Biomolecular hydration in protein-protein interaction, protein stability and aggregation and lead optimization : computational studies

Mohan Rajan Pradhan

2017

Mohan Rajan Pradhan. (2017). Biomolecular hydration in protein-protein interaction, protein stability and aggregation and lead optimization : computational studies. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/69994>

<https://doi.org/10.32657/10356/69994>

BIOMOLECULAR HYDRATION IN PROTEIN-PROTEIN INTERACTION, PROTEIN STABILITY AND AGGREGATION AND LEAD OPTIMIZATION: COMPUTATIONAL STUDIES

MOHAN RAJAN PRADHAN

2016

BIOMOLECULAR HYDRATION IN PROTEIN-PROTEIN INTERACTION, PROTEIN STABILITY AND AGGREGATION AND LEAD OPTIMIZATION: COMPUTATIONAL STUDIES

MOHAN RAJAN PRADHAN

Thesis Supervisors-
Dr. Chandra Verma and Dr. Kwoh Chee Keong

**Bioinformatics Institute (A*STAR) – School of Computer Science
and Engineering (NTU)**

**A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy**

2016

ACKNOWLEDGEMENTS

Firstly, I wish to thank the Bioinformatics Institute, A*STAR and the School of Computer Science and Engineering, NTU for their support in different forms at several stages of my PhD and my supervisors without whom my PhD would not have been possible. I am extremely grateful to BII for the research scholarship that allowed me to pursue my PhD and gave me the opportunity to present my work at two international conferences.

I wish to express my gratitude to my supervisor Dr. Chandra Verma for giving me the opportunity to pursue my passion for research in computational chemistry and structural biology. I feel extremely lucky to be part of his research group. I had the freedom to explore many things that I found interesting along the way. He often pushed me to ‘go extra mile(s)’ ☺ particularly while preparing the manuscripts, and though it took me some time, I realized that it is always worth the effort. He also motivated me to go deeper in the process of making scientific enquiries. This enormously helped me push my scientific endeavors beyond what was needed to publish papers. I sincerely thank him for this. Finally, I am grateful to him for the patience that he had while helping me improve my presentation skills- oral as well as written.

I am equally grateful to my supervisor Dr. Kwoh Chee Keong for his mentorship. He has been very kind and understanding. He allowed me the freedom to explore science in an area that interests me. He was always available for

discussions and I thank him for his time and support. Finally, his valuable suggestions helped me improve my thesis a lot.

I offer my heartiest thanks to the Verma group members. I think being part of the Verma group that comprised of 20+ members was a treat. The elaborate discussions on minute details of diverse topics that were presented during the weekly group meetings surely helped me grow a lot. I particularly learnt a lot from my colleague researchers and friends- Raghav, Adelene, Steve, Samuel, Thomas, Minh, Zohra, Chinh, Harini to name a few. However, I perhaps learnt the most from Dilraj Lama with whom I co-worked on the eIF4E project (Chapter5). The scientific discussions I had with him were very productive and insightful. I am very grateful to him. Also, I am equally thankful to Garima for the numerous discussions we had that surely helped my understanding.

I would also like to thank many unnamed researchers in the field of computational biophysics and structural biology whose scientific contributions have always motivated me.

I am very grateful to Ms. Betty Kee, Ms. Christine Low and Ms. Mun Kit for administrative support and to Mr. Yong Tai Pang, Mr. Robin Yee and Mr. Chee Siang for IT support at BII.

Last but not the least; I acknowledge my parents and my brother for their constant love and support without which this work would not have been possible.

TABLE OF CONTENTS

Acknowledgements	5
Table of contents	9
Abstract	13
List of figures	16
List of tables	21
List of algorithms	23
List of abbreviations	24
1. Biomolecular hydration: Introduction and Background	26
1.1 The intracellular water	26
1.2 Biomolecular role of water	27
1.2.1 Protein-protein/ligand interactions	29
1.2.2 Protein aggregation	32
1.3 Challenges and the need to further explore the biomolecular role of water	33
1.4 MD Simulations	34
1.5 Summary of Chapters	39
2. Designing a hydration metric to predict protein aggregation	42
2.1 Introduction	43
2.2 Methods	46
2.2.1 Preparation of protein dataset	46
2.2.2 Calculation of dehydrons	48
2.2.3 Calculation of evolutionary conservation of sequences	50
2.2.4 MD Simulations	53
2.2.5 Predicting APRs in proteins	56
2.3 Results and Discussion	56
2.3.1 Dehydron residue conservation in proteins	56
2.3.2 Dehydron residue conservation at APRs	58
2.3.3 Dehydron analysis in MD simulations	67
2.3.4 Comparison with sequence and structure based APR predictors	75
2.4 Conclusion	80

3. A combinatorial approach to probing the dynamics of aggregating mutants of p53 DNA Binding Domain reveals a novel ‘druggable’ pocket	85
3.1 Introduction	86
3.2 Methods	95
3.2.1 MD simulation	95
3.2.2 Umbrella sampling simulations	97
3.2.3 Benzene mapping simulations	98
3.2.4 MUTINF- Divergence calculations	99
3.2.5 Structure based pocket/cavity searching	99
3.2.6 Docking	100
3.3 Results and Discussion	101
3.3.1 Differential conformational sampling of S6-S7 turn in the wild and destabilizing mutants of p53 DBD	101
3.3.2 Biological/Functional implications of the ‘ <i>open</i> ’ conformation of the S6-S7 turn of the DBD	109
3.3.3 A network of interacting residues that regulate the conformation of the S6-S7 turn	118
3.3.4 Energetic barrier between the ‘ <i>open</i> ’ and ‘ <i>closed</i> ’ conformational states of the S6-S7 turn	130
3.3.5 D207A as a global site suppressor mutation for the destabilizing mutants of p53 DBD	131
3.3.6 Identification and characterization of a pocket that is predominantly formed in the destabilizing mutants	135
3.3.7 Is the S6-S7 pocket of p53 DBD druggable and functional?	136
3.3.8 Search for compounds that could dock in the S6-S7 pocket of p53 DBD	139
3.3.9 Testable hypothesis to probe the loss of function in destabilizing mutants of p53 DBD	143
3.4 Conclusion	144
 Preface to Chapter 4	 147
4. ‘JAL’: A software package for efficient characterization of hydration properties in structural ensembles of biomolecules	148
4.1 Introduction	149

4.2 Methods	155
4.2.1 Hydration properties	155
4.2.2 MD simulations	164
4.2.3 Buried water clusters	166
4.3 Results and Discussion	168
4.3.1 Case study 1: Role of buried waters in the stability of p53 DNA Binding Domain	168
4.3.2 Case study 2: Residence time of water molecules at the eIF4E-mRNA cap binding interface reveals a water bridge	176
4.3.3 Case study 3: Structure based search to displace interfacial waters in Scytalone dehydratase, EGFR kinase and eIF4E	178
4.3.4 Performance assessment	187
4.4 Conclusion	188
 Preface to Chapter 5	 191
 5. Systematic data mining reveals a water-bridge that mediates recognition of mRNA cap in eIF4E	 192
5.1 Introduction	193
5.2 Methods	197
5.2.1 Analysis of available crystallographic data on eIF4E	197
5.2.2 Structural alignment and detection of water clusters	205
5.2.3 System preparation for simulation	206
5.2.4 MD simulations	207
5.2.5 Water occupancy map and total residence time	208
5.2.6 Design and construction of surface entropically reduced eIF4E (23-217) for crystallization	209
5.2.7 Protein expression and purification	210
5.2.8 Crystallization	210
5.2.9 Data collection and refinement	210
5.3 Results and Discussion	212
5.3.1 Characterization of structural waters at the cap-binding interface	212
5.3.2 Water mediated interaction between protein and cap	216
5.3.3 W1 and W2 are held by interactions with residues W166 and N155	218
5.3.4 Preformed docking site at the cap binding interface	221
5.3.5 Crystal structure of apo eIF4E supports the presence of the water-bridge recognition module	227
5.4 Conclusion	230

6. Conclusions: summary, critiques and future directions	234
6.1 Hydration metrics to predict protein aggregation	234
6.1.1 Summary	234
6.1.2 Critiques and future directions	235
6.2 Molecular insights into the loss of function of tumor suppressor p53 in cancer	236
6.2.1 Summary	236
6.2.2 Critiques and future directions	236
6.3 Efficient characterization of hydration properties in structural ensemble of molecules	240
6.3.1 Summary	240
6.3.2 Critiques and future directions	241
6.4 Water-bridge mediates recognition of mRNA cap in eIF4E	242
6.4.1 Summary	242
6.4.2 Critiques and future directions	243
6.5 Concluding remarks	244
 Appendix	 246
Author's publications	248
References	249

ABSTRACT

The role of intracellular water in biomolecular function is well recognized and yet quantifying this is challenging. This is largely because protein hydration is extremely sensitive to the protein surface chemistry and topology and also to the environmental conditions. This makes protein hydration dependent on the protein being studied, and so it is hard to formulate general rule of how water affects protein function. There is no single theory that describes the precise nature of the microscopic role of water in protein structure stability, interaction, dynamics and function. Interrogating this, in my PhD I have focused on (1) Developing hypotheses/algorithms/programs of biomolecular hydration and (2) Using these to probe the role of water in protein function.

After providing a brief overview of biomolecular hydration and outlining the need for new insights to explain experimental observations such as aggregation ([Chapter 1](#)), a hydration metric is identified that is linked to Aggregation Prone Regions (APRs) in proteins. This hydration feature can provide guidance to complement existing strategies designed to inhibit protein aggregation ([Chapter 2](#)). Subsequently, I apply this feature to understand the loss of function through aggregation in the tumor suppressor protein p53 ([Chapter 3](#)), which allows us to model initial conformational events leading to the exposure of an aggregation prone region in p53 that could initiate the onset of aggregation. This also led to the identification of a novel ‘druggable’ pocket in p53. Driven by the observation that restoring the wild type function of mutant p53 is known to prevent tumors, I am in the process of screening libraries of compounds against this new pocket to identify binders that may guide the development of molecules to restabilize mutant p53. During the course of exploring

the hydration dynamics of proteins, I realized the need for a comprehensive, free, rapid and easy to use software package for characterizing hydration properties in a structural ensemble of a protein as derived from molecular dynamics simulation, for which a comprehensive software package, termed as ‘JAL’, is developed ([Chapter 4](#)). ‘JAL’ provides 3 novel utilities that have not yet been reported in other software packages: (1) computation of residence times of waters which are then used to rapidly identify water bridges involving multiple waters (2) rapid computation of buried waters in an explicit solvent MD trajectory using a novel “evaporation” technique and (3) a structure based search to compute the likelihood of favorably displacing interfacial waters; this method contrasts well with other available sophisticated but computationally expensive free energy methods. During the process of drug discovery, the question of which waters that exist in a binding site need to be displaced by the designed ligand is of critical importance and as yet has no universal solution. The utility of ‘JAL’ is demonstrated using case studies of 4 proteins for which a variety of relevant experimental data is available: the DNA Binding Domain of the transcription factor p53, the translation initiation factor eIF4E, Scytalone dehydratase which catalyzes the biosynthesis of melanin, and the kinase domain of the signaling protein EGFR. One of these proteins, eIF4E, is highly hydrated in its nucleotide binding pocket as is seen in several crystal structures, and presented us with a good opportunity to explore the importance of these water molecules using ‘JAL’ ([Chapter 5](#)). Combining systematic data mining and quantification of these hydration sites using ‘JAL’ a novel two water-bridge motif is identified that appears to be central to the nucleotide recognition by eIF4E. This is a novel finding and I believe will facilitate the design of specific inhibitors of eIF4E which is a therapeutic target in oncology.

