5084 cystines from 1303 high resolution protein crystal structures. Part of the dataset was obtained from our previous work, containing 700 disulfide-bridged cystine residues from 175 protein crystal structures.<sup>9</sup> Remaining 4384 cystines were curated from recent PDB<sup>17</sup> entries, from January 2004 to April 2016. Protein crystal structures with resolution better than 2 Å and sequence similarity <30% were selected. These proteins were not complexed with nucleic acid molecules. Proteins containing modified residues were not included in the dataset. The current selection resulted into 9090 protein structures. Out of this current selection, only 1128 proteins were detected with cystines from PDB header files using SSBOND keyword (see Supporting Information for PDB IDs, Table SI).

The microenvironment around each cystine was described by the amino acid fragments within a given radius, where the radius varies from one atom type to the other [Fig. 1(a)]. The microenvironment space around individual cystine was described using two parameters, “buried fraction” (BF) and hydrophobicity/hydrophilicity (rHpy).<sup>9</sup> Buried fraction is the fraction of the side-chain of an amino acid which is buried inside the protein and not exposed to outside environment. This parameter was computed using the program GEPOL93 that computes the solvent-excluded surface by filling the solvent-inaccessible spaces with a new set of spheres.<sup>19</sup> Three parameters were used in the computation, the number of triangles per sphere, maximum overlap among the new spheres and the size of the smallest sphere. Default values for these parameters, 3, 0.8, and 0.5, were used. The amino acid side-chain which is completely embedded inside a protein structure will have buried fraction value equal to one and zero when



**Figure 1**

(a) Schematic representation of microenvironment around disulfide-bridged cystine residue. The cystine is shown as tube representation. The microenvironment around the cystine is shown in ball and stick representation. Oxygen atoms of water molecules are shown as isolated red balls. The protein background is shown as surface representation. The schema is created using VMD.<sup>18</sup> (b) Schematic representation of different dihedral angles observed in disulfide-bridged cystine structure. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

completely exposed to the solvent. “rHpy” is a quantitative property descriptor (QPD) of microenvironment around a side-chain or fragment of an amino acid.<sup>9</sup> This QPD estimates the relative hydrophobicity around the amino acid fragment immersed in protein environment with respect to the same in water environment, within a given distance threshold, (that is the radius of the first contact shell around the atom). For example, the microenvironment radius described around a carbon atom is 4.475 Å.<sup>1</sup> From the definition of rHpy, an amino acid side-chain have maximum rHpy value of one when it is completely exposed to water. By definition, there is no absolute minimum rHpy value. Minimum rHpy value varies among proteins and that is decided by the hydrophobicity of individual protein interior. The lowest rHpy value observed in this microenvironment dataset is  $-0.4$ .

#### Curation of cystine functionalities in protein structures from literature

Functions of 5084 disulfide-bridged cystines were extracted from published literature, PDB header files,<sup>17</sup> Catalytic Site Atlas (CSA)<sup>20</sup> and PDBSUM<sup>38</sup> database. Each sulfur atom from a cystine is defined as a half-cystine through-out this article. The half-cystines were defined functional when those are marked as “catalytic sites” or as “active site” in PDBSUM website.<sup>38</sup> Catalytic sites were defined by the Catalytic Site Atlas (CSA).<sup>20</sup> Active sites were defined by PDBSUM<sup>38</sup> from the SITE records of the PDB header files.<sup>17</sup> The analysis of functional half-cystines has been performed over all the available enzyme classes in the current data set. A total of 617 enzymes containing cystines were analyzed (Supporting Information Table SII).

#### Clustering of microenvironment space

Clustering methods are mainly of two types: partitioning and hierarchical.<sup>21</sup> In this study, agglomerative clustering, a subclass of hierarchical clustering has been employed.<sup>22</sup> Agglomerative clustering is a type of hierarchical clustering where each observation is initially considered as a single cluster; based on the distance proximity of the nearest cluster centers, the smaller clusters combine to form large dissimilar clusters. Initially, the microenvironment space was divided into small bins with equal spacing [(BF, rHpy) = (0.1, 0.1)]. Agglomerative clustering method incorporates small microenvironment bins into larger dissimilar clusters where the Euclidean distance of each bin from cluster center is minimal with respect to other cluster centers. Ward’s method<sup>23</sup> of agglomerative clustering was employed.<sup>24</sup>

#### Identification of cystine residues in different structural classification of protein (SCOP) classes

Protein classes for all the 617 enzymes were attempted to extract from Structural Classification of Proteins (SCOP) database.<sup>25</sup> As latest SCOP database entry dated back to 2009 and our current selection dated, April 2016, SCOP classification was not available from the same database, for some PDB entries. In those cases, structural classification was done based on PDBeFold.<sup>26</sup> The secondary structures were calculated for individual proteins using Kabsch and Sander algorithm<sup>27</sup> in the Dictionary of Protein Secondary Structure (DSSP) calculation.<sup>28</sup>

#### Analysis of cystine geometry in hydrophobic and hydrophilic microenvironments

Disulfide-bridged cystine structures are governed by their internal conformations, the five dihedral angles,

**Table 1**Description of Different Microenvironment Clusters Around Half-Cystines, (S-S)<sub>1/2</sub>

Cluster	Center <sup>a</sup>	D <sup>b</sup> (Å)	V <sup>c</sup>	Cluster-size	N <sup>d</sup>
Buried hydrophobic	0.947, 0.164	0.212	0.063	8549	1071 (12.5)
Buried hydrophilic	0.770, 0.482	0.186	0.043	1465	960 (66)
Exposed hydrophilic	0.328, 0.720	0.145	0.031	154	142 (92)

Clusters are defined in terms of buried fraction and rHpy using agglomerative hierarchical clustering.<sup>21</sup> Number of other half-cysteine (partner in the disulfide) is also reported along with the normalized values in parenthesis. Clusters are arranged, according to descending order of hydrophobicity, measured by cluster center values.

<sup>a</sup>Centroid values of buried fraction, rHpy.

<sup>b</sup>Average distance to center.

<sup>c</sup>Within-class variance.

<sup>d</sup>No. of cystines present in different clusters (Normalized value in %).

namely chi1, chi2, chi3, chi1' and chi2' [Fig. 1(b)]. Based on these dihedral angles, dihedral strain energies<sup>29</sup> were computed for all the 5084 cystines in the current dataset. The program used to compute the disulfide dihedral strain energy (DSE) and is publicly available as a server (<http://services.mbi.ucla.edu/disulfide/>).

### Sequence and structural alignment of proteins in the dataset

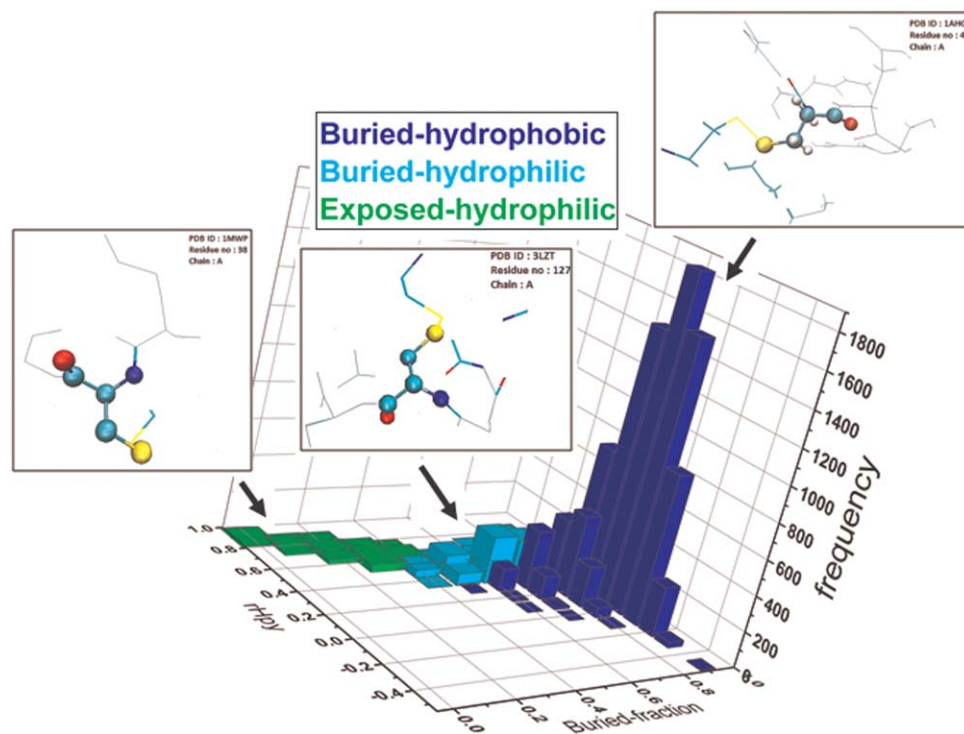
Sequence and structural alignment were performed for two enzyme classes; oxidoreductases and hydrolases. Global

sequence alignment was performed using T-COFFEE<sup>30</sup> and local alignment was performed using COBALT.<sup>31</sup> The structural alignment was performed using SALIGN.<sup>32</sup>

## RESULTS AND DISCUSSIONS

### Validation of the updated microenvironment dataset around disulphide (TD<sup>S</sup>) group of cystine

Microenvironment dataset around disulphide group of cystine amino acid, generated based on the current PDB

**Figure 2**

Distribution of different microenvironment clusters (obtained from hierarchical clustering) around all the half-cystines from 1478 different proteins in our updated microenvironment dataset. Microenvironments are being clustered in the space of buried fraction and rHpy values. Frequencies in each cluster are shown along z-axis. The plot was generated using Origin [Origin (OriginLab, Northampton, MA)]. Insets show ball and stick representations of representative microenvironments around half-cystines from the 3 different clusters. Thicker sticks highlight bonds within the microenvironment around the sulfur atoms and thin sticks represent the extended microenvironment around the cystine residue.

**Table II**

Distribution of Secondary Structures in Different Microenvironment Clusters, Results Obtained from DSSP Analysis

	Buried hydrophobic	Buried hydrophilic	Exposed hydrophilic	Total
Helix (G+H+I)	1976 (0.23)	322 (0.22)	33 (0.21)	2331
Beta strand (B+E)	2968 (0.35)	268 (0.18)	32 (0.21)	3268
Turn (T)	630 (0.07)	183 (0.12)	18 (0.12)	831
Coil (C)	2251 (0.26)	536 (0.37)	58 (0.38)	2845
Bend (S)	724 (0.08)	156 (0.11)	13 (0.08)	893

The values given in parenthesis are normalized with respect to the total number of (S-S)<sub>1/2</sub> in a particular cluster.

entries, was validated with respect to our previously observed dataset.<sup>9</sup> Buried fraction and rHpy, the two parameters used throughout this work to describe microenvironment, were compared across the previous and the current datasets. Mean and standard deviation values (given within the parenthesis) for buried fraction and rHpy, in the current dataset are 0.91 (0.13) and 0.22 (0.18), respectively. These are in close agreement to the values from the previous dataset (0.92 (0.12) and 0.21 (0.16) respectively; Table III in reference 1 for TD group). Agreement of the microenvironment descriptors in current and previous datasets indicates that despite of the dataset size, overall nature of the protein microenvironments around cystines remain the same. In addition to the mean values of these property descriptors, percentage of cystines microenvironment outliers, (outliers are defined as those outside of one standard deviation with respect to the mean value) are also compared across previous and current datasets. Total number of outliers in the current dataset is 906; 8.9% of the total dataset; this value is fairly in agreement with that in the previous dataset (9.3%; Table I in reference 1).