Finally, I summarize and critique the work and discuss possible future directions
(Chapter 6).

LIST OF FIGURES

Figure 1.1	Algorithm of MD simulations	37
Figure 1.2	TIP3P water model	38
Figure 2.1	Distribution of wrappers in PDB	49
Figure 2.2	Distribution of dehydron-residue conservation in PDB	57
Figure 2.3	Proportion of conserved and non-conserved dehydron-residues in the vicinity of APRs	61
Figure 2.4	Dehydron residues at oligomerization and aggregation interfaces of hemoglobin	63
Figure 2.5	Dehydron residue conservation in p53 DBD	64
Figure 2.6	Prediction of APR in p53 DBD using Tango	66
Figure 2.7	Dehydrons in p53 DBD and its homologs	67
Figure 2.8	RMSD and dehydron distribution in aggregating proteins	68
Figure 2.9	MD simulations of p53 DBD	69
Figure 2.10	Percentage of APR covered by dehydrons in crystal and MD of some aggregating proteins	71
Figure 2.11	Frequency distribution of dehydrons across MD of p53 DBD	72
Figure 2.12	Dehydrons and RMSF of p53 DBD at its interaction interfaces	73
Figure 2.13	Dehydrons in p53 DBD across MD with respect to its interacting interfaces	74
Figure 2.14	APR prediction using existing predictors and using	

	dehydrons	76
Figure 2.15	Illustration of APR prediction using dehydrons	78
Figure 2.16	SAP calculations for hemoglobin and p53 DBD	79
Figure 3.1	p53 DBD with secondary structures labeled	88
Figure 3.2	Dominant conformational states of S6-S7 turn in p53 DBD	102
Figure 3.3	Quantifying the conformational states of S6-S7 turn in p53 DBD	104
Figure 3.4	Conformational states of S6-S7 turn in WT and mutant p53 DBD, p63 DBD and p73 DBD	108
Figure 3.5	Benzene density near APRs in p53 DBD	109
Figure 3.6	Interactions that stabilize the closed and open state of S6-S7 turn in p53 DBD	111
Figure 3.7	Antibody binding epitopes mapped on to p53 DBD	112
Figure 3.8	SASA of the antibody binding epitopes of WT and mutant p53 DBD	114
Figure 3.9	Conformation of N-terminus across simulations of WT and mutant p53 DBD	116
Figure 3.10	Conformational state of S6-S7 turn in WT and mutant p53 DBD with extended N-terminus	117
Figure 3.11	p53 DBD colored according to KL-div	119
Figure 3.12	Sequence of events leading to the change in conformation of S6-S7 turn	125
Figure 3.13	Conformation of S6-S7 turn in R174A-E180A mutant of p53	

	DBD	126
Figure 3.14	Illustration of the extended N-terminus and the interaction triad	127
Figure 3.15	MSA of p53 homologs	129
Figure 3.16	Energy landscape across the conformational states of S6-S7 turn	130
Figure 3.17	Schematic representation of the proposed secondary site suppressor mutation D207A	131
Figure 3.18	Conformational state of S6-S7 turn in WT and mutant variants of p53 DBD	133
Figure 3.19	Identification and Characterization of a ‘novel’ pocket in mutant p53 DBD	135
Figure 3.20	Dehydrons in p53 DBD	138
Figure 3.21	Residues lining the S6-S7 pocket	139
Figure 3.22	3 compounds identified from the structure based search	140
Figure 3.23	Docking and simulation of DTA in the S6-S7 pocket of p53 DBD	142
Figure 4.1	Hydration metrics in the software package	155
Figure 4.2	Benchmarking computations of buried waters in BPTI	160
Figure 4.3	Buried water in p53 DBD	169
Figure 4.4	Buried water in loop L3 of p53 DBD	171
Figure 4.5	RMSF of loops L2 and L3 in simulations of p53 DBD	173
Figure 4.6	Density of Zn across simulations of p53 DBD	174

Figure 4.7	Using residence times of water to map multi-water bridges in eIF4E	176
Figure 4.8	Structure based search of an interfacial water molecule in Scytalone dehydratase and EGFR kinase	180
Figure 4.9	Structure based search of an interfacial water molecule in eIF4E	184
Figure 5.1	Crystal structure of m7GTP bound to eIF4E	195
Figure 5.2	Number of water molecules around the cap binding interface of eIF4E	204
Figure 5.3	Crystal structure of eIF4E (PDB ID- 2w97)	205
Figure 5.4	Water clusters, occupancy maps and B-factors	212
Figure 5.5	Water mediated hydrogen bond network	214
Figure 5.6	Water clusters and occupancy map	215
Figure 5.7	Solvent occupancy at W1 and W2 in WT and mutants	218
Figure 5.8	Water bridge motif, Pillar residue conservation and RMSF	221
Figure 5.9	MSA of eIF4E and its homologs	222
Figure 5.10	Conformation of W166 and N155 and solvent density as a function of double point mutations in pillar residues	224
Figure 5.11	Conformation of W166 and N155 and solvent density as a function of single point mutations in pillar residues	226
Figure 5.12	Crystal structure of Apo eIF4E and the water bridge module	227
Figure 6.1	S6-S7 pocket and conformation of the N-terminus in WT and mutant p53 DBD	238

Figure 6.2	Modeling a disulphide bond in p53 DBD	239
Figure 6.3	Rationale behind W146H mutation.	240
Figure 6.4	Fragment screening in the nucleotide binding pocket of eIF4E	244
Figure A1	Distance distribution characterizing the conformational state of S6-S7 turn in simulations of wild type and mutants of p53 DBD	247

LIST OF TABLES

Table 2.1	Aggregating proteins with distinct folds	47
Table 2.2	Identification of APRs from MD trajectories of aggregating proteins using dehydrons	70
Table 2.3	Comparison of methods that predict APRs	75
Table 3.1	Number and types of MD simulations of p53 DBD	92
Table 3.2	KL-div values for destabilizing mutants of p53 DBD	120
Table 3.3	Quantifying the open state of S6-S7 pocket in WT and destabilizing mutants of p53 DBD	136
Table 4.1	Crystal structures of p53 DBD used for superimposition	167
Table 4.2	Percentage occupancy of atom types in a structure based search of waters in Scytalone dehydratase and EGFR kinase	183
Table 4.3	Percentage occupancy of atom types in a structure based search of water molecule in eIF4E	186
Table 4.4	Computational cost associated with the hydration calculations	187
Table 5.1	Crystal structures of eIF4E considered for the analysis of conserved waters	197
Table 5.2	Conserved waters in the cap-binding interface of cap-bound eIF4E	199
Table 5.3	Conserved waters in the cap-binding interface of cap-unbound eIF4E	202
Table 5.4	Hydration properties of water at positions W1 and W2	219

LIST OF ALGORITHMS

Algorithm 2.1	Computing dehydrons in a protein structure	50
Algorithm 2.2	Computing Evolutionary conservation of protein residues from Multiple Sequence Alignment (MSA)	53
Algorithm 4.1	Computing buried waters in an MD trajectory of a protein	159
Algorithm 4.2	Computing favorable/unfavorable solute/solvent interactions	161
Algorithm 4.3	Computing displaceability of interfacial water	164

LIST OF ABBREVIATIONS

DBD	DNA Binding Domain
WT	Wild Type
MD	Molecular Dynamics
APR	Aggregation Prone Region
PPI	Protein-Protein Interaction
PDB	Protein Data Bank
Hb	Hemoglobin
IARC	International Agency for Research on Cancer
TRT	Total Residence Time
m ⁷ GTP	7-methylguanosine triphosphate
RESP	Restricted Electrostatic Potential
U	Potential energy
SASA	Solvent Accessible Surface Area
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuations
BPTI	Bovine Pancreatic Trypsin Inhibitor
MSA	Multiple Sequence Alignment
KL-div	Kullback-Leibler Divergence
MCC	Mathews Correlation Coefficient
TTR	Transthyretin
SAP	Spatial Aggregation Propensity
ZMC	Zinc-metallachaperone

DTA	Dodecane trimethylamine
PMEMD	Particle Mesh Ewald Molecular Dynamics
WAP	Water Analysis Package
PSAP	Protein Structure Analysis Package
SO	Structure Overlap
NR	Non-redundant
ΔG_{desol}	Free energy of desolvation
NMR	Nuclear Magnetic Resonance
AMBER	Assisted Model Building with Energy Refinement
COM	Center Of Mass
RISM	Reference Interaction Site Model
GIST	Grid Inhomogeneous Solvation Theory
IFST	Inhomogeneous Fluid Solvation Theory
GCMC	Grand Canonical Monte Carlo
JAWS	Just Add Water Molecules
ML	Machine Learning
CSD	Cambridge Structural Database
LSH	Loop-Sheet-Helix motif
WHAM	Weighted Histogram Analysis Method

Chapter 1: Biomolecular Hydration:

Introduction and Background

This chapter reviews some of the studies in the past that probed the role of water molecules in protein structure and function. The focus is particularly on the water molecules which are present at the binding interfaces in proteins. This discussion will bring out the need for further efforts in this field. A brief introduction to Molecular Dynamics (henceforth referred to as MD) simulations, that are predominantly employed to study the role of water in protein hydration in this thesis, is also presented. This chapter concludes with a brief summary of all the chapters in the thesis.

1.1 Intracellular water

Water constitutes ~65% by mass of a human cell ([Ball, 2008a](#)). Given the dense network of organelles, macromolecules and ions within a cell, the water in a cell is dispersed non-uniformly ([Ball, 2003](#)). The average intramolecular distance between biological macromolecules is $\sim 10\text{\AA} - 20\text{\AA}$ ([Ball, 2008a](#)). This ‘crowding’ within cells is known to impart peculiar properties to intracellular water making it distinctly different from pure/distilled water. It is now known that intracellular water is more viscous and diffuses roughly an order slower than distilled water. Kuffel et al. theorized the nature of solute-solvent interactions to be responsible for

the higher density of intracellular water ([Kuffel & Zielkiewicz, 2012](#)). Intracellular water can be categorized into 2 classes- ‘vicinal or biological’ and ‘bulk’ water. Vicinal waters are those which are present within 5 Å of the biomolecule such that they comprise the 1st (<3.4 Å) and the 2nd (3.4 Å - 5 Å) solvation shells of a biomolecule. Any water molecule farther than this contributes to bulk water.

Optimal levels of cellular hydration are important and deviations from this are known to have deleterious effects on a cell. For instance, the assembly of a eukaryotic cell would be impossible without the hydrophobically driven assembly of the lipid bilayer. Similarly, enhanced cellular hydration is a characteristic feature of a cancerous cell. This led many research groups to propose how increased free water levels within a cell could be linked to cancer. The most widely held belief is that excessive solvation could accelerate the cellular metabolic rate to abnormal levels. This could result in speeding respiration, enabling cancer cells to take up more nutrients; in other words, water mediated processes will occur at a much faster rate ([Chaplin, 2006](#)).

1.2 Biomolecular role of water

It is intriguing to think how a protein (or any other biomolecule) finds its suitable interaction partner in such a ‘crowded’ intracellular environment within a cell. Clearly long range electrostatic interactions play a role. Additionally, correlated solvent flows around proteins within a 10 pico-second time scale ([Umezawa et al, 2007](#)) and 12 Å length scale have been observed suggesting that

fluctuations in these local correlated solvent flows could drive solute - solute interactions. In addition, chains of water molecules ('water-wire') have been implicated in connecting the donor and acceptor sites during proton-electron transfer, mediated by the flipping of hydrogen bonds along the chain. Thus, the stronger the bound water molecules in the water-wire, the more efficient is the proton-electron transfer ([Chaplin, 2006](#)). This facilitates long distance communication between distant parts of the cell especially where simple diffusion fails due to the absence of a charge concentration gradient. Water can also act as a catalyst in sampling protein conformations and folding. Given the enormous spatial and orientational sampling space available for solvent, water molecules should be able to rearrange themselves quickly to decrease the energy barrier associated with protein motions, thus smoothening the free energy surface of a protein ([Chaplin, 2006](#)). Additionally, water plays a crucial role in mediating protein-protein interactions (PPI) or protein-ligand interactions, which form nodes in biochemical pathways in cells. Together, these show how water modulates biomolecular processes inside the cell through its high dipole moment, multiple hydrogen bonding nature and rotational freedom ([Fernandez, 2004b](#)). The traditional view that water is an inactive background in which bio-macromolecules are dispersed, is rapidly being replaced by the realization that it has an active role at the molecular level ([Ball, 2003](#)). Any study of the biomolecular structure, interaction and dynamics will be incomplete without factoring in the role of vicinal waters. In the following sub-sections, the focus is particularly on the role of water in protein structure and function.

1.2.1 Protein-protein/ligand interactions

Water molecules have been observed at protein-protein/ligand interfaces. In some cases though, the interaction interface needs to be completely dehydrated. The desolvation free energy which is governed by a very subtle balance in the protein surface chemistry and topology is known to regulate the extent of interfacial hydration. Selective displacement of water molecules from the ligand binding sites in proteins is known to dominate ligand binding thermodynamics and kinetics. Additionally, water interactions with backbone hydrogen bonds in proteins are known to form the basis of a charge-hydrophobe interactions at protein-protein/ligand interfaces ([Fernandez, 2004a](#); [Fernandez & Berry, 2002](#); [Fernandez & Crespo, 2008](#); [Fernandez & Scheraga, 2003](#); [Fernandez & Scott, 2003](#)), which have significant implications in structure guided ligand design.

Exclusion of interfacial water upon biomolecular interactions- When two proteins come together to interact, the intervening water is excluded in several cases. It has recently been established that this displacement contributes maximally to the binding affinity in protein interactions. The idea is that this interfacial water interacts with the protein and is thus constrained in its mobility. The release of this bound water into the bulk solvent results in gain in entropy for that water molecule and hence the overall process, by increasing the degrees of freedom. This is particularly strong in cases where electrostatics and van der Waals interactions between water and protein residues are weaker or absent, for example, when the protein surface is hydrophobic. This is because the displacing water doesn't have to

overcome much enthalpic forces arising from its interactions with the protein ([Barillari et al, 2007](#)).

Water mediated biomolecular interactions- There are several situations where the water molecules are retained at the interface of proteins or protein-ligand complexes. These waters likely form hydrogen bonds with both partners. These are termed as water mediated intermolecular interactions. It is known that such waters play an important role in the recognition process. For example, interfacial waters have been known to mediate interactions between molecules that do not have optimal shape complementarity at their interfaces ([Barakat et al, 2011](#)). The nature of interface solvation has also been reported to affect the stability of the multimeric state of a protein. For example, a study of the solvation properties of insulin oligomerization interfaces revealed how dehydration affects the stability of the multimeric interface and is linked to increase in blood sugar levels in Diabetes mellitus ([Bagchi & Roy, 2014](#)). There appears to be a correlation between the residence time of water and water hydrogen bond relaxation with differential stabilities of the dimeric and hexameric interfaces of insulin. Similarly, it is suggested that interfacial waters at the hemoglobin dimeric interface could facilitate co-operativity associated with oxygen binding ([Royer et al, 1996](#)). Systematic mutagenesis studies lead to the identification of hydrogen bonded cluster of water molecules between hemoglobin dimers and confirmed its role in communication between subunits associated with oxygen binding co-operativity.

Interfacial water is also thought to facilitate non-specific binding; for instance the binding of a receptor to a range of ligands. The oligopeptide binding

protein (OppA) is able to recognize and interact with 2-5 residue peptides regardless of the peptide sequence in most cases. This is reportedly facilitated by the water molecules in the peptide binding site of the protein (Tame et al, 1994; Tame et al, 1996). In contrast, the presence of water molecules in the ligand binding cavity of the Thyroid hormone receptor has been reported to confer binding selectivity (Martinez et al, 2009). Similarly, an invariant water network is thought to be important for selectivity in the binding of bosutinib to a kinase (Levinson & Boxer, 2014). In yet another instance, the presence of waters in the binding cavity of arabinose binding protein explains its selectivity towards L-arabinose over D-fucose and D-galactose. Water molecules in the binding cavity provide optimal interactions and a deviation from this (as in case of other non-specific ligands) could affect binding (Quiocho et al, 1989).

Indirect involvement of water at protein-protein/ligand interfaces- While the more common charge-charge and hydrophobe-hydrophobe interactions have been known, the existence of a charge-hydrophobe interaction was elucidated by Fernandez (Fernandez, 2004a) almost a decade ago. This interaction, which has its origin in protein backbone hydrogen bonds, can solely be attributed to the presence of water in the surrounding environment. Fernandez developed the concept of a Dehydron or an Underwrapped Backbone Hydrogen Bond (UDHB) which was defined as a protein backbone hydrogen bond which is accessible to water (Fernandez & Berry, 2002). This occurs when there are not enough hydrophobic side chain groups packing the backbone hydrogen bond. This renders the hydrogen bond accessible to solvent, which can compete for hydrogen bonding with the

donor/acceptor site in the backbone and weakens the backbone hydrogen bond, resulting in perturbations to the local secondary structure and protein stability.

Dehydrons have been implicated in several different ways. Dehydron containing sites on proteins are hypothesized to have evolved to be interacting sites, where a hydrophobic entity from the interaction partner provides the shielding required for blocking water molecules from the solvent. This feature has been exploited to achieve binding specificity (more details in [section 2.1](#)). The extent of protein backbone hydration imposes a constraint on protein folding. In another work ([Fernandez, 2004b](#)), the manipulation of hydration of a glucagon helix was shown to affect its membrane permeability- a strategy that could be employed to engineer molecules to cross biomembranes. Excessive underwrapping has also been observed in amyloid proteins.

1.2.2 Protein aggregation

Another PPI (Protein-Protein Interaction) mediated by water, and is much more difficult to characterize, is protein aggregation. Aggregation is usually deleterious to the cell and leads to loss of protein functions culminating in disorders such as Alzheimer's. Usually proteins contain regions which have the potential to nucleate aggregation. Such regions are termed as Aggregation Prone Regions (APRs). They are usually buried in the protein core, and upon exposure, lead to self assemble into large β -sheet structures. Therefore, it is important to identify such aggregation prone regions in proteins, understand their mechanism and design molecules that prevent aggregation where relevant. As discussed earlier, aberrant solvation during protein folding could lead to a misfolded or

partially folded protein, exposing these APRs and this sometimes culminates in aggregation. The driving force for the process is dehydration of water between aggregation prone regions, which could be hydrophobic or hydrophilic ([Thirumalai et al, 2012](#)).

1.3 Challenges and need to further explore the biomolecular role of water

Developing theories of protein hydration is extremely challenging largely because hydration of proteins is extremely sensitive to protein surface chemistry and topology which themselves are very diverse. Furthermore, the role of hydration seems not only dependent on the protein system being studied but also on environmental conditions like temperature, pressure, pH, etc. Das et al. ([Das & Matysiak, 2012](#)) showed that the impact of low temperatures & pressure on the solvation of hydrophobic polymers, and solute-solvent interactions are believed to be the major contributor to cold-pressure mediated protein denaturation. Although there are several theories suggesting that such confined/restrained water is responsible for the differences between the behavior of biological and bulk water, there is as yet no consensus ([Ball, 2003](#)). Understanding the entire spectrum of processes governed by water and its mechanisms will require much more time and effort ([Levy & Onuchic, 2004](#)). But with advances in experimental and computing technologies that can probe events at the nanoscale level, the time is right to probe and develop detailed insights into biomolecular hydration ([Ball, 2008a](#); [Ball,](#)

2008b). This thesis develops methods based on MD simulations to interrogate biomolecular hydration towards developing new insights into its effects of protein form and function.

1.4 MD Simulations

MD simulations are computational techniques for studying the time evolution of the motions of atoms and molecules under a set of predefined conditions. MD yields fluctuations and conformational changes in a molecule with respect to time, which is governed by chemical interactions between the atoms of the system. The ensemble of structures generated using MD will be used to quantify aspects of the biomolecular role of water (as will be discussed in [Chapters 2, 3, 4 and 5](#)).

The fundamental assumption underlying MD, which is known as Ergodic hypothesis, is that a system's path in phase space (a function of position and momenta of particles in a system) on a surface of constant energy will, in the long run, visit each region (microstate) of that surface equally often. In other words, over a long simulation, the system behaves similarly when averaged over time as averaged over space. The Ergodic hypothesis forms the foundation for the theorem of equipartition of energy which states that molecules in thermal equilibrium have the same average energy associated with each independent degree of freedom of their motion. A degree of freedom is any independent variable that affects the range of states in which a system may exist, in particular any of the directions in which independent motion can occur. In a state of thermal equilibrium, all the

accessible microstates (any point in phase space) of a system are equally probable and the time averages of all the macroscopic quantities (e.g. total energy, pressure, temperature, volume, density, etc.) are time independent. Of course, this happens when the underlying probability distributions of these macroscopic quantities are themselves independent of time.

The process of carrying out an MD simulation involves defining the initial coordinates of the biomolecule (e.g. protein). These are available as X-ray or NMR structures and can be downloaded from the Protein Data Bank (PDB) ([Berman et al, 2000](#)). Alternatively, if the experimentally determined coordinates are not available, these can be modeled using molecular modeling approaches such as homology modeling ([Sali & Blundell, 1993](#)).