### Microenvironment clusters around cystine residues distributed in high-resolution crystal structures with 25–30% sequence similarity

Microenvironment clusters around half-cystines -(S-S)<sub>1/2</sub>- were obtained using agglomerative clustering, resulting into three categories (i) buried-hydrophobic,

(ii) buried-hydrophilic, and (iii) exposed-hydrophilic (Table I). These microenvironment clusters gradually progress from protein interior to surface (decrease in BF) with increasing hydrophilic character (increase in rHpy) (Fig. 2).

Cystine residues in buried-hydrophobic cluster are observed inside the protein interior, (average buried fraction, 0.95) (Supporting Information Fig. S1). The buried-hydrophobic cluster mainly contains half-cystines which are buried deep inside the protein core and embedded in hydrophobic microenvironment (average rHpy 0.16) (Fig. 2 and Supporting Information Fig. S1). Conversely, half-cystines in the buried-hydrophilic cluster are buried inside the protein (average buried fraction, 0.77) which constitutes relatively hydrophilic protein microenvironment (average rHpy, 0.48). Half-cystines in the exposed-hydrophilic cluster are observed on the outer surface of proteins (average buried fraction of 0.33) and have relatively high hydrophilic microenvironment (average rHpy 0.72). Interestingly, cluster sizes (number of half-cystines in each cluster) gradually decreases with hydrophilicity in the clusters (Table I), indicating that cystine residues tend to be in buried hydrophobic microenvironments.<sup>9</sup>

Both half-cystines derived from the same disulfide bonded cystine not necessarily fall into the same microenvironment cluster. The second sulfur atom from the disulfide bond tend to stay in the same microenvironment cluster, provided the later is hydrophobic (Table I).

### Correlation between secondary, super-secondary structures and microenvironment clusters around half-cystines

The correlation between secondary structures of half-cystines and respective protein folds was known.<sup>33</sup> Here we aim to find out whether half-cystines with similar secondary structures (e.g., alpha helix, beta sheet, or coil) populate into similar microenvironment clusters. Microenvironment can be viewed as the local tertiary structure around a particular amino acid which might include different components of secondary structures. Therefore, there is a possibility of underlying relationship between

**Table III**Description of Different Microenvironment Clusters Populated with -(S-S)<sub>1/2</sub> - from Different Enzyme Classes

Cluster	E	Oxidoreductases <sup>a</sup>	Hydrolase	Lyase	Transferase	Isomerase
Buried hydrophobic	3571 <sup>b</sup>	571 (0.16)	2471 (0.69)	97 (0.03)	355 (0.1)	59 (0.02)
Buried hydrophilic	717 <sup>c</sup>	104 (0.15)	520 (0.73)	29 (0.04)	54 (0.08)	4 (0.01)
Exposed hydrophilic	31	0 (0.0)	28 (0.90)	0 (0.0)	2 (0.06)	1 (0.03)

The total numbers of -(S-S)<sub>1/2</sub> -, present in an enzyme class in each cluster (E), are reported. The number of -(S-S)<sub>1/2</sub> - in individual enzyme classes are also reported, along with their normalized values, given in parenthesis. The values are normalized with respect to the total number of half-cystines present in a particular microenvironment cluster as part of an enzymes class (E).

<sup>a</sup>And Electron transport proteins.

<sup>b</sup>18 cystines were observed as part of ligase.

<sup>c</sup>6 cystines were observed as a part of ligase.

**Table IV**

Number of  $-(S-S)_{1/2}$  - with Different Main Chain Conformations Present in Enzyme Classes—Hydrolases, Oxidoreductases and Transferases

Enzyme class	Total number of cysteines	Alpha helix	Beta sheet	Loop
Hydrolases	3019	656	869	1495
Oxidoreductases and 675 Electron transport proteins	209		91	375
Transferases	411	77	124	210

super-secondary structures (protein classes) and amino acid microenvironment, particularly, around disulphide-bridged cysteines which dictates protein scaffolds. It has been shown earlier that similar disulphide bonding patterns lead to similar fold, families and super-families of proteins, despite of their low sequence identity.<sup>34</sup> To understand the main-chain conformations of half-cysteines in different microenvironments, DSSP secondary structure analyses were performed on our protein dataset. Preferences of different secondary structures around half-cysteines in three different microenvironment clusters were observed and are described below (Table II).

The buried-hydrophobic cluster possesses higher percentage of beta secondary structures (and to lesser extent

alpha helical structures). The buried-hydrophilic cluster predominates in coil conformations. The exposed-hydrophilic cluster mainly possesses cysteines in coil conformations. These results have suggested that the half-cysteines buried in hydrophobic microenvironment are mainly comprised of alpha and beta secondary structures, whereas, partly exposed half-cysteines embedded in hydrophilic microenvironment, predominate in flexible conformations, like coils. Here we have attempted to identify microenvironment specific functions of cysteines observed with different secondary structures. As the functions of cysteines are mainly associated with enzymes, in the following section we have analyzed cystine functions for 617 enzymes from all the enzyme classes<sup>14</sup> present in our current dataset.

### Buried-hydrophobic microenvironment cluster hosts half-cysteines from redox-active -C-x-x-C- motif

The half-cysteines present in buried-hydrophobic microenvironment cluster, mainly come from (i) oxidoreductase and electron transport proteins, (ii) hydrolase, (iii) lyase, (iv) transferase, and (v) isomerase enzyme classes (Table III). Hydrolase is the most frequently

**Table V**

Buried Hydrophobic Microenvironment Cluster: Secondary Structures and Strain Energies of Redox-Active Cysteines (Part of C-x-x-C motif), Present in Different Oxidoreductases Enzymes and Electron Transport Proteins

PDB	Fold SCOP or PDBeFOLD <sup>a</sup>	Cystine Numbers (Chain)	Half-cystine in helix (Chain) [cluster no.]	Half-cystine in beta turn <sup>b</sup> (Chain) [cluster no.]	Strain Energy (KJ/mol)	BF	rHpy
3ZIT	Thioredoxin <sup>a</sup>	12 (A)–15(A)	15 (A) [2]	12 (A) [1]	11.4	0.732	0.375
3GWN	a <sup>c</sup>	80(A)–83(A)	83 (A) [2]	80 (A) [2] <sup>d</sup>	19.5	0.751	0.275
3GWN	a <sup>c</sup>	80(B)–83(B)	83 (B) [2]	80 (B) [2] <sup>d</sup>	21.5	0.800	0.228
4HS1	Thioredoxin	11(A)–14(A)	14 (A) [2]	11 (A) [2]	13.0	0.804	0.194
3POK	b <sup>e</sup>	155 (A)–158(A)	158 (A) [2]	155 (A) [2]	8.23	0.805	0.162
3ZIT	Thioredoxin <sup>a</sup>	12 (B)–15(B)	15 (B) [2]	12 (B) [1]	11.6	0.871	0.33
3FZ4	Thioredoxin <sup>a</sup>	10 (A)–13(A)	13(A) [2]	10 (A) [2]	10.5	0.926	0.095
1THX	Thioredoxin	32(A)–35(A)	35(A) [2]	32 (A) [1]	15.4	0.953	0.106
1JR8	a <sup>c</sup>	54(A)–57(A)	57 (A) [1]	54 (A) [2] <sup>d</sup>	19.1	0.960	0.045
1H75	Thioredoxin	11(A)–14(A)	14 (A) [2]	11 (A) [2]	13.5	0.967	0.192
2I4A	Thioredoxin <sup>a</sup>	32(A)–35(A)	35 (A) [2]	32 (A) [2]	11.4	0.978	0.029
2B1L	Thioredoxin	80(B)–83(B)	83 (B) [2]	80 (B) [1]	25.8	0.984	0.149
2B1L	Thioredoxin	80(A)–83(A)	83 (A) [2]	80 (A) [1]	25.0	0.986	0.142
1FVK	Thioredoxin	30(B)–33(B),	33(B) [2]	30 (A) [1]	14.3	0.988	0.065
1FVK	Thioredoxin	30(A)–33(A),	33(A) [2]	30 (A) [1]	19.2	0.993	–0.021
1KNG	Thioredoxin	92(A)–95(A)	95 (A) [2]	92 (A) [2]	19.3	0.996	–0.017
1ABA	Thioredoxin	14(A)–17(A)	17 (A) [2]	14 (A) [1]	12.4	0.997	0.130
2HLS	Thioredoxin <sup>a</sup>	150(B)–153(B)	153 (B) [2]	150 (B) [1]	10.3	0.998	0.078
1ST9	Thioredoxin	73(A)–76(A)	76 (A) [2]	73 (A) [2]	14.8	0.999	0.077
1ST9	Thioredoxin	73(B)–76(B)	76 (B) [2]	73 (A) [2]	13.0	0.999	0.119
2HLS	Thioredoxin <sup>a</sup>	150(A)–153(A)	153 (A) [2]	150 (A) [1]	12.2	1.000	0.117
3IOS	Thioredoxin <sup>a</sup>	81 (A)–84(A)	84 (A) [2]	81 (A) [2]	- <sup>f</sup>	1,000	0.015

Residue numbers of half-cysteines corresponding to each cystine are shown (column 3). Respective cluster numbers are given in parenthesis. Buried hydrophobic cluster is referred to as 2 and buried hydrophilic cluster is referred to as 1 here. Folds of individual residues are mentioned either from SCOP<sup>25</sup> and PDBeFOLD.<sup>26</sup>

<sup>a</sup>Obtained from PDBeFold.<sup>26</sup>

<sup>b</sup>Unless otherwise mentioned; according to DSSP program.<sup>27,28</sup>

<sup>c</sup>Represents Four-bundle-up-and-down bundle fold.

<sup>d</sup>Half-cystine present in Gamma turn instead of beta turn.

<sup>e</sup>Represents spectrin repeat like fold.