Given the coordinates of the atoms that constitute a protein, molecular mechanics force fields are then used to estimate the forces between the particles (here atoms) and the associated potential energy. The potential energy (U) of a system of molecules is calculated as shown in [eqn. 1.1, 1.2, 1.3 and 1.4](#)-

$$U_{total} = U_{covalent} + U_{noncovalent} \quad \dots (1.1)$$

The covalent (defines the atoms that separated by one or two bonds) and non-covalent (the atoms that are separated by 3 or more bonds) terms can be further decomposed as follows-

$$U_{covalent} = U_{bond} + U_{angle} + U_{dihedral} \quad \dots (1.2)$$

$$U_{noncovalent} = U_{electrostatics} + U_{van\ der\ waals} \quad \dots (1.3)$$

The following functional form of the potential energy (U) (eqn. 1.4) (Weiner et al, 1984) shows how each of the above terms are accounted for.

$$U = \sum_{\text{bonds}} \frac{1}{2} k_b (r - r_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{\text{torsions}} k_\phi [1 + \cos(n\phi - \delta)] \\ + \sum_{i < j} \sum \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{i < j} \sum 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \dots (1.4)$$

k_a , k_b and k_ϕ are the force constants associated with bond lengths, bond angles and dihedral angles. r is the bond length, r_0 is the equilibrium bond length. θ is the bond angle and θ_0 is the equilibrium bond angle. Φ is the dihedral angle, δ is the phase angle. σ_{ij} is the distance at which the van der Waals energy is zero. q is the atomic charge, r_{ij} is the interatomic distance and ϵ_{ij} is the depth of the van der Waals energy of interaction well and corresponds to the optimum van der Waals interaction between two atoms.

Then, to compute the motions of the atoms under the above constraints, at a given temperature, Newton's second law of motion (eqn. 1.5) is solved, which is $\mathbf{F} = m\mathbf{a}$, where 'F' is the force exerted on a particle of mass 'm' having an acceleration 'a'. In the 1st step of the interaction, initial velocities are assigned to the atoms that make up the protein according to the Maxwell-Boltzmann probability distribution at a given temperature.

$$\frac{d^2 x_i}{dt^2} = - \frac{F_{x_i}}{m_i} \dots (1.5)$$

Given that we know the force on each particle of our system, one can now compute acceleration. These values are used to get the coordinates and velocities at

the successive time steps using the following integration equations (eqn. 1.6 and 1.7)-

$$r(t + \delta t) = r(t) + v(t)\delta t + \left(\frac{1}{2}\right)a(t)\delta t^2 \quad \dots (1.6)$$

$$v(t + \delta t) = v(t) + \left(\frac{1}{2}\right)[a(t) + a(t + \delta t)]\delta t \quad \dots (1.7)$$

The basic concept is to split the integration into many small intervals, each separated by a constant time δt . For a finite δt , this is called Euler's approximation. Using this approach, given the initial positions and velocities, and a formulation for computing the force on each atom of a protein, we have the ability to simulate a phase space trajectory. Figure 1.1 summarizes the algorithm of MD as:

Figure 1.1: Algorithm of MD simulations

Setting up a MD simulation involves several technicalities. For instance, ACE (a neutral acetyl group) and NME (a neutral methylamine group) moieties are used to cap the charges at the N-terminal and C-terminal residues respectively. The biomolecular environment being at least partly aqueous, protein (may include

ligands) is simulated in a cubic box of TIP3P water molecules (Jorgensen et al, 1983a). Figure 2 represents the TIP3P water model. It is a 3 point model where the 3 atoms of the water molecule correspond to the 3 interaction points. Other than the parameters shown in Figure 2, the model also has values for the non-bonded parameters- ϵ and σ (eqn. 1.4). But what happens to the atoms situated at the edge of the box when the system is simulated? Periodic boundary conditions are used to correct for inaccuracies introduced due to water molecules at the edge of the box. The rate-limiting step in MD is the computation of long range electrostatics which decay as an inverse of the distance between the particles ($1/r_{ij}$) (eqn. 1.4). Therefore, Particle Mesh Ewald method (Darden et al, 1993b) is used to calculate long range electrostatic interactions beyond a threshold distance ($\sim 8 \text{ \AA}$) which reduces the time complexity to $O(n \log(n))$ as opposed to $O(n^2)$ when all the pairwise interactions (n) are accounted for explicitly.

Figure 1.2: TIP3P water model. Charges on atoms are labeled in blue. Bond length and angle are labeled in black.

However, before an MD simulation is set up, we need to relax the structure in order to remove atomic clashes, if any; this is particularly so when one is simulating a crystal structure where the crystal contacts often introduces close packing which must be relieved. For this purpose, we minimize the potential

energy (eqn. 1.4) of the system using Gradient descent algorithms. Energy minimization involves moving the atoms according to the force acting on them to locations corresponding to a net zero force on the atoms (Leach, 2001). The derivatives of potential energy with respect to the coordinates are calculated numerically. After energy minimization, the system reaches the nearest local minima on the potential energy surface. A potential energy surface is a 'hypersurface' with $3N-6$ dimensions ('3' refers to the 3 coordinates- x, y and z; 'N' refers to the number of atoms in a system; '6' refers to 3 rotations and 3 translations of the whole system which do not change the potential energy of the system).

Once the MD simulations have been carried out for a period of time, the positions of the atoms are collected and subject to various enquiries (discussed in Chapters 2, 3, 4 and 5) which are then interpreted in developing hypotheses linking the motions with function.

1.5 Summary of Chapters

The thesis contains 6 chapters which are summarized as below-

In **Chapter 2 - Designing a hydration metric to predict protein aggregation**, I study aggregating proteins and demonstrate that a feature, obtained by combining 2 simple sequence and structure evaluation parameters- evolutionary conservation and the extent of backbone hydration, respectively, can be suggestive of probable aggregation associated regions of proteins. Comparison with the existing predictors

of Aggregation Prone Regions (APRs) leads me to conclude that this feature can complement present approaches designed to study and inhibit protein aggregation.

In Chapter 3 - A combinatorial approach to probing the dynamics of aggregating mutants of p53 DNA Binding Domain reveals a novel ‘druggable’ pocket, the feature developed in Chapter 2 is used to postulate the molecular events in the onset of aggregation in the oncogenic mutants of the p53 DNA Binding Domain (DBD). A novel ‘druggable’ pocket in the mutant forms of the protein is observed. It is currently being probed if this pocket can be targeted with compounds designed to restore wild type like native structure in mutant DBD.

Chapter 4 - ‘JAL’: A software package for efficient characterization of hydration properties in structural ensembles of biomolecules, I present a comprehensive software package that can compute hydration properties in a structural ensemble from MD simulations. The package encompasses 3 novel algorithms- (1) to compute buried waters, (2) to identify multi-water bridges in proteins and (3) to evaluate the displaceability of interfacial waters.

In Chapter 5 - Systematic data mining reveals a water-bridge that mediates recognition of mRNA cap in eIF4E, I use the algorithms and tools developed in Chapter 4 to probe the role of water molecules at the mRNA cap-binding interface of the protein the translation initiation factor eIF4E in translation. A systematic characterization of hydration properties reveals a water bridge at the interface that

appears to be crucial for ligand recognition. This knowledge will be used to design novel inhibitors of eIF4E, which is also an established cancer target.

Chapter 6 - Conclusions: summary, critiques and future directions. This chapter lists the main conclusions and limitations of the work, and briefly discusses future directions.

Chapter 2: Designing a hydration metrics to predict protein aggregation

Aggregation is an irreversible form of protein complexation and often toxic to cells. The process entails partial or major unfolding that is largely driven by hydration. In this chapter, attempts are made to model the role of hydration in aggregation using ‘Dehydrons’. ‘Dehydrons’ are unsatisfied backbone hydrogen bonds in proteins that seek shielding from water molecules by associating with ligands or proteins.

I find that the residues at aggregation interfaces have hydrated backbones, and in contrast to other forms of protein-protein interactions, are under less evolutionary pressure to be conserved. Combining evolutionary conservation of residues and extent of backbone hydration allows us to distinguish regions on proteins associated with aggregation (non-conserved dehydron-residues) from other interaction interfaces (conserved dehydron-residues). I believe that this novel feature can complement the existing strategies used to investigate protein aggregation/complexation.

2.1 Introduction

Correctly folded native state structures of proteins are crucial for their functions. However, local destabilization/misfolding can result in non-specific irreversible self-association between protein monomers, leading to aggregation (Buck et al, 2012). Protein aggregation is complex and diverse, often governed by a delicate balance of factors such as solvent, cellular milieu, mutations, hydrophobicity, chaperons, pH, temperature, pressure, the extent of unfolding required to form an aggregation competent state, etc (Aguzzi & O'Connor, 2010; Chaplin, 2006; Chong & Ham, 2014; Ross & Poirier, 2005; Thirumalai et al, 2012). Depending on the structure of the aggregate, protein aggregation is classified as amyloid and amorphous. Amyloids are highly ordered β -rich structures and often toxic to cells. In contrast, amorphous aggregation is a more common form of irreversible protein self-association, where the aggregates lack a defined structure. Although, amorphous aggregates are usually non-toxic, the associated loss of protein function can predispose the cell to a disease phenotype. Aggregation has been associated with a variety of disease conditions like Alzheimer's, Parkinson's, Type II diabetes and cancer (Barakat et al, 2011; Chiti & Dobson, 2009; Xu et al, 2011). Aggregation also thwarts protein design and production efforts (Buck et al, 2012). However, it has a critical role in regulating levels of a protein within a cell (Wolozin, 2012). A deep understanding of the mechanism of protein aggregation thus is essential. Aggregation of proteins usually results from nucleation of segments that are largely hydrophobic, and are often buried in the protein core. Such regions are referred to as Aggregation Prone

Regions (APRs). Thermal fluctuations or major unfolding events ([Chiti & Dobson, 2009](#)) can trigger their exposure, which then cascades into β -sheet containing aggregates. Understanding features that predispose the exposure of these APRs could thus provide valuable insights into the mechanism of aggregation and guide therapeutic intervention.

Several sequence and structure based approaches have attempted to identify APRs in proteins ([Buck et al, 2012](#); [Tsolis et al, 2013](#)). Sequence based methods evaluate the propensity of different sequence segments to form APRs. The maximum value of Mathews Correlation Coefficient (MCC) that has been achieved for this class of predictors is 0.24 ([Walsh et al, 2014](#)). This suggests the need for improvement in APR prediction methods. Structure based methods predict APRs based on the analysis of features that are derived from protein 3D structures. Most structure based attempts have been focused at the peptide level ([Cellmer et al, 2007](#); [Dima & Thirumalai, 2002](#); [Gsponer et al, 2003](#); [Klimov & Thirumalai, 2003](#)). It is clear that a more comprehensive understanding of the process of aggregation is needed ([Thirumalai et al, 2012](#)). To further probe this problem, the role of specific water-protein interactions in aggregation is examined.

The tertiary structures of proteins are mostly stabilized by backbone hydrogen bonds between the carbonyl and amide groups. These bonds are subject to perturbation by the environment, notably by the highly polar water molecules. Water molecules can interact electrostatically with the charges on such backbone carbonyl and amide groups, resulting in the transient breaking of the backbone hydrogen bonds. This compromises the local secondary structure, destabilizing the

protein structure. Extensive structural analysis of proteins has revealed that often the backbone hydrogen bonds are shielded from solvent by non-polar groups from the surrounding side-chains. Insufficient shielding results in exposing these backbone atoms to solvent water, leading to instability. Backbone hydrogen bonds ($\text{N-H}\cdots\text{O}=\text{C}$) in proteins that are accessible to solvent are referred to as ‘Dehydrons’ (or Under-dehydrated hydrogen bonds, UDHB) and the non-polar groups as ‘wrappers’ (Fernandez & Berry, 2002; Fernandez et al, 2003).

Dehydrons are shielded from solvent water by engaging in interactions with other proteins or ligands, making them sticky in nature. Dehydrons have been widely used to understand protein structure and function (Fernandez, 2004a; Fernandez & Crespo, 2008; Fernandez et al, 2003). For example, they have been used to identify protein interaction interfaces (Fernandez & Scheraga, 2003); the number of dehydrons in a protein (or domain) has been found to be directly proportional to the number of its known interacting partners. This naturally lead to the realization that dehydron-residues must be evolutionarily conserved (Fernandez, 2004a). The average number of wrappers for all the backbone hydrogen bonds in proteins has been found to be fairly constant and suggestive of protein stability (Fernandez & Berry, 2002). In this context, it was shown that most of the backbone hydrogen bonds in soluble proteins are well protected by side-chain hydrophobic groups. Wrapping has also been hypothesized to modulate protein folding (Fernandez & Scheraga, 2003). Moreover, dehydrons have also been implicated in drug discovery. Recently a small molecule was designed to wrap a dehydron in a protein cavity of an oncogenic p53 mutant in an attempt to

rescue protein function ([Accordini et al, 2013](#); [Fernandez & Scheraga, 2003](#)). Dehydrons have also been exploited to enhance selectivity of protein kinase inhibitors, where a kinase inhibitor was successfully reengineered to increase its selectivity as a treatment against gastrointestinal stromal tumors ([Fernandez et al, 2007](#)).

Here, I focus on aggregation and dehydrons and enquire whether the APRs of proteins are distinct from other interacting regions of proteins by examining the evolutionary conservation of dehydron-residues. I show that by combining evolutionary conservation information together with backbone hydration enables the identification of regions in proteins that are associated with aggregation. I demonstrate this using two well studied examples, hemoglobin and p53. I further show how this method can complement existing methods to predict APRs in proteins.

2.2 Methods

2.2.1 Preparation of protein dataset

A high quality non-redundant dataset of PDB proteins was compiled using the PISCES web server ([Wang & Dunbrack, 2005](#)). This set consists of ~1200 structures with resolution higher than 2 Å and a maximum pairwise sequence identity of 30% with any other protein in this dataset. Structures with any missing atoms/residues are excluded to avoid modeling errors. A non-redundant dataset of 27 aggregating proteins with distinct folds was primarily compiled from AMYPdb

database (Pawlicki et al, 2008) (Table 2.1). Additionally, I examine two proteins which are known to aggregate upon mutations: hemoglobin and p53.

Sr. No.	Protein	Chain Length	Technique	Resolution (Å)
1	1A1N_B	99	X-ray	2
2	1BEB_A	162	X-ray	1.8
3	1AV1_A	201	X-ray	4
4	1BPH_B	30	X-ray	2
5	1BYV_A	33	NMR	NA
6	1CEW_I	108	X-ray	2
7	1CKT_A	71	X-ray	2.5
8	1DZG_I	432	X-ray	2.8
9	1F9B_A	695	X-ray	2.7
10	1FZA_A	87	X-ray	2.9
11	1G6Y_B	261	X-ray	2.8
12	1G96_A	120	X-ray	2.5
13	1GHL_A	130	X-ray	2.1
14	1HKQ_A	132	X-ray	2.75
15	1I4M_A	108	X-ray	2
16	1IE4_A	127	X-ray	2.5
17	1KCQ_A	104	X-ray	1.65

18	1MG8_A	78	NMR	NA
19	1MWP_A	96	X-ray	1.8
20	1R5N_A	467	X-ray	2.9
21	1RW5_A	199	NMR	NA
22	1T34_A	435	X-ray	2.95
23	1WLAA	153	X-ray	1.7
24	1X3BA	146	NMR	NA
25	1XQ8A	140	NMR	NA
26	2A01A	243	X-ray	2.4
27	3PF4A	67	X-ray	1.38

Table 2.1: Aggregating proteins with distinct folds mined from the primary dataset. This dataset of aggregating proteins is used for the statistical study of dehydron-residue conservation.

2.2.2 Calculation of dehydrons

Dehydrons are calculated according to the method proposed by Ariel Fernandez: counting the number of hydrophobic groups within a fixed volume around backbone H-bonds (Fernandez, 2004a; Fernandez & Berry, 2002; Fernandez & Crespo, 2008; Fernandez & Scheraga, 2003; Fernandez & Scott, 2003). The donor-acceptor distance, for the identification of H-bonds, was $< 3.5 \text{ \AA}$. Dehydrons were calculated using the ‘Dehydron’ plugin of PyMOL with the standard parameter values. The radius of the desolvation spheres, which defines the local environment around a backbone hydrogen bond, was fixed at $6.5 \text{ \AA}$ around the CA atoms of residues involved. A backbone hydrogen bond was classified as a

dehydron if the number of non-polar groups viz. CH_n (n= 1, 2 or 3) within the desolvation domain was less than 19. To benchmark these calculations, the data on the distribution of wrappers around backbone hydrogen bonds in the PDB has also been reproduced (Figure 2.1) (Algorithm 2.1).

Figure 2.1: Overall distribution of wrappers around backbone hydrogen bonds in the PDB.

Algorithm 2.1: Calculating dehydrons in a protein structure

Input: A protein structure

Output: Dehydrons

```
1  For each Backbone-hydrogen bond bh in the input protein structure
2      com  $\leftarrow$  COM (O, N)      ### COM refers to Center of Mass
3      wrapper_count  $\leftarrow$  0
4      For all non-hydrogen atoms nh in the input protein structure
5          If dist (com – nh)  $\leq$  6.5 Å:
6              wrapper_count  $\leftarrow$  wrapper_count + 1
7          If wrapper_count < 19:
8              bh  $\rightarrow$  dehydron
9          Else:
10             bh  $\rightarrow$  not_a_dehydron
11  End For
12  End
```

2.2.3 Calculation of evolutionary conservation of sequences

In the spirit of the method of Sander and Schneider (Sander & Schneider, 1994), Shannon entropy of aligned amino acid sequences is used to quantify the evolutionary conservation of dehydron residues (Algorithm 2.2). For each protein in the aggregating and random datasets, a multiple alignment of at least 60 sequences displaying more than 30% identity over the whole chain was extracted from the HSSP database. The Shannon entropy s_i at a position i was then calculated as:

$$S_i = -\sum p_k \ln(p_k) \quad \dots(2.1)$$

where p_k is the probability that the i th position contains a residue belonging to class 'k', the amino acids being grouped into 7 classes based on the similarity of their environment in protein structures (Guharoy & Chakrabarti, 2005). The 7 classes are: Class I (Ala, Val, Leu, Ile, Met, Cys); Class II (Gly, Ser, Thr); Class III (Asp, Glu); Class IV (Asn, Gln); Class V (Arg, Lys); Class VI (Pro, Phe, Tyr and Trp); and, Class VII (His). Mutations within each class are not expected to alter the biochemical environment, hence such substitutions are assumed to be conserved (Mirny & Shakhnovich, 1999). As the sequence entropy is a measure of the divergence at each position in the alignment, a value of zero indicates that the position is fully conserved, i.e. the position is under an evolutionary pressure to retain a particular type of residue, whereas a higher value of s_i signifies greater divergence. Because s_i depends on the sample of aligned homologous sequences and the position, its value is normalized by dividing by the average entropy of all the residues (n) in the dataset:

$$S_i = s_i / \left(\frac{\sum s_i}{n} \right) \quad \dots (2.2)$$

Hence, $S_i < 1$ implies that the position 'i' is better conserved (grey in the figures which refer to evolutionary conservation of residues in this chapter) than the dataset average, while $S_i > 1$, implies that the position 'i' is more divergent (orange in the figures which refer to evolutionary conservation of residues in this chapter). Shannon entropy is also computed without considering the above seven amino acid classes, such that a mutation to any other residue will be regarded as a

substitution. I find that the normalized entropies, computed using either of these methods, are very similar.

The Tango algorithm ([Fernandez-Escamilla et al, 2004](#)) was used to calculate the aggregation propensity in the p53 DBD Binding Domain. This algorithm predicts β -aggregation propensities for protein/peptide residues given an input amino acid sequence. It calculates the side chain – side chain interaction energies to generate a phase space for the α -helix, β -turn and β sheet aggregate population. Based on this, it calculates the relative propensities of a residue to occupy each state which is output as a score on a scale of 1-100 for each residue. Figures were prepared using PyMOL ([Schrodinger, 2010](#)) and Chimera ([Pettersen et al, 2004](#)).

Algorithm 2.2: Computing Evolutionary conservation of protein residues from Multiple Sequence Alignment (MSA)

Input: MSA of homologs of non-redundant dataset proteins from HSSP database

Output: Evolutionary conservation score for protein residues

```
1      Input multiple sequence alignments for non-redundant dataset from HSSP
      database
2      For each protein  $j$  in the non-redundant dataset
3           $counter \leftarrow 0$ 
4          For each residue  $i$  in protein  $j$ 
5               $s(i) \leftarrow - \sum p_k \ln(p_k)$  ,  $k = 1$  to  $7$  ( $k$  is amino-acid class)
              #Shannon-entropy
6               $E_{tot} \leftarrow \sum s(i)$ 
7               $counter \leftarrow counter + 1$ 
8          End For
9           $E_{avg} \leftarrow E_{tot} / counter$       # Average entropy of a protein  $j$ 
10         For each residue  $i$  in protein  $j$ 
11              $S(i) \leftarrow s(i) / E_{avg}$       # Normalized entropy
12         End For
13 End For
```

2.2.4 MD Simulations

Dehydrons are used to model a structural property. Given that the crystallographic data represents a single snapshot, it was decided to explore the variations in dehydrons as a function of conformational variations. To investigate this, all-atom explicit solvent simulations of aggregating proteins in their apo states is carried out. For this, the following crystallographic structures of proteins are

used because experimental data on their APRs is known: PDB IDs and their resolution- 1a1nB ([Smith et al, 1996](#)) [2 Å], 1bebA ([Brownlow et al, 1997](#)) [1.8 Å], 1hkqA ([Forest & Filutowicz, 2003](#)) [2.75 Å], 1ie4A ([Wojtczak et al, 2001](#)) [2.4 Å], 1kcqA ([Kazmirski et al, 2002](#)) [1.65 Å], 1wlaA ([Maurus et al, 1997](#)) [1.7 Å], 3pf4A ([Sachs et al, 2012](#)) [1.38 Å]. These proteins were chosen based on high resolution of their crystallographic structures and very few missing atoms in their crystallographic structure. Coordinates of the missing atoms were generated using the tleap module in Amber. Simulations were carried out using the Amber ff99SB force field ([Hornak et al, 2006](#)). The N and C termini were capped with ACE and NME moieties respectively. Crystallographic waters were retained and the protein systems were solvated in a TIP3P ([Jorgensen et al, 1983a](#)) water box with a minimum separation of 10 Å between the protein atoms and the edge of the water box. The solvated protein systems were neutralized with the addition of appropriate number of counter ions. The structures were then minimized for 1000 steps each of steepest descent followed by conjugate gradient. The proteins were then gradually heated to 300 K in an NVT ensemble followed by 2ns of equilibration in the NPT ensemble. Finally, a 50 ns production run in NPT was carried out for each system using the GPU implementation of PMEMD ([Gotz et al, 2012](#)). A cut off of 8 Å was used to evaluate the Lennard-Jones (LJ) and short-range electrostatic interactions, while the long-range electrostatic interactions were evaluated using the particle-mesh Ewald (PME) method with a grid spacing of 0.1 nm and a fourth-order interpolation ([Darden et al, 1993b](#)). Temperature was regulated using Langevin thermostat ([Loncharich et al, 1992](#)). Constant pressure (1.0 bar) using weak

coupling with a relaxation time of 1 ps was used in the simulations. SHAKE algorithm ([Ryckaert et al, 1977a](#)) was used to constrain the bond lengths with dangling hydrogen atoms in amino acids, thus enabling an integration time step of 2 fs; coordinates were saved at an interval of 5 ps.