<sup>f</sup>Strain energy cannot be calculated by the "Disulfide Bond Dihedral Angle Energy Server."<sup>35</sup>

**Table VI**

Half-Cystines (First Out of the Pair of Half-cystines from a Disulphide Shown in the Table) Present in Different Enzyme Active Site or its Embedded Microenvironment, Reported from Buried-Hydrophilic Cluster Protruding Toward Protein Surface (Buried-Fraction < 0.78 and rHpy > 0.40)

PDB ID	cystine (chain)	2° structures	Sequence conservations <sup>a</sup>	Strain energy (KJ/mol)	Active residues in 4.5 Å region in half-cystine	BF	rHpy
<b>Cystines in active site in hydrolases</b>							
3EDH	65(A)–64(A)	Coil-bend	High-high	13	C64	0.438	0.602
1K7C	232(A)–214(A)	COIL-helix <sup>b</sup>	Low-high	6.3	C232	0.471	0.575
2FHF	643(A)–644(A)	bend-bend	High-medium	9.8	C643,T642,H607	0.512	0.547
3B8Z	376(A)–371(A)	Turn-beta <sup>c</sup>	Medium-low	13.4	C376,L370	0.528	0.567
4XOJ	124(A)–225(A)	Coil-helix	Medium-medium	10.9	C124,A125	0.531	0.538
1QNR	284(A)–334(A)	Helix-helix	Medium-low	5.5	C284	0.534	0.512
3B8Z	376(B)–371(B)	Turn-beta	Medium-low	13.2	C376,L370	0.545	0.553
3EQA	473(A)–246(A)	Coil-helix	High-low	15.4	C473	0.572	0.534
3HHI	215(B)–154(B)	Coil-helix	High-high	6.3	C154	0.586	0.521
3KUV	73(A)–73(B)	Beta-beta	Low-low	14.5	C73,v74	0.596	0.547
3EDH	64(A)–65(A)	Bend-coil	High-high	13	C64	0.598	0.415
3TBJ	45(A)–28(A)	Turn-coil	High-high	8.9	C45	0.602	0.546
4HWX	31(A)–46(A)	Beta-helix	High-high	7.6	C31,A32	0.605	0.474
1LNI	96(B)–7(B)	Coil-beta	High-low	10.7	C96	0.622	0.505
2WBF	636(X)–593(X)	Turn-helix	High-high	5.3	C636,R635	0.638	0.477
1LNI	7(A)–96(A)	Beta-coil	Low-high	12	C7	0.639	0.515
3ARX	121(A)–116(A)	Beta-coil	High-high	15.5	C121	0.645	0.424
1G6X	14(A)–38(A)	Coil-bend	High-high	12.7	C14	0.649	0.42
3WQB	301(A)–326(A)	Turn-coil	High-high	8.7	C326	0.678	0.526
3TBJ	44(A)–149(A)	Turn-helix	High-high	4	C45	0.71	0.44
4D04	127(B)–158(B)	Turn-turn	High-high	14.8	C127,D80	0.718	0.476
4CPY	336(B)–317(B)	Turn-coil	High-high	12.3	C336	0.721	0.406
4D04	127(A)–158(A)	Turn-turn	High-high	14.5	C127,D79	0.724	0.47
3A21	604(B)–585(B)	Coil-beta	Medium-low	14.7	C604,E577	0.74	0.468
3ZFP	69(A)–168(A)	Turn-coil	High-high	15.1	C69	0.74	0.509
3A21	604(A)–585(A)	Coil-beta	Medium-low	14.9	C604,E577	0.749	0.46
4Y5L	255(A)–290(A)	Turn-beta	High-high	14	C255	0.759	0.406
<b>Cystines in the active site of Oxidoreductases</b>							
1ZK7	465(A)–464(A)	Bend-bend	Medium-medium	23.4	C465,A466,C464	0.484	0.534
1ZK7	464(A)–465(A)	Turn-coil	Medium-medium	23.4	C464,	0.503	0.566
4OZ7	4(B)–10(B)	Coil-coil	NA <sup>d</sup> -NA <sup>d</sup>	16.4	C4,C10,S5	0.528	0.568
4OZ7	10(B)–4(B)	Coil-coil	NA <sup>d</sup> -NA <sup>d</sup>	16.4	C10,C4,S5,P8	0.583	0.543
5DQY	62(A)–69(A)	Bend-coil	NA <sup>d</sup> -NA <sup>d</sup>	9.3	C62,Q63,S67	0.586	0.49
4OZ7	10(A)–4(A)	Coil-coil	NA <sup>d</sup> -NA <sup>d</sup>	17.3	C10	0.595	0.534
4OZ7	4(A)–10(A)	Coil-coil	NA <sup>d</sup> -NA <sup>d</sup>	17.3	C4	0.597	0.476
4NTC	148(A)–145(A)	Turn-coil	NA <sup>d</sup> -NA <sup>d</sup>	14.8	C148,H144	0.643	0.435
2Q0L	136(A)–133(A)	Helix-coil	High-medium	- <sup>e</sup>	C136,Q289	0.649	0.51
2Q0L	133(A)–136(A)	Coil-helix	Medium-high	- <sup>e</sup>	C133,T132,C136, K288,Q289	0.658	0.537
<b>Cystines in the active site of transferases</b>							
4WMA	385(A)–356(A)	Turn-coil	NA <sup>d</sup> -NA <sup>d</sup>	20.4	C385,W358	0.587	0.473
2APC	115(A)–145(A)	Bend-turn	Medium-medium	8.6	C115,A114	0.622	0.419
<b>Cystines in the microenvironment of active site residues of hydrolases</b>							
2VB1	6(A)–127(A)	Helix-turn	High-high	5.1	E7,I124	0.367	0.719
3RLG	53(A)–201(A)	Turn-turn	High-high	14.4	N200	0.447	0.652
3LUM	269(C)–297(C)	Coil-helix	High-high	14.1	P249	0.557	0.476
3KUV	73(B)–73(A)	Beta-beta	Low-low	14.5	C73,V74	0.567	0.539
3C1U	96(A)–88(A)	Beta-helix	Medium-high	9.8	Y97	0.588	0.444
3NKQ	801(A)–413(A)	Turn-bend	High-high	3.6	S800	0.617	0.462
1KNM	119(A)–100(A)	COIL-Beta	High-high	13.2	Y117	0.62	0.489
1EB6	117(A)–177(A)	Turn-Coil	Medium-high	30	H118	0.624	0.409
1G66	2(A)–79(A)	Coil-Bend	High-high	6	S1,S74	0.631	0.554
2NLR	69(A)–64(A)	Beta- Beta	Medium-medium	19.8	H69	0.635	0.467
2D1Z	425(A)–406(A)	Coil-beta	Medium-medium	13.2	Y423	0.645	0.437
3TRS	101(D)–18(D)	Coil-bend	High-high	13.9	G15	0.646	0.458
2WJ9	13(B)–125(A)	Bend-coil	Low-medium	7.3	C129	0.647	0.524
3EQN	77(A)–73(A)	Coil-bend	Medium-high	13.4	D78	0.65	0.491
2XXL	197(A)–188(A)	Beta-beta	High-low	18.2	D187	0.671	0.476
3LUM	250(A)–277(A)	Turn-turn	High-high	2.9	P249	0.673	0.439
3LUM	250(C)–277(C)	Turn-turn	High-high	3	P249	0.675	0.437
3LUM	250(B)–277(B)	Turn-turn	High-high	3.2	P249	0.676	0.437

**Table VI**  
(Continued)

PDB ID	cystine (chain)	2° structures	Sequence conservations <sup>a</sup>	Strain energy (KJ/mol)	Active residues in 4.5 Å region in half-cystine	BF	rHpy
2WJ9	13(A)–125(B)	Coil-coil	Low-medium	7.4	C129	0.677	0.456
3LUM	250(D)–277(D)	Turn-turn	High-high	3.7	P249	0.679	0.434
3LZT	6(A)–127(A)	Helix-coil	High-high	5.78	R128	0.68	0.441
4M1U	56(F)–3(F)	Coil-beta	Medium-medium	7.6	S54	0.68	0.526
2XXL	197(B)–188(B)	Beta-bend	High-low	18.2	D187	0.69	0.499
1LBU	81(A)–3(A)	Bend-bend	Medium-medium	11.1	D80	0.691	0.43
3A21	647(B)–628(B)	Coil-beta	Low-medium	17.1	G626,W645	0.693	0.432
4HYQ	53(A)–28(A)	Coil-helix	High-high	10.8	S54	0.693	0.467
1I71	1(A)–78(A)	Beta-Coil	High-high	7.5	P79	0.699	0.44
4H04	589(B)–564(B)	Coil-beta	High-high	11.4	N588	0.703	0.557
4H04	589(A)–564(A)	Coil-beta	High-high	11.5	N588	0.705	0.489
3TRS	101(B)–18(B)	Coil-coil	High-high	15.6	G15	0.71	0.455
4EMN	291(C)–355(C)	Helix-helix	High-high	8.1	E292,C355	0.711	0.41
2IFR	148(A)–141(A)	Beta-beta	Low-medium	13.9	E139	0.72	0.402
1Y43	101(B)–18(B)	Coil-coil	High-high	11.8	E19	0.735	0.436
2D1Z	842(B)–823(B)	Coil-beta	High-High	12.3	H843,Y840	0.737	0.436
2FHF	644(A)–643(A)	Bend-beta	Medium-high	9.8	T642,H607	0.74	0.401
3B8Z	371(B)–376(B)	Beta-beta	Low-medium	13.2	L370,C376	0.749	0.487
3A21	562(A)–543(A)	Coil-beta	Medium-low	13.3	N563,W560	0.752	0.464
2I9A	50(D)–131(D)	Bend-bend	High-high	8.1	H129	0.753	0.436
2XU3	65(A)–22(A)	Turn-helix	High-high	7.7	G23	0.757	0.403
4HYQ	28(A)–53(A)	Helix-coil	High-high	10.8	S54	0.766	0.421
3B8Z	371(A)–376(A)	Beta-turn	Low-medium	13.4	L370,C376	0.785	0.4
Cystine in microenvironment of active site residues in Transferases							
4WTP	264(A)–218(A)	Coil-helix	Medium-high	12.7	K221,A262	0.642	0.402
4EBY	93(A)–25(A)	Beta-coil	High-high	8.3	F98	0.64	0.461
Cystine in microenvironment of active site residues in Lyases							
4Q8K	365(A)–377(A)	Coil-helix	High-high	9.1	K364	0.636	0.47
3KBR	131(A)–210(A)	Helix-beta	High-medium	7.7	H213,P214,N215	0.642	0.542
Cystine in microenvironment of active site residues in Electron transport proteins							
2Z8Q	48(A)–21(A)	Helix-high	Low-coil	11.3	L20	0.601	0.543
Cystine in microenvironment of active site residues in Isomerases							
3O22	89(A)–186(A)	Beta-bend	High-high	6.1	F83	0.257	0.75

Secondary structures and sequence conservation of each half-cystines were reported along with the strain energy of the cystine disulphide bonds. PDB IDs are arranged according to the increasing values of buried fraction (second last column). Amino acids and their positions involved in enzyme active sites are mentioned in third last column.