2.2.5 Predicting APRs in proteins

APRs in proteins are examined by using the following predictors: Tango ([Fernandez-Escamilla et al, 2004](#)), Waltz ([Maurer-Stroh et al, 2010](#)), Amlypred ([Tsolis et al, 2013](#)), PASTA ([Walsh et al, 2014](#)) and ZipperDB ([Goldschmidt et al, 2010](#)). These methods constitute the sequence based predictors that evaluate the propensity of protein segments to be an APR; respective web servers were used with default parameters. In addition, Spatial Aggregation Propensity (SAP) ([Chennamsetty et al, 2009](#)), is used for the structure based predictions of APRs. SAP combines residue hydrophobicity and solvent accessibility to provide a measure of their likelihood to behave as APRs. The SAP measure for APRs was calculated for every atom by constructing a sphere of radius 10 Å around it.

Figures were made using PyMOL ([Schrodinger, 2010](#)) and Chimera molecular visualizers ([Pettersen et al, 2004](#)). Trajectories were analyzed using ptraj and VMD ([Humphrey et al, 1996](#)).

2.3 Results and discussion

2.3.1 Dehydron-residue conservation in proteins

Two possible manifestations of dehydrons have been reported– (a) the dehydrons engage in protein-protein interactions (PPI) as a way to circumvent local destabilization ([Fernandez & Scheraga, 2003](#)) thus leading to the proposition that dehydron-residues are conserved ([Fernandez, 2004a](#)); (b) dehydrons induce local unfolding and predispose the protein to aggregation ([Fernandez et al, 2003](#)).

Given that aggregation can be deleterious to the cell ([Chiti & Dobson, 2006](#); [Selkoe, 2003](#)), are residues that are dehydronic likely to be non-conserved? This motivated us to systematically examine a non-redundant set of proteins from the PDB to probe regions of proteins that are involved in PPIs and those that are involved in aggregation.

I define a protein residue as conserved or non-conserved based on its class average entropy value that is shown as the dotted line in [Figure 2.2A](#) (refer Experimental Procedures for details). The plot follows a bimodal distribution for the conservation of dehydron-residues, with a sharp peak of conserved dehydron-residues centered at ~0.2 and a broader peak of non-conserved dehydron-residues centred at ~1.9. This is in contrast to previous reports that the dehydron-residues tend to be evolutionarily conserved ([Fernandez, 2004a](#)).

To test the statistical significance of this distribution, all the entropy values are split into random non-overlapping subsets ([Figure 2.2B, 2.2C and 2.2D respectively](#)). Since the shape of the distribution is invariant across the subsets, it

appears that the distribution is statistically significant and perhaps biologically relevant.

Figure 2.2: (A) Distribution of dehydron-residue conservation for non-redundant set of PDB proteins. Dehydron-residue conservation for the 10 (B), 5 (C) and 20 (D) random non-overlapping subsets derived from the primary set. The dotted red line in (A), (B), (C) and (D) marks the average residue conservation for all residues across the dataset, discriminating conserved and non-conserved residues.

The peak of conserved dehydron-residues supports earlier observations made regarding the presence of dehydrons at PPI interfaces by Ariel Fernandez (Fernandez, 2004a; Fernandez & Crespo, 2008; Fernandez & Scheraga, 2003). However, the peak of non-conserved dehydron-residues has not been previously

reported and the engagement of such residues in aggregation is investigated in the following sections.

Theoretically, all proteins are capable of misfolding and aggregating under physiological conditions, depending on their innate (in)stability and the accessibility of locally unfolded states due to thermal fluctuations ([Thirumalai et al, 2012](#)). Therefore, the intention is not to classify proteins as aggregating and non-aggregating based on dehydron-residue conservation. Instead, this signature is used to investigate the regions in a protein structure that are vulnerable to undergo local unfolding and destabilization, thus signaling the possible onset of aggregation.

2.3.2 Dehydron-residue conservation at APRs

A statistical study of dehydron-residue conservation of PDB proteins outlined in the previous section allowed us to hypothesize that a considerable proportion of non-conserved dehydron-residues are implicated in aggregation. To test this hypothesis, the dehydron-residue conservation at APRs of aggregating proteins is studied for which experimental information on the location of APRs within the proteins is available. In addition, two proteins, Hemoglobin and the p53 DNA Binding Domain (DBD) are examined which are structurally and functionally well-annotated. These two proteins are chosen because they have been hypothesized to follow different mechanisms of aggregation. In the native state of Hemoglobin, the APR is exposed to solvent and the protein is not known to undergo major conformational changes in its transition to an aggregation

competent state ([Padlan & Love, 1985](#)). In contrast, in p53 DBD, the APR is largely buried in the native state and significant conformational changes are required to expose the APR in the aggregation competent state ([Reddy et al, 2009](#)). These two proteins were not part of the aggregation dataset used.

A literature survey revealed 27 aggregating proteins with distinct folds ([Table 2.1](#)) of which, the APRs of 15 proteins are known. The APRs are mapped on their structures to study the conservation of residues involving dehydrons that are in the vicinity of the APRs ([Figure 2.3](#)). A dehydron-residue is identified as being in the vicinity of an APR if it is within a 3 residue distance (sequence wise) or within 5 Å of the APR residues (structurally). This distance corresponds approximately to the volume of the desolvation domain as defined by Ariel Fernandez ([Fernandez, 2004a](#); [Fernandez & Berry, 2002](#); [Fernandez & Crespo, 2008](#); [Fernandez & Scheraga, 2003](#)). It is observed that, for most proteins, the dehydron-residues that are in the vicinity of APRs are predominantly non-conserved, with the exception of Transthyretin (TTR) where the conserved dehydron-residues are more in number. The possible reason for this anomaly may be the fact that part of the APR in TTR is buried and hence fibril formation would require partial unfolding and exposure of APR ([Reddy et al, 2009](#)). A limit of this approach is that if the APR is buried, a complete knowledge of all the dehydrons in the vicinity of APRs and those that may necessarily be relevant to the exposure, may not always be possible. For instance, Cystatin C, has been crystallized in 2 conformations Pdb ID 1cewI (chicken) and Pdb id 1g96A (human), and an examination of their dehydrons demonstrate this. The domain-swapped

conformation of human cystatin C has been suggested to be a conformation en-route to aggregation ([Janowski et al, 2001](#)). This is also the conformation that has more non-conserved dehydron-residues in the vicinity of the APR. In contrast, the conformation of chicken cystatin C does not show the presence of dehydrons in the vicinity of the APR. Similarly, Lactoferrin (PDB ID- 1f9bA) has no dehydrons in the vicinity of APRs, as they are partially buried in the protein core. Nevertheless, the results support the hypothesis that non-conserved dehydron-residues are indicative of APRs in proteins. And, when the APR is buried, these non-conserved dehydron-residues could signal regions where water attack could induce local destabilization and unfolding, which can in turn expose the APRs. This is further elaborated using the examples of hemoglobin and p53 DBD.

Figure 2.3: Proportion of conserved and non-conserved dehydron residues in the vicinity of APRs in some aggregating proteins whose structures are available. Green arrow highlights the example of Transthyretin which is an anomaly, in that, it has more number of conserved dehydron-residues than non-conserved dehydron-residues in the vicinity of its APR.

Case study1: Dehydron-residue conservation at APR in Hemoglobin

Hemoglobin is an oxygen transport protein that exists as a tetramer of 2 α and 2 β chains. A Glu6Val mutation in β -globin chain of hemoglobin (now referred to as Hemoglobin S) is known to cause aggregation in erythrocytes, resulting in a disorder known as sickle cell anaemia (Ingram, 1957).

I study the dehydron-residue conservation in the β -globin chain of Hemoglobin S (Harrington et al, 1997) (PDB ID 2HBS, resolution of 2.05 Å where the authors have identified the oligomerization interface (Figure 2.4A) and the

aggregation interface ([Figure 2.4B](#)). These interfaces are highlighted on to the hemoglobin β -chain ([Figure 2.4C](#)). It can be observed that the dehydron-residues A86, E90, D94 & K95 are mostly conserved (grey sphere representation) and present at the oligomerization interface, while dehydron-residues P5, S9, K65, G69, D73 & A76 are non-conserved (orange sphere representation) and occupy the aggregation interface. This contrasts with the observations of Fernandez who had reported that all the residues involved in dehydrons in this protein chain are conserved ([Fernandez, 2004a](#)) The possible reason for this new finding is the availability of new protein homologs from other organisms over the past decade which has led us to more accurately establish some of these residues as non-conserved.

Figure 2.4: Dehydran residues from oligomerization and aggregation interfaces of haemoglobin (PDB code 2HBS). (A) oligomerization interface, (B) aggregation interface, (C) β -chain: dehydran residues (in spheres) from (A) and (B) are highlighted in blue and red, respectively, and coloured according to their conservation score.

Case study2: Dehydron-residue conservation at APR in p53 DNA Binding Domain (DBD)

The DNA Binding Domain (DBD) of the tumor suppressor p53 binds DNA in a sequence specific manner leading to transcription. The DBD is inherently unstable, unfolding at temperatures slightly above physiological temperatures (~42-44°C), and is mutated in 90% of p53-mutant cancers (Joerger & Fersht, 2008). 30% of all cancerous mutations cause unfolding of the protein, leading to aggregation. This loss of p53 function has been shown to partly result from aggregation of a β -strand (residues 251-257) in the DBD (Ano Bom et al, 2012; Xu et al, 2012).

I associate most of the dehydrons formed between conserved residues with known binding events as shown earlier (Pietrosemoli et al., 2007) (Figure 2.5).

Figure 2.5: Dehydrons (blue lines) at the DNA binding and dimerization interfaces of p53 DBD (PDB code- 3IGK_A) are formed largely between conserved residues (grey). Dehydrons involving non-conserved residues (orange, within circles and residues are labelled) are found to be close to the ‘aggregation strand’.