<sup>a</sup>Sequence conservation obtained from PdbSum.<sup>38</sup>

<sup>b</sup>Includes 3(10)-helix, alpha-helix and  $\pi$ -helix secondary structures calculated by DSSP.<sup>27,28</sup>

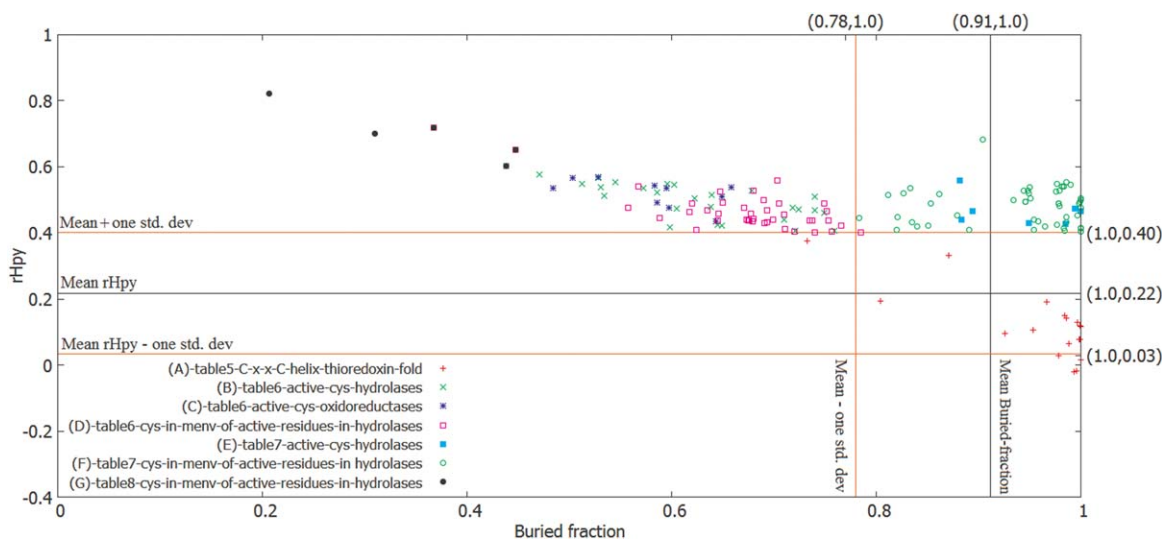
<sup>c</sup>Includes beta sheet, beta bridge and beta strand secondary structures calculated by DSSP program.<sup>27,28</sup>

<sup>d</sup>Sequence conservation was not reported by "The ConSurf Server."<sup>36</sup>

<sup>e</sup>Strain energy cannot be calculated by the "Disulfide Bond Dihedral Angle Energy Server."<sup>35</sup>

observed enzyme class in this cluster, as this enzyme class constitutes the maximum population size. Conversely, oxidoreductase and electron transport proteins were exclusively present in buried clusters, either hydrophobic or hydrophilic (0 half-cystines were observed in exposed-hydrophilic cluster from oxidoreductase and electron-transport proteins). In the following part of this section, we have attempted to answer the question why oxidoreductase and electron transport proteins are selective toward buried clusters only. Out of the 675 half-cystines in oxidoreductase and electron-transport proteins, 209 belong to alpha helical secondary structure (Table IV). Twenty two cystines in alpha helical conformations were part of redox-active C-x-x-C- motif (Table V). 18 out of 22 cystines, belongs to thioredoxin fold. All these 22 cystines have higher strain energies compared to remaining

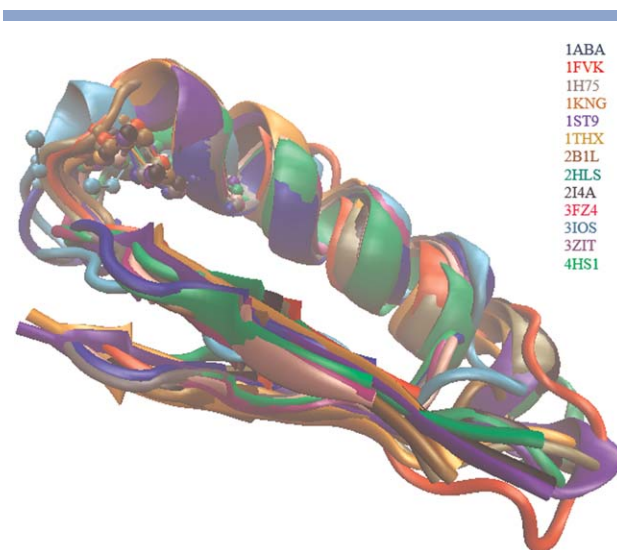
disulphides in buried-hydrophobic cluster (average value of strain energy in buried-hydrophobic cluster is 11.13 KJ/mol). The above redox active half-cystines with alpha helical conformations in thioredoxin fold span a small region within buried-hydrophobic cluster, with few exceptions (Table VI). One prominent exception is, Cys15 (chain A) of a thioredoxin like mutant protein (PDB ID:3ZIT) that significantly differ from other thioredoxin proteins due to the presence of a Thr53 residue (not conserved) within the microenvironment of -C-x-x-C-motif.<sup>37</sup> The second Cys from -C-x-x-C- motif belongs to beta turn secondary structure for all the above thioredoxin folds (as obtained from PDBSUM<sup>38</sup>) (Table V). The active half-cystines, part of four-helical bundle folds, possess the second Cys in gamma turn secondary structure. It has been proposed earlier that redox active



**Figure 3**

Microenvironments of half-cystines depicted in buried fraction and rHpy space—(a) (red diamond) as a part of -C-x-x-C- motif in thioredoxin fold with alpha helical geometry (values given in Table V); (b) (green cross) as a part of the active sites in hydrolase enzyme class present in buried hydrophilic cluster protruding toward protein surface (values given in Table VI); (c) (blue star) as a part of the active sites in oxidoreductase enzyme class present in buried hydrophilic cluster protruding toward protein surface (values given in Table VI); (d) (magenta square) as a part of the microenvironment around the active sites in hydrolase enzyme class present in buried hydrophilic cluster protruding toward protein surface (values given in Table VI); (e) (cyan filled square) as a part of the active sites in hydrolase enzyme class present in buried hydrophilic cluster protruding toward protein interior (values given in Table VII); (f) (green circle) as a part of microenvironment surrounding the active sites in hydrolase enzyme class present in buried hydrophilic cluster protruding toward protein interior (values given in Table VII); (g) (black filled circle) as a part of the active sites in hydrolase enzyme class present in exposed hydrophilic cluster (values given in Table VIII). Mean values of buried fraction and rHpy are shown by black lines. One standard deviation values with respect to buried fraction and rHpy are shown in orange. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

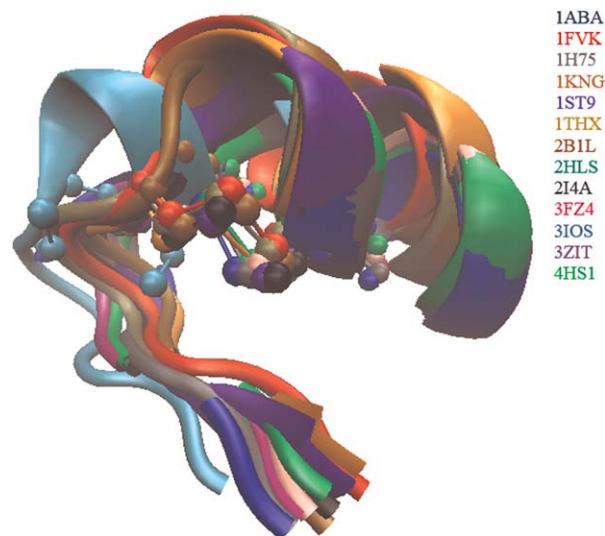
cystines in oxidoreductases prefer alpha helical conformation due to their easy cleavage.<sup>39</sup> To analyze the residues involved in -C-x-x-C- motif of thioredoxin fold, sequence and structural alignment of these proteins (listed in Table V) were performed. Global sequence alignment (using T-COFFEE<sup>30</sup>) revealed 37% of sequence identity among all these 13 sequences. The local sequence alignment shows the conservation of -C-x-x-C- motif in all these proteins, with an exception of one protein, PDB ID: 1FVK (Supporting Information Fig. S2). However, structural alignment of the above full proteins resulted into C-alpha RMSD value of 4.4 Å in the structurally aligned regions (as obtained from SALIGN<sup>32</sup>) It was reported earlier that the redox active cystines, part of -C-x-x-C- motif in thioredoxin folds, were mainly observed in beta-alpha-beta fold region.<sup>15</sup> Superimposition of beta-alpha-beta regions from above 13 proteins shows Cα RMSD of 1.99 Å (Fig. 3). The Cα RMSD of the superimposed cystine residues in those proteins is 1.4 Å. The Cα RMSD of the microenvironment around the cystines in -C-x-x-C- motif of thioredoxin fold is 1.7 Å. The novel finding in this work is that these half-cystines at the active site share near-identical microenvironment as a consequence of having a common fold (Fig. 4).



**Figure 4**

Superimposed beta-alpha-beta region from 13 proteins with thioredoxin fold. Half-cystine from different proteins in this aligned regions are depicted by different colors. Disulphide bridges are shown by spheres connected by sticks. Beta-alpha-beta regions are represented through cartoons. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]





**Figure 5**

Conservation of microenvironment around disulphide bridge (-S-S-) in -C-x-x-C- motif from 13 proteins in alpha-helical conformations of thioredoxin folds. Microenvironment in each case have alpha-helical conformation on one side and beta turn on the other side. -C-x-x-C motif in all these proteins are encompassed within the cleft created by alpha helix and beta turn region. Disulphide bond is represented as spheres connected by sticks. The cartoon diagram represents alpha helix and beta turn. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

#### Buried-hydrophilic cluster hosts half-cystines from enzyme active sites or from surrounding microenvironments of the catalytic (or active) sites

Buried-hydrophilic microenvironment cluster contradicts with the intrinsic hydrophobic character of disulphide-bridged cystine residues. The average buried fraction and rHpy values of cystines in the total dataset (total 10,168 half-cystines) are 0.91 and 0.22 respectively. Respective one standard deviation values for buried fraction and rHpy are 0.13 and 0.18 (Fig. 5). Cystines outside one standard deviation values of buried fraction and rHpy were defined to have mismatched microenvironment.<sup>9</sup> Cystines from buried hydrophilic cluster, encountering mismatched microenvironment (total 963 half-cystines), are mostly exposed toward the surface of the protein (buried fraction  $<0.78$  and rHpy  $>0.40$ ). Out of these, 445 belongs to enzymes; mostly from hydrolase enzyme class (332 half-cystines, constituting 74% of the enzyme class in buried hydrophilic cluster). Total 39 half-cystines were detected as part of the enzyme active sites and 47 in the microenvironments around enzyme active sites (Table VI). Two-third of the active half-cystines in hydrolases have high sequence conservation according to ConSurf Server.<sup>40</sup> There were only three active half-cystines in hydrolases with low sequence conservation. No active half-cystines in oxidoreductases

exhibited low sequence conservation (Table VI). In terms of the secondary structures of active half-cystines, those from oxidoreductases mainly adopted coil or turn conformations and exposed toward the protein surface (Fig. 5). Hydrolases, being the largest class of enzyme in buried-hydrophilic cluster, secondary structures of half-cystine pairs (from a disulphide) have wider variations in comparison to those in oxidoreductases. However, active half-cystines from hydrolases cluster together into smaller sub-groups (Fig. 5). The one toward the protein interior is more densely packed. All the half-cystines from this sub-group share similar secondary structure with their counter-parts in the disulphide (Table VI). Remaining half-cystines from hydrolases are more scattered (toward exposed region) in buried-fraction—rHpy space (Fig. 5), that can presumably be attributed to two different secondary structures of half-cystines in a disulphide (Table VI).

Apart from the half-cystines protruding toward protein surface (BF  $<0.78$  and rHpy  $>0.4$ ) in buried hydrophilic cluster, we have also explored the functionalities of half-cystines embedded in hydrophilic microenvironment protruding toward protein interior (buried fraction  $>0.78$  and rHpy  $>0.40$ ). These cystines also exhibited catalytic activities but to lesser extent compared to the cystines those discussed in the above paragraph. Total 652 such half-cystines were obtained. Out these 337 belong to enzyme classes. Only seven half-cystine were part of active site and 63 were reported as part of the microenvironment around the active sites (Table VII). One compact sub-group of half-cystines from microenvironment of the active site was observed (buried fraction  $\sim 1$ ) (Fig. 5). Within this sub-group, almost all the half-cystines share the same secondary structures with its counterpart in the disulphide (Table VII). This observation is similar to our observation in the previous paragraph for densely packed sub-group (Table VI and Fig. 5). This analysis has shown that the cystines in the microenvironment of catalytic residues, on average, have high strain energies compared to the average strain energy in this microenvironment cluster (11.0 KJ/mol). Despite of low or medium sequence conservation in some of the half-cystines, all these half-cystines share very similar microenvironment (Fig. 5). One of the reasons for low sequence identity of these cystine residues is their occurrences in irregular secondary structures (coil, turn etc.) in contrast to cystines in -C-x-x-C- motif of thioredoxin fold (fully conserved). It has been shown in Table V that the cystines in -C-x-x-C- motifs from oxidoreductase or electron transport proteins mainly occupy alpha-helical structure to facilitate easy cleavage of the disulfide bond, with high strain energy values. These observations together suggest that cystines within the microenvironment of catalytic residues of hydrolase enzyme do not require easy cleavage of the disulfide bond as in thioredoxin folds.

**Table VII**

Half-Cystines (First Out of the Pair of Half-Cystines From a Disulphide Shown in the Table) Present in Different Enzyme Active Site or its Embedded Microenvironment, Reported from Buried-Hydrophilic Cluster Protruding Toward Protein Interior (Buried-Fraction > 0.78 and rHpy > 0.40)

PDB ID	Cystine (chain)	2° structures	Sequence conservations <sup>a</sup>	Strain energy (KJ/mol)	Active residues in 4.5 Å region in half-cystine	BF	rHpy
Active cystines in hydrolases							
4ZA3	206(A)–189(A)	Coil-beta <sup>b</sup>	High-high	8.4	C206	0.882	0.559
2QTW	358(B)–323(B)	Turn-helix <sup>c</sup>	High-high	11.4	C358	0.883	0.439
2JJB	539(C)–533(C)	coil-bend	Low-medium	19.4 C539, P534, A544, C541	0.894	0.465	
1G66	52(A)–46(A)	Helix-coil	Low-medium	17.1 T13, G47, Q49, S50, Q91	0.949	0.43	
1G66	179(A)–147	Bend-coil	Low-low	13.6 D175, T146, D172, A173,	0.985	0.427	
1G66	46(A)–52(A)	Coil-Helix	Medium-low	17.1 T13, G47, Q49, S50, Q91, S174, Y177	1	0.464	
1GPI	225(A)–245(A)	Beta-Beta	High-high	16.3	G207	0.994	0.472
Cystines in microenvironment of active site residues in Hydrolases							
2D1Z	342(A)–323(A)	Coil-beta	Medium-low	11.4	Y340, H343	0.783	0.445
4YEO	127(A)–6(A)	Coil-helix	High-high	3.8	R125	0.812	0.514
3WMT	297(B)–308(B)	Coil-coil	High-high	8.9	L296	0.82	0.408
3E2V	303(B)–331(B)	Helix-coil	High-high	13	E332	0.821	0.447
3E2V	303(A)–331(A)	Helix-coil	High-high	12	E332	0.827	0.518
1XKG	183(A)–145(A)	Coil-helix	High-high	8.5	R184	0.833	0.536
3WMT	297(A)–308(A)	Coil-bend	High-high	8.2	L296	0.835	0.432
2X5X	36(A)–85(A)	Bend-turn	Medium-medium	12	S35	0.84	0.419
300D	43(A)–47(A)	Coil-helix	High-high	16.8	N42	0.851	0.421
2QTW	323(B)–358(B)	Helix-turn	High-high	11.4	C358	0.853	0.489
2XXL	349(B)–314(A)	Turn-helix	High-high	14.7	T348	0.862	0.517
5AR6	64(A)–71(A)	Coil-beta	High-high	17	N70	0.879	0.451
3VPL	159(A)–189(A)	Turn-helix	High-medium	15.7	Y158	0.891	0.409
2JJB	539(B)–539(B)	Coil-coil	Medium-low	24.6	P534, A544, C540	0.904	0.681
1ROR	16(I)–35(I)	Helix-helix	High-high	7.7	F37	0.934	0.498
3WMT	308(A)–297(A)	Bend-coil	High-high	8.2	L296	0.944	0.526
3WMT	308(B)–297(B)	Coil-coil	High-high	8.9	L296	0.946	0.494
3WMT	308(B)–297(B)	Coil-coil	High-high	8.9	L296	0.946	0.494
1KNM	76(A)–59(A)	Coil-Beta	High-high	14.9	W77	0.948	0.527
3B7E	318(B)–336(B)	Coil-bend	High-high	23.4	D387	0.949	0.519
3B7E	318(A)–336(A)	Coil-bend	High-high	22.3	D387	0.95	0.538
2XXL	314(B)–349(A)	Helix-turn	High-high	14.7	T348, V347	0.951	0.505
3MWWQ	197(A)–204(A)	Coil-beta	NA <sup>d</sup> -NA <sup>d</sup>	14.8	D253	0.954	0.439
3TRS	18(B)–101(B)	Coil-coil	High-high	15.6	G15	0.954	0.408
2X5X	85(A)–36(A)	Turn-bend	Medium-medium	12	D84	0.958	0.434
3LUM	277(D)–250(A)	Turn-turn	High-high	2.9	Q262, A263	0.965	0.419
2UWA	218(B)–226(B)	Beta-coil	High-low	9.8	T219	0.976	0.524
2D1Z	382(A)–365(A)	Coil-beta	Medium-medium	14.1	E357, R359	0.977	0.548
1LLF	60(A)–97(A)	Coil-coil	High-high	4.5	R361	0.977	0.423
3NKQ	73(A)–86(A)	Coil-coil	High-medium	7.3	N53	0.978	0.487
3NKQ	73(A)–86(A)	Coil-coil	High-medium	7.3	N53	0.978	0.487
1KLI	91(L)–102(L)	Turn-beta	High-high	16.7	E94, S89	0.979	0.481
2UWA	218(C)–226(C)	Beta-coil	High-low	7.2	T219	0.979	0.526
2UWA	218(A)–226(A)	Beta-coil	High-low	8.0	T219	0.982	0.539
2JJB	533(A)–539(A)	Bend-coil	Medium-low	21.6	P534, A544, C539	0.983	0.54
2JJB	533(D)–539(D)	Bend-coil	Medium-low	23.6	P534, A544, C541	0.983	0.54
3B8Z	342(A)–394(A)	Helix-beta	High-high	23.3	C394	0.983	0.413
2ODP	655(A)–685(A)	Coil-turn	High-high	7.5	R694	0.984	0.407
4D04	49(A)–42(A)	Bend-turn	High-high	10.6	C42, I40	0.986	0.553
3PMS	231(A)–252(A)	Helix-coil	High-high	6.4	P246	0.986	0.447
4O36	65(B)–72(B)	Coil-beta	High-high	14.6	Q69, N71, C72	0.986	0.432
4D04	49(B)–42(B)	Bend-turn	High-high	11.4	C42, I40	0.99	0.544
3PMS	208(A)–204(A)	Bend-beta	High-high	15	Y161	0.997	0.452
2GMN	181(B)–201(B)	Coil-coil	High-high	29.7	H101, H177	0.998	0.489
2GMN	181(A)–201(A)	Coil-coil	High-high	29.8	H101, H177	0.999	0.493
3WVC	111(A)–99(A)	Turn-helix	High-high	14.2	S66	1	0.498
4UZ1	413(A)–432(A)	Helix-beta	High-high	16.1	R409, H412	1	0.503
3EQN	698(A)–692(A)	Bend-helix	High-high	7.5	F684	1	0.414
3EQN	692(B)–692(B)	Helix-helix	High-high	7.9	F684	1	0.414
4BDX	33(A)–5(A)	Beta-beta	NA <sup>d</sup> -NA <sup>d</sup>	6.2	L4	1	0.476
3LZT	127(A)–6(A)	Coil-helix	high-High	5.8	R128	1	0.402
Cystine in the microenvironment of Ligase							

**Table VII**  
(Continued)

PDB ID	Cystine (chain)	2° structures	Sequence conservations <sup>a</sup>	Strain energy (KJ/mol)	Active residues in 4.5 Å region in half-cystine	BF	rHpy
2PHN	248(B)–244(B)	Helix-helix	High-medium	12.8	C244	0.889	0.501
2PHN	248(A)–244(B)	Helix-helix	High-medium	12.8	C244	0.895	0.474
Cystine in the microenvironment of Llyases							
1Y7W	221(A)–31(A)	Coil-helix	High-high	11.1	L216	1	0.566
1Y7W	221(B)–31(B)	Coil-helix	High-high	11.4	L216	1	0.596
Cystine in the microenvironment of Transferases							
5AJ0	473(A)–456(A)	Coil-beta	NA <sup>d</sup> -NA <sup>d</sup>	12.4	Q451	0.804	0.43
4WMA	356(A)–385(A)	Coil-turn	NA <sup>d</sup> -NA <sup>d</sup>	20.3	N384	0.955	0.444
5FOE	414(A)–407(A)	Coil-coil	High-high	9.7	T1045, I1022	0.996	0.525
Cystine in microenvironment of Oxidoreductases							
3NT1	36(B)–47(B)	Turn-beta	Low-medium	10.9	Y55	0.828	0.418

Secondary structures and sequence conservation of each half-cystines were reported along with the strain energy of the cystine disulphide bonds. PDB IDs are arranged according to the increasing values of buried fraction (second last column). Amino acids and their positions involved in enzyme active sites are mentioned in third last column.

<sup>a</sup>Sequence conservation obtained from PdbSum.<sup>38</sup>

<sup>b</sup>Includes beta sheet, beta bridge and beta strand secondary structures calculated by DSSP program.<sup>27,28</sup>

<sup>c</sup>Includes 3(10)-helix, alpha-helix and  $\pi$ -helix secondary structures calculated by DSSP program.<sup>27,28</sup>

<sup>d</sup>Sequence conservation is not available by ConSurf server.<sup>36</sup>

### Exposed-hydrophilic microenvironment cluster hosts half-cystines as a part of enzyme active sites or catalytic sites

Physico-chemical properties (in terms of buried fraction and rHpy values) of this microenvironment cluster are most dissimilar with respect to the embedded cystine residue. Average buried fraction and rHpy values for this cluster are 0.328 and 0.720 respectively (Table I), in contrast to 0.91 and 0.22 for current cystine microenvironment dataset. As this microenvironment cluster is extremely mismatched to the overall hydrophobicity of cystine, very less data points were recorded in this cluster, compared to other two clusters (Table I). There were only 31 half-cystines recorded as part of enzymes, 28 out of those were hydrolases (Table III). Out of these 28, only 5 half cystines were detected as a part of the microenvironment of active site in hydrolases (Table VIII). No half-cystines were reported as part of the active sites in

hydrolases, in contrary to those in buried hydrophilic cluster. This presumably indicates that half-cystines occupying the exposed surface of the protein rarely participate in direct enzymatic reactions, however, those are engaged in protecting (via microenvironment) the enzyme active sites. All these half-cystines are highly scattered in the buried fraction, rHpy space (Fig. 5 and Table VIII).