I also find that the dehydrons at D259–N263 (S9-S10 turn) and at Y107–D148 are formed with non-conserved residues with no experimental evidence of a binding event. Among these, the former dehydron is in the immediate vicinity of strand S9 (residues 251-257) which is reported to be the aggregating strand ([Ano Bom et al, 2012](#); [Xu et al, 2012](#)) and the latter is in the neighbourhood of strand S3 (residues 141-146) which has minimal aggregation propensity (was predicted by ‘Tango’ at residues 144-148 ([Figure 2.6A](#)), other prediction methods did not show this). There is no experimental evidence of this strand’s involvement in aggregation. This occurrence of non-conserved dehydron-residues in the immediate vicinity of the aggregation strand &/or signal may suggest a role for local dynamics in the aggregation process. Crystal and NMR structures show unambiguously that the aggregation strand S9 involving residues 251-258 is only partly exposed to the solvent ([Figure 2.5](#)) and for this strand to nucleate aggregation, clearly dynamic changes of the surrounding residues must occur for this strand to get exposed ([Wang & Fersht, 2012](#)). A recent study ([Wang & Fersht, 2015](#)) used Φ -value analysis, which gauges the extent of local unfolding, to study the aggregation kinetics of a destabilizing mutant of p53. It reported a high value of Φ for the S9-S10 turn implying that this region undergoes major conformational changes in the transition. Moreover, for the dehydron (Y107–D148) involving non-conserved residues, a mutation D148V is identified, through exhaustive computational mutagenesis of all possible point mutations of the dehydron-residues viz. Y107 and D148, that drastically increases the aggregation propensity of strand S3 as predicted by Tango ([Figure 2.6A and 2.6B](#)). Thus, just as in hemoglobin, there

seems to be a strong correlation of non-conserved dehydron-residues and the aggregation propensity in p53 DBD too. This may have a bearing on the biological modulation of p53. p53 is an essential protein that is stable only under a small temperature range, and needs to be subject to rapid degradation. It is known that this regulation is carried out by the ligase MDM2 through the proteasomal pathway, which in turn may be facilitated by its property to readily unfold ([Brown et al, 2011a](#)). This property is further examined in the homologs p63 and p73.

Figure 2.6: Prediction of APR in p53 DBD, for (A) wild type (WT) and (B) D148V mutant, using the Tango algorithm. There is a significant change in the total aggregation propensity of p53 DBD on incorporation of a single mutation, D148V.

p63 and p73 are structural and functional homologs of p53, sharing ~60% sequence identity with p53. Both p63 and p73 have been shown to be more thermostable than p53 ([Brandt et al, 2012](#)). It is known that p53 DBD comprises of an APR that has been reported to nucleate in oncogenic structural mutants of the protein ([Xu et al, 2011](#)). Comparing the DBDs in these homologous proteins, it is observed that the S9-S10 turn is evolutionarily less conserved as compared to the S9 and S10 strands ([Figure 2.7A](#)). The backbone hydrogen bonds in the S9-S10

turn of DBD are prone to water attack in p53 while they are well wrapped in the case of p63 and p73, as suggested by the presence of dehydrons only in p53 (Figure 2.7B). The absence of these dehydrons in the more thermostable p63 and p73 suggests that the S9-S10 turn may modulate initial unfolding and aggregation events in p53.

Figure 2.7: Dehydrons in p53 and its homologs. (A) MSA of human p53, p63 and p73 DBDs in strand S9 (residues 251-257), S9-S10 (residues 258-263) and S10 turn (residues 264-274) is shown. (B) Dehydrons (shown with dashes on protein structures) in p53 (2AHI_A), p63 (3QYM_A) and p73 (3VD2_A) are shown. Yellow circles highlight the turn S9-S10, which is adjacent to aggregation strand S9, and are composed of relatively less conserved residues. Dehydrons from this turn are found only in p53 and they form between two non-conserved residues D259 and N263 (highlighted with arrows in (A)).

2.3.3 Dehydron analysis of MD simulations

The crystal structure study suggested that the non-conserved dehydron-residues could be used as markers to identify APRs in proteins. This property is now investigated in an ensemble of structures that have been generated using

molecular dynamics (MD) simulation. 7 aggregating proteins (PDB IDs 1a1nB, 1bebA, 1hkqA, 1ie4A, 1kcqA, 1wlaA, 3pf4A) with experimentally known APRs are simulated to test the robustness of using the presence of dehydron for the identification of APRs. Significant dehydrons are identified as those with at least 10% occupancy across well equilibrated trajectories (suggested by stable RMSD values) (Figure 2.8A). It is worth noting that the average number of dehydrons calculated per snapshot is more in number than in their respective static crystal structures (Figure 2.8B). Most of the major conformational changes across the trajectory arise from loop motions. It appears that backbone motions along with dynamic side-chains create new dehydrons.

Figure 2.8: (A) RMSD plot for 8 aggregating proteins over 50ns simulations. (B) Distribution of dehydrons calculated over ensemble structures from MD for the 8 aggregating proteins. The dotted line shows the number of dehydrons as calculated from their crystal structures.

Moreover, it is observed that most of these additional dehydrons identified from the simulated structures lay within the interaction sites and at APRs, which enables a comprehensive analysis. This data is supported using the MD simulations of p53 DBD (Figure 2.9A and 2.9B) and 7 aggregating proteins (Table 2.2).

Figure 2.9: MD simulation of p53 DBD (A) Distribution of number of dehydrons across simulation. The red dotted line shows the number of dehydrons identified in crystal structure. (B) p53 DNA Binding Domain showing APR residues in red and residues involved in dehydrons (>10% occupancy across simulation) in blue.

For these 7 proteins, the percentage of APRs identified by dehydrons increased significantly in 5 cases (highlighted as green in [Table 2.2](#) and [Figure 2.10](#)). For the remaining two proteins, no change is seen which could well reflect the limitations of the dehydron hypothesis, i.e. either necessarily being at sites of PPIs or that water interactions at the backbone alone lead to destabilization, opening the possibility of as yet undiscovered factors. For p53 DBD, the dehydrons cover the DNA binding interface and the dimerization interface more extensively than the one covered by dehydrons that are calculated from the crystal structures. However, for p53 DBD, the percentage of APRs covered by vicinal dehydrons, which are dehydrons within 5 Å of the APR residues or between residues that are not farther than 3 residues of the APR residue, remains unchanged. This is not surprising because the calculations were carried out on wild type p53 and normal MD simulations do not lead to unfolding events.

Protein name & PDB ID	Percent of APR covered by dehydrons (crystal)	Percent of APR covered by dehydrons (MD)
Beta2-Microglobulin (1a1nB)	64	100
Beta-lactoglobulin (1bebA)	21	21
RepA (1hkqA)	22	22
Transthyretin (1ie4A)	36	50
Gelsolin (1kcqA)	20	70
Myoglobin (1wlaA)	65	70
Cold shock protein CspB (3pf4A)	50	65

Table 2.2: Identification of APRs from MD trajectories of aggregating proteins using dehydrons. Table shows significant increase in the extent of APRs covered by dehydrons calculated in ensemble structures.

Figure 2.10: Percentage of APR covered by dehydrons in the crystal (blue) and MD (red). This figure conveys the same information as [Table 2.2](#). The figure however, makes it easier to observe that there is a significant increase in the extent of APRs covered by dehydrons calculated in ensemble structures of 5 of the 7 proteins studied.

I examine the persistence (occupancy) of dehydrons that were observed in the crystal structure of p53 DBD as well as the dehydrons that are formed during the MD simulations. The highly significant dehydrons in the p53 DBD are defined as those dehydrons which are present in more than 50% of the MD trajectory. I observe ([Figure 2.11](#)) that in a static crystal structure, I am able to identify up to 11 dehydrons which persist in the MD simulations for more than half of the simulation time. However, during the MD simulations I see the formation of an extra 10

persistent dehydrons. This suggests that MD simulations can be used to complement the dehydron picture that is deduced from the X-ray structure.

Figure 2.11: Figure shows frequency of highly significant dehydrons (occupancy in MD > 50%) in p53 DBD.

These highly significant dehydrons are present in regions with high fluctuations (Figure 2.12A). This is not surprising given that these regions are under conformational flux. It has been shown (Accordini et al, 2012) that drugs disrupting protein-protein interactions result in wrapping by the drugs or wrapping by intramolecular changes in the target protein. This behavior is also seen in the complexes of the p53 DBD with other macromolecules that are available in the PDB. These include complexes with the DBD monomer as part of the DBD dimer (PDB- 2AHI), DNA (PDB- 2AHI), BRCT domain of the apoptotic 53BP1 protein (PDB- 1GZH), nanobody Nb139 (PDB- 4QO1) and Bcl-xl (PDB- 2MEJ) (Figure 2.12 B, C, D, E, and F respectively). It is compelling that several dehydrons that

are seen in the MD simulations appear at or in the vicinity of the interfaces of these complexes.

Figure 2.12: (A) Mode of interaction of p53 DBD (PDB Id- 2AHI chain A, colored according to the fluctuations of the residues from MD simulations of the monomer (low fluctuation in blue to high fluctuations in red) shown in complex with (B) a 10-mer double stranded piece of DNA, (C) DBD monomer as part of the DBD dimer, (D) molecule of the BRCT domain of the apoptotic 53BP1 protein, (E) nanobody Nb139 and (F) molecule of Bcl-xl. The C α atoms of residues that engage in highly significant dehydrons (>50% occupancy in the MD simulations) are mapped onto the crystal structure of p53 DBD. Red spheres represent dehydron-residues that are identified from the crystal structure while green spheres represent dehydron-residues that are additionally observed in the MD.

A comparison of these structures with uncomplexed monomeric p53 DBD taken from the crystal structure 2XWR.pdb show that the interfacial sidechains in p53 DBD undergo conformational changes of the order of $\sim 1\text{-}2\text{ \AA}$. Hence it is clear that dehydrons seen in MD simulations do seem to appear as a result of conformational changes that occur in the p53 DBD in its monomeric state in solution, and are seen clustered at sites of intermolecular interactions. Combining these observations, it can then be hypothesized that the dehydrons on the p53 DBD surface (Figure 2.13) that are not at the known interaction sites, are likely to be sites of interactions with other macromolecules; I wait with optimism for these sites to be validated as interaction sites.

Figure 2.13: Binding interfaces of the p53 DBD with the 5 molecules shown in Figure 9 are mapped onto the surface representation of the p53 DBD. The C α atoms of residues that engage in highly significant dehydrons ($>50\%$ occupancy in MD) are mapped onto the crystal structure of p53 DBD and are colored using the same convention as in figure 2.12. The regions containing dehydrons that are not overlapping with the interfaces are hypothesized to be the likely interaction sites for other molecules.

Studying dehydron-residue conservation in the vicinity of APRs in aggregating proteins has revealed that non-conserved dehydron-residues do appear

to modulate the process of aggregation. Next I use the dehydron-residue conservation feature as a tool for APR identification and compare the results with other known APR predictors.

2.3.4 Comparison with sequence and structure based APR predictors

I next evaluate the dehydron-residue conservation feature developed in this thesis with the performance of available in silico predictors by computing the percentage of APRs that are correctly identified by each predictor (true positives). Among the available predictors I use Tango (Fernandez-Escamilla et al, 2004), Waltz (Maurer-Stroh et al, 2010), PASTA (Walsh et al, 2014), ZipperDB (Goldschmidt et al, 2010) and AmlyPred 2 (Tsolis et al, 2013) and a structure based APR, Spatial Aggregation Propensity (SAP) (Chennamsetty et al, 2009). Data on known APRs was taken from (Tsolis et al., 2013). The comparisons are shown in (Table 2.3).