## CONCLUSION

It was already known that similar folds lead to similar functions (enzyme activities) in proteins.<sup>41–43</sup> In this study we have shown that cystines from different enzymes, evolved within similar microenvironments, when present at the active site of same enzyme class. Our underlying aim was to test the hypothesis that cystines with similar functions (enzyme activities) from

**Table VIII**

Half-Cystines (First Out of the Pair of Half-Cystines from a Disulphide Shown in the Table) Present in Embedded Microenvironment of Active Sites in Hydrolase Enzyme Class, Reported from Exposed-Hydrophilic Cluster

PDB ID	cystine (chain)	2° structures	Sequence conservations <sup>a</sup>	Strain energy (KJ/mol)	Active residues in 4.5 Å region in half-cystine	BF	rHpy
1ZGX	96(B)–7(B)	Coil-coil	High-low	11.4	K94	0.206	0.821
2G58	96(A)–84(A)	Helix <sup>b</sup> -helix	High-high	10.8	K74	0.31	0.699
2VB1	6(A)–127(A)	Helix-helix	High-high	5.1	E7, I124	0.367	0.719
3EDH	65(A)–64(A)	Coil-coil	High-high	13	C64	0.438	0.602
3RLG	53(A)–201(A)	Turn-turn	High-high	14.4	N200	0.447	0.652

Secondary structures and sequence conservation of each half-cystines were reported along with the strain energy of the cystine disulphide bond. PDB IDs are arranged according to the increasing values of buried fraction (second last column). Amino acids and their positions involved in enzyme active sites are mentioned in third last column.

<sup>a</sup>Sequence conservation obtained from PdbSum.<sup>38</sup>

<sup>b</sup>Includes 3(10)-helix, alpha-helix and  $\pi$ -helix secondary structures calculated by DSSP program.<sup>27,28</sup>

different proteins will belong to similar microenvironment clusters. To this end, by using hierarchical clustering method, we have identified three different microenvironment clusters: I) buried-hydrophobic, ii) buried-hydrophilic and exposed-hydrophilic, and correlated these with their enzymatic functions. In our analysis, special emphasis has been given to cystines from enzyme classes. As demonstrated here, cystines from -C-x-x-C- motifs in Oxidoreductase enzyme class all have very similar microenvironment, that is, buried and hydrophobic. The catalytic cystines from the hydrolase enzyme class always prefer partly exposed hydrophilic microenvironment (that is buried hydrophilic cluster). Other cystines from hydrolase enzymes which participate in stabilizing catalytic or active sites were detected in both buried hydrophilic and exposed hydrophilic microenvironment clusters. Despite of low to medium sequence conservation in some of the cystines, active cystines from hydrolase enzymes share similar microenvironments. This report illustrates that cystine residues share similar microenvironments at active sites of specific enzyme classes despite of variation in their sequence similarities and sequence position conservations, thus validating the working hypothesis. We believe this conclusion should be further verified for other amino acids, particularly, titratable amino acids, like Aspartic acid, Glutamic acid, Arginine and so forth. Titratable amino acids are expected to be more sensitive toward change in microenvironment due to alteration in their protonation states. Hence, microenvironment modulated switching of protonation states in titratable amino acids would be of interest in major biochemical reactions, like photosynthesis. For example, carbon dioxide fixing enzyme, Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCo), involves seven charged residues in its active site. It is known that switching some of these protonation states could lower the transition state, thus assist to prevent backward reaction and trap more carbon dioxide.<sup>44</sup> This can be further verified by modifying the protein structures with altered protonation states of charged residues and test the capacity of carbon dioxide fixation.

Amino acids in heterogeneous protein microenvironments are analogous to amino acids in different solvents with variable dielectric media. As solubility, bond dissociation energy and spectral properties of cystine vary from hydrophilic to hydrophobic solvents, the same is also expected when cystine is transferred from hydrophilic part of the protein microenvironment to its hydrophobic part. This could be exploited to guide experimentation in to the local dielectric medium within protein microenvironments. We have shown here that disulfide-bridged cystine molecule is a good model system to examine the effect of various dielectric medium on S-S bond dissociation energy and can be extended for experimental verifications.

## ACKNOWLEDGMENTS

Authors sincerely acknowledge Dr. Marcin Apostol, ADRx. Inc. Thousand Oaks, California, USA, for kindly providing the program for dihedral strain energy calculations.

## REFERENCES

1. Rekker RF. The hydrophobic fragmental constant. Amsterdam: Elsevier; 1977.
2. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. *Nature* 1986;319:199–203.
3. Wimley WC, Creamer TP, White SH. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* 1996;35:5109–5124.
4. Mehler EL, Guarnieri F. A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophys J* 1999;77:3–22.
5. MacCallum JL, Bennett WFD, Tieleman DP. Partitioning of amino acid side chains into lipid bilayers: results from computer simulations and comparison to experiment. *J Gen Physiol* 2007;129:371–377.
6. McIntosh TJ, Simon SA. Bilayers as protein solvents: role of bilayer structure and elastic properties. *J Gen Physiol* 2007;130:225–227.
7. White SH. Membrane protein insertion: the biology-physics nexus. *J Gen Physiol* 2007;129:363–369.
8. Wolfenden R. Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins. *J Gen Physiol* 2007;129:357–362.
9. Bandyopadhyay D, Mehler EL. Quantitative expression of protein heterogeneity: response of amino acid side chains to their local environment. *Proteins* 2008;72:646–659.
10. Jiang Y, Ruta V, Chen J, Lee A, Mackinnon R. The principle of gating charge movement in a voltage-dependent K<sup>+</sup> channel. *Nature* 2003;423:42–48.
11. Moukhametzianov R, Klare JP, Efremov R, Baeken C, Göppner A, Labahn J, Engelhard M, Büldt G, Gordeliy VI. Development of the signal in sensory rhodopsin and its transfer to the cognate transducer. *Nature* 2006;440:115–119.
12. Harris TK, Turner GJ. Structural basis of perturbed pK a values of catalytic groups in enzyme active sites. *IUBMB Life* 2002;53:85–98.
13. Ray S, Bhattacharyya M, Chakrabarti A. Conformational study of spectrin in presence of submolar concentrations of denaturants. *J Fluoresc* 2005;15:61–70.
14. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;28:304–305.
15. Branden C, Tooze J. Introduction to protein structure, 2nd ed. New York: Garland Publishing Taylor and Francis group; 1991.
16. Lundstrom-Ljung J, Holmgren A. Prolyl hydrolase, protein disulfide isomerase, and other structurally related proteins. New York: CRC Press; 1998.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000;28:235–242.
18. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph* 1996;14:33–38, 27–28.
19. Pascual-ahuir JL, Silla E, Tunon I. GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface. *J Comput Chem* 1994;15:1127–1138.
20. Furnham N, Holliday GL, de Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res* 2014;42:D485–D489.
21. Hartigan JA. Clustering algorithms. New York: Wiley; 1975.
22. Tryon RC, Bailey DE. Cluster analysis. New York: McGraw-Hill; 1973.

23. Ward J. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 1963;58:236–244.
24. Addinsoft. <http://www.xlstat.com> XLSTAT 2014, Data analysis and statistics software for Microsoft Excel; Paris, France, 2014.
25. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 2008;36:D419–D425.
26. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr Sect D Biol Crystallogr* 2004;60:2256–2268.
27. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
28. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G. A series of PDB related databases for everyday needs. *Nucleic Acids Res* 2011;39:D411–D419.
29. Schmidt B, Ho L, Hogg PJ. Allosteric disulfide bonds. *Biochemistry* 2006;45:7429–7433.
30. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* 2015;43:W580–W584.
31. Papadopoulos JS, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* 2007;23:1073–1079.
32. Braberg H, Webb BM, Tjioe E, Pieper U, Sali A, Madhusudhan MS. SALIGN: a web server for alignment of multiple protein sequences and structures. *Bioinformatics* 2012;28:2072–2073.
33. Harrison PM, Sternberg MJ. The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J Mol Biol* 1996;264:603–623.
34. Chuang C-C, Chen C-Y, Yang J-M, Lyu P-C, Hwang J-K. Relationship between protein structures and disulfide-bonding patterns. *Proteins* 2003;53:1–5.
35. Katz BA, Kossiakoff A. The crystallographically determined structures of atypical strained disulfides engineered into subtilisin. *J Biol Chem* 1986;261:15480–15485.
36. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 2005;33:W299–W302.
37. Rohr ÅK, Hammerstad M, Andersson KK. Tuning of thioredoxin redox properties by intramolecular hydrogen bonds. *PLoS One* 2013;8:e69411.
38. de Beer TAP, Berka K, Thornton JM, Laskowski RA. PDBsum additions. *Nucleic Acids Res* 2014;42:D292–D296.
39. Simone ADE, Berisio R, Zagari A, Vitagliano L. Limited tendency of  $\alpha$ -helical residues to form disulfide bridges: a structural explanation. *J Pept Sci* 2006;12:740–747.
40. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 2010;38:W529–W533.
41. Chen S, Bahar I. Mining frequent patterns in protein structures: a study of protease families. *Bioinformatics* 2004;20:18.
42. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 2004;14:208–216.
43. Holbourn KP, Acharya KR, Perbal B. The CCN family of proteins: structure-function relationships. *Trends Biochem Sci* 2008;33:461–473.
44. Collings AF, Critchley C. Artificial photosynthesis: from basic biology to industrial application. In: Anthony F. Collings et al., editors. *Weinheim, Germany, Wiley; 2007. pp 280–282.*

# Role of microenvironment in modulating function of Cystine residues in high-resolution protein structures

## *Modulating effect of micro-environment on Cystine function*

Akshay Bhatnagar and Debashree Bandyopadhyay\*

*Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, Hyderabad campus, Hyderabad, A.P, INDIA, Pin 500078*

\*e-mail of the corresponding author: [banerjee\\_debi@yahoo.com](mailto:banerjee_debi@yahoo.com); [banerjee.debi@hyderabad.bits-pilani.ac.in](mailto:banerjee.debi@hyderabad.bits-pilani.ac.in)

**Abstract:** Three dimensional spatial arrangements of hydrophilic and hydrophobic functional groups constitute the micro-environment around an amino acid within a protein structure. This micro-environment strongly influences various structure and functional aspects of that particular amino acid which, in turn, contribute to the overall structure and function of the protein. Here we have exploited the micro-environment database, generated for Cystine disulphide-bridged residues in the context of high-resolution crystal structure, to understand the implicit relationship between micro-environment and the functional aspects of that particular Cystine residue embedded in the micro-environment. Individual residue functions are curated from the large literature database. Following rules are being deduced from this study regarding the Cystine residue function and its implication in protein structure. i) Disulphide linked Cystine residues are mainly involved in structural stabilization via formation of intra and inter chain disulphide bond. ii) Cystine residues participate in protein function only via formation of few structural motifs and iii) by participating in the protein active site. iv) Unlike the hydrophilic residues, Cystine residue is in functional form only when it is deeply buried inside a protein structure.

**Keywords:** *Protein micro-environment, high-resolution crystal structure, buried fraction, rHpy, micro-environment space, oxidized Cystine, reduced Cystine, structural rules, SCOP classes, normalized occupancies, text-mining.*

### I. Introduction

Quantitative characterization of local microenvironment around any amino acid side chain in proteins has an implication on precisely finding the causative effect of the microenvironment on structure-functionally important amino acids (Lakowicz 1983<sup>[1]</sup>; Haque et.al. 2000<sup>[2]</sup>; Ray et.al. 2005<sup>[3]</sup>; Minh and Madhusudhan 2012<sup>[4]</sup>). Microenvironment can be described as three dimensional spatial arrangement of atoms around any given functional group within a given distance cutoff in a biomolecule (Figure 1). The microenvironment could be highly hydrophobic to highly hydrophilic in nature depending upon the protein class, location of that particular functional group, etc., (Mehler and Guarinerri 1999<sup>[5]</sup>). Protein microenvironments are crucial i) in modulating the protonation states of titrable amino acids, like Glutamic acid, ii) conserving the neighborhood around conserved residues in protein families iii)

maintaining structural heterogeneity of protein tertiary structure and iv) controlling molecular interactions between different macro-molecular complexes, such as – protein-protein, protein-DNA, protein-ligand complexes. Despite of crucial role of protein micro-environment, characterization of this property is limited due to experimental constraints. Experimentally, protein microenvironment can only be characterized for certain amino acids whose fluorescence quenching can be measured (Lakowicz 1983<sup>[1]</sup>). Hence, computational approaches are indispensable to explore micro-environment around individual amino acids. Surprisingly, there were only very few computational attempts to quantitatively characterize protein microenvironment, in general (Ponnuswamy 1980 et.al.<sup>[6]</sup>, Eisenberg and McLachlan 1986<sup>[7]</sup>). Our recent development of microenvironment quantitative property descriptor (QPD), based on physico-chemical properties (Rekker 1977<sup>[8]</sup>), allow quantification of microenvironment around any amino acid, anywhere inside the protein (Bandyopadhyay and Mehler 2008<sup>[9]</sup>). This QPD was applied to a non-redundant database of high-resolution crystal structures (Bandyopadhyay and Mehler 2008<sup>[9]</sup>). Thus, all possible microenvironments for individual amino acids in the context of high-resolution protein structures have been explored. Each amino acid embedded in a microenvironment of protein structure supposed to have its unique role in protein structure and function. Recalling the fact of micro-environment conservation around conserved amino acids in protein families (Bandyopadhyay and Mehler 2008<sup>[9]</sup>), here we have hypothesized that embedded amino acids in similar microenvironment will perform similar functions. This hypothesis is being tested throughout the article.

In this study we have attempted to find out microenvironment dependent function of individual amino acid and its relationship with protein classification. The analysis is being carried out for disulphide bridged Cystine residues which are mainly responsible for structural stability in many proteins. The microenvironment space is generated here based on the buried fraction of the amino acid concerned and the QPD generated for that amino acid for a given location. This microenvironment space was found to be clearly divided into eight different clusters. The structure-functional role of disulphide-bridged

Cystines embedded in those eight different micro-environment regions was explored. This study clearly show discrimination of protein SCOP classes towards different disulphide-bridged Cystine micro-environments.

## II. Method

### A. Description of microenvironment property descriptor:

The micro-environment space is described using two parameters, buried fraction of that particular amino acid and the quantitative property descriptor parameter, rHpy (Bandyopadhyay and Mehler 2008<sup>[9]</sup>). This rHpy is a relative estimate of hydrophobicity value for the microenvironment of an amino acid, computed based on physico-chemical properties of neighboring atoms. The relative contributions come from the neighboring atoms either in protein environment or in bulk water environment

$$rHpy = \frac{Thpy}{Hpys} \quad \text{Equation (1)}$$

Where, Thpy is the microenvironment contribution from protein environment and Hpys is the same contribution from bulk water environment.

### B. Database generation and curation:

In our previous work (Bandyopadhyay and Mehler 2008<sup>[9]</sup>), a non-redundant dataset of high-resolution crystal structures were curated from PDB database. Microenvironments for the side chains of all nineteen amino acids are computed over the entire dataset of protein structures. Seven hundred and one pairs of intra-molecular disulphide-bridged Cystine micro-environments were obtained from that analysis. That dataset for disulphide-bridged Cystine microenvironment has been used here to elucidate role of amino acid microenvironment in protein structure and function. The information on the disulphide connectivity was extracted from respective PDB (Berman et. al. 2000<sup>[10]</sup>) header files using PERL program.

### C. Binning of micro-environment space:

For the analysis of microenvironment distribution, the entire space is theoretically divided into small bins. Range of the buried fraction for any amino acid is zero to one. Upper limit of rHpy is one which is the limiting value of aqueous microenvironment. However, there is no lower limit for rHpy. The minimum value of rHpy is indicative of highly hydrophobic environment and it was observed to assume negative fractional values. The microenvironment space is divided into uniform bins with bin size (0.1, 0.1). (To note, both the quantities, rHpy and buried fraction are dimensionless.) Depending on the observation it has been found that many adjacent bins are almost equally populated. This adjacent bins are taken together to redefine new cluster. Hence, the new clusters are of unequal size and redefined based on their positional proximity and normalized frequency values.

### D. Curation of residue functions associated with each bin:

The functions associated with individual Cystine residues were manually curated from literature. Number of different residue functions was counted in each bin to estimate the dependency of residue

function on its surrounding microenvironment. Similarly, the SCOP (Murzin, et. al. 1995<sup>[11]</sup>, Andreeva et. al. 2007<sup>[12]</sup>) classification of proteins corresponding to individual Cystine residues was also noted to explore the relationship between amino acid micro-environment and SCOP classes.

## III. Results and Discussion

In order to explore the relationship between different functionalities of Cystine molecules and their surrounding microenvironments, systematic database analysis and literature mining were performed. Certain relationship between protein structural classes and amino acid functionalities has been deduced in the following subsections.

### A. Distribution patterns of different micro-environments around disulphide bridged Cystines in high-resolution crystal structures:

A wide range of micro-environment distribution has been observed around disulphide-bridged Cystine residues in the context of high-resolution protein crystal structures (Figure 2). It is interesting to note that no Cystine was found in any protein that was completely exposed to the solvent molecule. Moreover, depending on the position and population of different microenvironments, eight different clusters have been identified (Table 1). Cluster number 1 contains the most buried and hydrophobic micro-environment which gradually progress to solvent exposed and hydrophilic micro-environment in higher cluster numbers. Maximum number of Cystine residues was found in cluster number 1 which is deeply buried inside the protein interior with high hydrophobicity values. The populations gradually decrease in solvent exposed hydrophilic clusters (higher cluster numbers). This is due to mismatch of amino acid hydrophobicity character to that of the micro-environment. There are few intervening small clusters (cluster no. 2 and 6) which are separated because of their physical positions in micro-environment space.

### B. Structure-function role of specific Cystine residue in a particular protein structure and the nature of its embedded micro-environment:

Different micro-environment clusters discussed in the previous subsection gives a complex mosaic of the protein heterogeneity where individual Cystine residue is embedded. Due to the heterogeneity in physico-chemical characteristic of the surrounding, the function of the particular amino acid is expected to be different. Here, we have curated structural and functional contribution of each Cystine residue from literature. Those structure-function contributions of Cystine residues are compared across different micro-environment clusters (Table 2). The main structural role of disulphide-bridged Cystine is to form intra-chain disulphide bonds. In case of multi-domain proteins, few inter-chain disulphide bridges are also formed. Apart from intra and inter chain disulphide bonding, Cystine residues also participate in bridging different secondary structural elements in a protein<sup>[13-</sup>

<sup>23]</sup>Functional aspects of Cystine residues are i) formation of structural motifs<sup>[24-26]</sup> and participating in proteins active site<sup>[27-31]</sup>. All the functional Cystine residues are found only in cluster 1, which is completely buried inside the protein interior and are highly hydrophobic in nature (Table 2). The role of solvent-exposed Cystine residues (found in clusters 7 and 8) are only structural. Hence, this analysis gives an important insight regarding the position and hydrophobicity of Cystine residue and its role in respective protein. To further understand the relationship of micro-environment and the preference of different protein classes, we have compared the SCOP classes in the following subsection.

C. Do protein SCOP classes, containing disulphide-bridged Cystine, prefer certain micro-environments over other?

To address the above question, we have compared the number of protein SCOP classes, containing Cystine residue, across different micro-environment clusters (Table 3). Cluster 1 harness maximum number of proteins. Small proteins prefer more hydrophilic micro-environment (cluster numbers 5 to 8) as shown by their relative occupancies in different clusters (Table 3). Involvement of protein classes other than small proteins is relatively small. However, all-alpha protein class shows reduced preference in their normalized occupancies towards hydrophilic micro-environment. Preferences for all-beta protein classes are equivalent throughout the micro-environment space. Notable preference is observed for alpha+beta class protein; higher preferences towards hydrophilic microenvironment (cluster number 7 and 8) over hydrophobic microenvironment. The alpha/beta class

slightly prefer buried environment (cluster 1 and 2) over exposed micro-environment. This study clearly shows discrimination of protein classes towards different micro-environments. Preference of small protein class towards solvent-exposed micro-environments is justified in terms of protein size.

#### IV. Conclusion

This study was attempted to find out the causative effect of micro-environment on structure and function of individual amino acid in the context of protein structure. Some important rules are being deduced from this study. i) Disulphide linked Cystine residues are mainly involved in structural stabilization via formation of intra and inter chain disulphide bond. ii) Cystine residues participate in protein function only via formation of few structural motifs and iii) by participating in the protein active site. iv) Unlike the hydrophilic residues, for example, Glutamic acid, Lysine, Arginine etc., Cystine residue is in functional form only when it is deeply buried inside a protein structure.

These rules are important characterization about Cystine micro-environment and its functional aspects. However, the role of free thiol (-SH) containing Cystine residue is not explored here. Micro-environment analysis of thiol containing Cystine, in combination with the present study, can give the mechanistic details of disulphide bridge formation and possible role of Cystine in protein structure, both in its reduced and oxidized form. Perhaps the future direction of this study can cast light on oxidation and reduction process of Cystine residues found in hard-keratins.



Table 1: Description of different micro-environment clusters, depicting i) the occupancies in each cluster by Cystine residues and ii) minimum and maximum values of buried fraction and rHpy which define the micro-environment space.

Cluster number	Total number of Cystines	Minimum buried fraction	Minimum rHpy	Maximum buried fraction	Maximum rHpy
8	38	0.3	0.2	0.6	0.7
7	69	0.6	0.3	0.6	0.6
6	2	0.7	0.1	0.7	0.1
5	94	0.7	0.2	0.7	0.4
4	8	0.7	0.0	0.8	0.5
3	156	0.8	0.1	0.8	0.4
2	7	0.8	-0.3	0.9	0.5
1	1028	0.9	-0.2	0.9	0.6

Table 2: Number of different functionalities (depicted along the elements of first row) in different microenvironment clusters (depicted along elements of first column) around disulphide bridged Cystine residues in high-resolution protein structure database. Cluster 8 is the most hydrophilic and solvent-exposed cluster which gradually changes to hydrophobic microenvironment in cluster 1.

Cluster number	Total no. of cysteine residues	Structural attribute				Functional attribute	
		Intra chain s-s bond	Inter chain s-s bond	Bridging secondary structures	loop	Motif	active site
9	38	35	2	1	0	0	0
8	69	69	2	1	0	0	0
7	2	1	0	0	0	0	0
6	94	8	0	1	0	0	0
5	8	8	0	1	0	0	0
4	156	141	3	2	3	1	0
3	7	5	1	0	0	0	0
2	1018	303	5	6	4	3	5
1	10	9	0	0	1	0	0

Table 3: Number of different protein SCOP classes in micro-environment clusters. Normalized occupancy values are shown in parentheses. Normalization done with respect to the total number of proteins in each cluster

Cluster no.	Total proteins	Small protein	All alpha protein	All beta protein	Alpha+beta protein	Alpha/beta protein	Coiled coil protein
8	30	7 (0.23)	2 (0.067)	9 (0.3)	6 (0.2)	5 (0.167)	1(0.033)
7	46	14 (0.304)	3 (0.065)	12 (0.261)	14 (0.304)	2 (0.043)	0
6	1	1	0	0	0	0	0
5	30	10 (0.333)	5 (0.166)	7 (0.232)	5 (0.166)	3 (0.1)	0
4	7	1 (0.14)	1 (0.14)	4 (0.57)	1 (0.14)	0	0
3	80	13 (0.225)	10 (0.212)	13 (0.25)	8 (0.175)	9 (0.012)	0
2	5	1 (0.2)	0	1 (0.2)	3 (0.6)	0	0

1	154	27 (0.181)	25 (0.188)	34 (0.268)	25 (0.188)	18 (0.148)	0
---	-----	------------	------------	------------	------------	------------	---

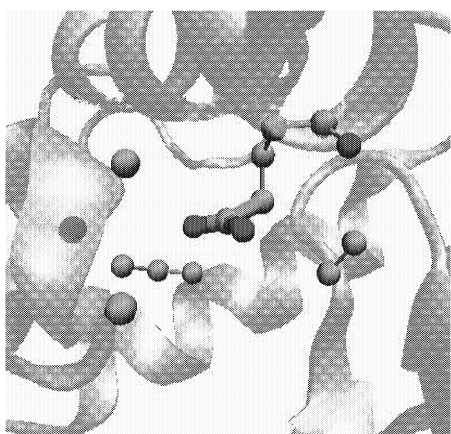


Figure 1: Schematic representation of micro-environment around a carboxylic acid functional group from a Glutamic acid residue. The carboxylic acid functional group is shown as ball and thick line representation. The surrounding micro-environment (three dimensional atomic arrangement), within 4.5 Å distance cutoff, is shown as ball and thin line representation. In this particular example, the Glutamic acid side chain is embedded in hydrophobic micro-environment created mainly by surrounding carbon atoms, represented in cyan. Only one nitrogen atom is being found in this micro-environment.

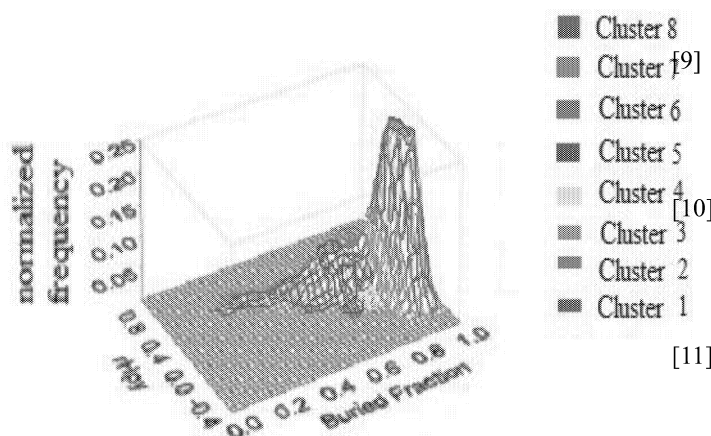


Figure 2: Population distribution of different clusters of micro-environment around disulphide-bridged Cysteine residues, obtained from high-resolution crystal structures. The micro-environment space is two-dimensional, defined by buried fraction and rHpy values. Clusters are depicted by different colors. With increase in cluster number the hydrophilic nature of the micro-environment increases.

### V. Acknowledgment

DB acknowledges funding of this work from Birla Institute of Technology and Science, Pilani, Research Initiation Grant.

### VI. References

- [1] Lakowicz JR. Principles of fluorescence spectroscopy. Plenum 1983.
- [2] Haque ME, Ray S, Chakrabarti A. Polarity estimate of the hydrophobic binding sites in

erythroid spectrin: a study by pyrene fluorescence. *J Fluoresc* 2000.

- [3] Ray S, Bhattacharyya M, Chakrabarti A. Conformational study of spectrin in presence of submolar concentrations of denaturants. *J Fluoresc* 2005.
- [4] Minh NN, Madhusudhan MS. Biological insights from topology independent comparison of protein 3D structures. *Nucl Acid Res* 201; 39(14): e94.
- [5] Mehler EL, Guarnieri F. A self-consistent microenvironment modulated screened coulomb potential approximation to calculate pH dependent electrostatic effects in proteins. *Biophysics J* 1999.
- [6] Ponnuswamy PK, Prabhakaran M, Manavalan P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim Biophys Acta* 1980.
- [7] Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. *Nature* 1986.
- [8] Rekker RF, Nauta WT. The hydrophobic fragmental constant. Elsevier 1977.
- [9] Bandyopadhyay D, Mehler EL. Quantitative expression of protein heterogeneity: Response of amino acid side chains to their local environment. *Proteins* 2008.
- [10] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov I.N, Bourne PE. "The Protein Data Bank" *Nucleic Acids Research* 2000.
- [11] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol* 1995.
- [12] Andreeva A, Howorth D, Chandonia JM, Brenner S.E, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucl. Acid Res* 2007.
- [13] Rossjohn J, Cappai R, Feil SC, Henry A, McKinstry WJ, Galatis D, Hesse L, Multhaup G, Beyreuther K, Masters CL, Parker MW. Crystal structure of the N-terminal, growth factor-like domain of Alzheimer amyloid precursor protein. *Nat Struct Biol* 1999.
- [14] Mueller-Dieckmann C, Ritter H, Haag F, Koch-Nolte F, Schulz GE. Structure of the ecto-ADP-ribosyl transferase ART2.2 from rat. *J Mol Biol* 2002.

- [15] Behnke CA, Yee VC, Trong IL, Pedersen LC, Stenkamp RE, Kim SS, Reeck GR, Teller DC. Structural determinants of the bifunctional corn Hageman factor inhibitor: x-ray crystal structure at 1.95 Å resolution. *Biochemistry* 1998.
- [16] Sandler BH, Nikonova L, Leal WS, Clardy J. Sexual attraction in the silkworm moth: structure of the pheromone-binding-protein-bombykol complex. *Chem Biol* 2000.
- [17] Kashiwagi T, Kunishima N, Suzuki C, Tsuchiya F, Nikkuni S, Arata Y, Morikawa K. The novel acidophilic structure of the killer toxin from halotolerant yeast demonstrates remarkable folding similarity with a fungal killer toxin. *Structure* 1997.
- [18] Yang X, Moffat K. Insights into specificity of cleavage and mechanism of cell entry from the crystal structure of the highly specific *Aspergillus* ribotoxin, restrictocin. *Structure* 1996.
- [19] Usón I, Sheldrick GM, de La Fortelle E, Bricogne G, Di Marco S, Priestle JP, Grütter MG, Mittl PR. The 1.2 Å crystal structure of hirutasin reveals the intrinsic flexibility of a family of highly disulphide-bridged inhibitors. *Structure* 1999.
- [20] Toth J, Cutforth T, Gelinis AD, Bethoney KA, Bard J, Harrison CJ. Crystal structure of an ephrin ectodomain. *Dev Cell* 2001.
- [21] Esposito L, Vitagliano L, Sica F, Sorrentino G, Zagari A, Mazzarella L. The ultrahigh resolution crystal structure of ribonuclease A containing an isoaspartyl residue: hydration and stereochemical analysis. *J Mol Biol* 2000.
- [22] Khademi S, Guarino LA, Watanabe H, Tokuda G, Meyer EF. Structure of an endoglucanase from termite, *Nasutitermes takasagoensis*. *Acta Crystallogr D Biol Crystallogr* 2002.
- [23] Song HK, Kim YS, Yang JK, Moon J, Lee JY, Suh SW. Crystal structure of a 16 kDa double-headed Bowman-Birk trypsin inhibitor from barley seeds at 1.9 Å resolution. *J Mol Biol* 1999.
- [24] Stehr M, Schneider G, Aslund F, Holmgren Lindqvist Y. Structural basis for the thioredoxin-like activity profile of the glutaredoxin-like NrdH-redoxin from *Escherichia coli*. *J Biol Chem* 2001.
- [25] Wilken J, Hoover D, Thompson DA, Barlow PN, McSparron H, Picard L, Wlodawer A, Lubkowski J, Kent SB. Total chemical synthesis and high-resolution crystal structure of the potent anti-HIV protein AOP-RANTES. *Chem Biol* 1999.
- [26] Crow A, Acheson RM, Le Brun NE, Oubrie A. Structural basis of Redox-coupled protein substrate selection by the cytochrome c biosynthesis protein ResA. *J Biol Chem* 2004.
- [27] Eklund H, Ingelman M, Söderberg BO, Uhlin T, Nordlund P, Nikkola M, Sonnerstam U, Joelson T, Petratos K. Structure of oxidized bacteriophage T4 glutaredoxin (thioredoxin). Refinement of native and mutant proteins. *J Mol Biol* 1992.
- [28] Dai S, Schwendtmayer C, Schürmann P, Ramaswamy S, Eklund H. Redox signaling in chloroplasts: cleavage of disulfides by an iron-sulfur cluster. *Science* 2000.
- [29] Guddat LW. Structural analysis of three His32 mutants of DsbA: support for an electrostatic role of His32 in DsbA stability Petersen TN, Henriksen A, Gajhede M. Structure of porcine pancreatic spasmolytic polypeptide at 1.95 Å resolution. *Acta Crystallogr D Biol Crystallogr* 1996.
- [30] Petersen TN, Henriksen A, Gajhede M. Structure of porcine pancreatic spasmolytic polypeptide at 1.95 Å resolution. *Acta Crystallogr D Biol Crystallogr* 1996.
- [31] Enguita FJ, Pohl E, Turner DL, Santos H, Carrondo MA. Structural evidence for a proton transfer pathway coupled with haem reduction of cytochrome c" from *Methylophilus methylotrophus*. *J Biol Inorg Chem* 2006.