Protein name and PDB ID	Percent of APR residues correctly predicted						Percent of non- conserved dehydron-residues in the vicinity of APR)
	Tango (value >10)	Waltz	Pasta	Amylpred	ZipperDB	SAP (if SAP of residue < -0.25)	Percent of APR covered by dehydrons
Beta2-Microglobulin (1a1nB)	17	24	0	56	45	11	64
Beta-lactoglobulin (1bebA)	21	16	26	73	26	4	21
Amphoterin (1cktA)	37	43	0	43	62	1	25
Lactoferrin (1f9bA)	50	75	50	25	75	0	0
Cystatin C (1g96A)	0	100	0	100	100	2	100
RepA (1hkqA)	100	55	0	100	88	2	22
Transthyretin (1ie4A)	27	0	0	63	59	7	36
Gelsolin (1kcqA)	0	10	0	12	34	6	20
Prolactin (1rw5A)	11	0	28	27	58	12	60
Myoglobin (1wlaA)	0	32	49	55	72	5	65
Kerato-epithelin (1x3bA)	44	0	0	44	88	3	100
Cold shock protein CspB (3pf4A)	0	33	15	19	54	11	50

Table 2.3: Comparison of methods that predict APRs. Table shows comparison of APRs identified using dehydron-residue conservation with those identified using some of the available APR predictors.

Figure 2.14: Percentage of APRs residues correctly predicted by different APR predictors (blue) and using dehydrons (red). Figure represents the same information as [Table 2.3](#).

Tango predicts the propensity of a region to form APRs based on the probability of that region for adopting a buried beta strand which can engage in the formation of beta-sheet aggregates. Waltz compares the input sequence with precomputed Position Specific Scoring Matrices to determine amyloid forming stretches. PASTA 2.0 relies on the computation of the stability and propensity of segments in the input protein to form amyloid cross-beta structures. The energy function is based on a pairwise interaction potential of amino acids. This, in turn, is derived from the observed propensities of two residues lying side-by-side in a beta-sheet structure in globular proteins. ZipperDB splits the input protein sequence into overlapping six residue segments and then fits it onto the known fibril forming NNQQNY peptide (from the sup35 prion protein) structure. It then evaluates the

propensity of this segment to form amyloids based on energies derived from the Rosetta Design program. AmylPred 2 is a consensus APR predictor that utilizes results from multiple independent predictors; it was shown to perform better than the individual predictors. SAP identifies APRs from an ensemble of protein structures obtained from MD. Originally developed for APR prediction on antibodies, it was further extrapolated for APR identification in other proteins as well and was demonstrated with the example of hemoglobin. However, the performance of this method was not calibrated against other known APR predictors.

Percentage of APRs covered by dehydrons is calculated by considering all the APR residues that lie within the dehydrons in the vicinity, as shown in [Figure 2.15](#). Overall, the method performs reasonably well. For some proteins, it performs better than other methods (highlighted as green boxes in [Table 2.3](#) and [Figure 2.14](#)). The method does not perform well for the following proteins: Lactoferrin, Beta-lactoglobulin, Amphoterin and RepA. There could be several reasons for this. As already discussed, for Lactoferrin, most of the APR is buried in the core of the protein native structure and hence dehydrons are not identified in the vicinity of its APR. In such cases, the proteins need to undergo a conformational change in order for the APR to expose itself to the solvent. The current method relies on the correct identification of dehydrons and hence its validity is limited by the availability of high resolution structural information. Moreover, it should be noted that there were significant number of false positives (incorrectly identified APRs) i.e. the method seems to over predict APRs. This could be due to the same cut off used to mark

protein residues as conserved and non-conserved; and since every protein family is under different evolutionary pressures to evolve, this cut off could be tuned by taking into consideration the evolutionary rate of the protein system under study. Variations of this cut off value of average entropy for p53 DBD, however, suggested that an entropy value of 1 seems to optimally distinguish PPI regions (as conserved dehydron-residues) and APRs (as non-conserved dehydron-residues).

Figure 2.15: Illustration of how the APR covered by dehydrons is calculated. Part of the Cystatin C structure (PDB ID 1G96 chain A) shown in tube-ribbon representation and colored by evolutionary conservation (grey is conserved and orange is non-conserved). Dehydrons are shown as blue lines. Experimentally known APR is shown in rectangle-ribbon representation. APR predicted using dehydron-residues is highlighted in green.

SAP is a structure based APR predictor, so I compare its results for hemoglobin and p53 DBD with the dehydron-residue conservation method. Below is the mathematical formulation of SAP APR predictor ([eqn. 2.3](#)). SAA refers to the Solvent Accessible Area. ‘A’ and ‘B’ in the [eqn. 2.3](#) are SASA of the side chain atoms within radius R and SASA of side chain atoms of fully exposed residue respectively.

$$SAP_{atom\ i} = \frac{1}{N} \sum_N \sum_{\substack{\text{Residues with at least} \\ \text{one side chain atom within } R}} \left(\frac{A}{B} \right) \quad \dots (2.3)$$

* **Residue Hydrophobicity**)

It is observed that SAP identifies only a part of the hemoglobin APR in helix H1 (Figure 2.16A) and fails to identify residues in helix H3, which also, upon structural visualization, appear to be at the aggregation interface. SAP does not identify the aggregation strand S9 as an APR in p53 DBD (Figure 2.16B), which is probably due to the fact that this strand is largely buried. Thus, dehydron-residue conservation can complement the existing strategies aimed at APR identification.

Figure 2.16: Spatial Aggregation Propensity (SAP) calculated for Hemoglobin and p53 DBD (A) Hemoglobin β -chain colored according to SAP scale, from red (high aggregation propensity) to blue (low aggregation propensity). APRs are shown in spheres. SAP calculation correctly predicts residues in helix H1 as APRs but not in helix H3. (B) p53 DBD colored according to SAP scale, from red (high aggregation propensity) to blue (low aggregation propensity). APRs are shown in spheres.

The dehydron-residue conservation feature was primarily tested for its occurrence in the vicinity of APRs in order to examine its link with aggregation associated regions in proteins. The results additionally suggest that this feature can also complement the existing APR prediction methods.

2.4 Conclusion

I report a statistical study of ‘dehydrons’ in proteins. The primary observation is that the dehydrons occur at highly conserved and non-conserved residues. This is in contrast to the previous understanding that dehydrons, being an indicator of PPI, exist exclusively between conserved residues (Fernandez, 2004a). Since enhanced backbone hydration is also believed to play an important role in destabilizing protein folds and subsequent aggregation (Fernandez et al, 2003; Thirumalai et al, 2012), the residue conservation at such regions is expected to be low. Hence, I systematically explore the role of non-conserved dehydron-residues in aggregation.

One way to test this hypothesis is to examine the conservation of dehydron-residues at the known Aggregation Prone Regions (APRs) in proteins. The results show that the residues with hydrated backbones near the vicinity of APRs in proteins are, to a major extent, fairly non-conserved. And so, it is demonstrated that complementing hydration data with the evolutionary conservation information, allows us to distinguish dehydrons which are involved in normal cellular function (binding sites as conserved hydrated residues) from those that are associated with

abnormalities like aggregation (aggregation associated sites as non-conserved hydrated residues). I, then, compare this APR predictor with other existing sequence and structure based APR predictors and showed that my method can perform better in some cases and can be used to complement other existing strategies that are designed to predict APRs in proteins. Earlier, Fernandez had hypothesized that proteins with localized concentrations of dehydrons are vulnerable to aggregation ([Fernandez et al, 2003](#)). Thus, though the extent of wrapping in a localized region is crucial, my analysis of aggregating proteins suggests that the nature (evolutionary conservation) of the dehydron-residues also seems to be distinct (less conserved). Finally, I try to rationalize the co-occurrence of aggregation associated regions in proteins with the presence of non-conserved dehydron-residues. It is likely that water attack at the backbone atoms of such residues predisposes the protein to local destabilization followed by unfolding and exposure of APRs, which culminates in aggregation. Thus, these residues must be under less evolutionary pressure to be conserved and, suggests that in the course of evolution, these residues would likely undergo mutations to residues with increased numbers of wrappers. I would like to state that, though non-conserved dehydron-residues could be suggestive of APRs (as shown here), there will be as yet undetermined factors that will undoubtedly contribute to the process of aggregation.

We further rationalize the mechanism underlying the dehydron-residue conservation feature introduced in the thesis. Residues and hence dehydrons that are present at interfaces that mediate protein-protein interactions (PPI) are

conserved because they are involved in protein-protein interactions (PPIs). These interactions form the nodes of essential pathways/networks that transmit diverse signals important for cell survival and health. These interactions combine specific electrostatic and non-polar interactions. In particular, the electrostatic interactions demand specific arrangements and interactions of atoms that ensure specificity and hence are under great evolutionary pressure to be conserved in order to ensure the proper transmission of signals and hence functioning of cells. It is this requirement that places constraints on the chemical environment, including specific geometries that enable salt bridges and hydrogen bonds involving sidechains, which characterizes the PPI interfaces, and hence establishes the basis for these dehydrons to be conserved. In contrast, the non-conserved dehydrons are found at protein aggregation sites because these regions are characterized mostly by hydrophobic amino acids. Protein aggregation is largely driven by non-specific hydrophobic interactions and either involve sidechain interactions (forming the so-called amorphous aggregates) or backbone-backbone hydrogen bonds between the β -strands (forming the so-called amyloid aggregates). In both these cases, there is no requirement of any specificity as hydrophobic interactions are mostly non-directional (in contrast to specific geometries required for electrostatic interactions seen at PPI interfaces) and this likely accounts for the observation that there is not much conservation of dehydrons required at aggregation interfaces; for the case of amyloid aggregates, where the assemblies are mediated through backbone hydrogen bonds, clearly there is not a very stringent need for sidechain conservation anyway and hence we do not observe many conserved dehydrons.

This is also apparent in the observation that there are larger numbers of dehydrons in the conserved regions. This is why adding dehydron counting to the other methods of predicting PPI and aggregation sites is a good strategy to improve the outcome of the predictions

It is now well accepted that aggregation is not a rare phenomenon ([Cohen et al, 2012](#)). Even under equilibrium conditions, many more proteins aggregate than are currently known and some of these have associated pathologies ([Chiti & Dobson, 2006](#)). Hence, one approach to prevent aggregation could be to understand and modulate the dynamics of such dehydron containing sites on proteins. This makes understanding regions on proteins associated with aggregation crucial in developing strategies to prevent this fate. Thus, this signature can provide clues to artificially tweaking the aggregation propensity. This method, to the best of my knowledge, is the only one which combines protein sequence and structure data to predict APRs and has promise of being a useful addition to the toolset used to understand the complexity of biomolecules.

Preface to Chapter 3

In next chapter, extensive Molecular Dynamics simulations are performed to investigate the role of non-conserved dehydron-residues in p53 DBD, which are present in the vicinity of the APR (residues 251-257), in unfolding and destabilization of the domain (Figure 2.5, 2.6 and 2.7). Probing the role of these yet unreported non-conserved dehydron-residues in the onset of destabilization in the domain serves as a case study for investigating the involvement of non-conserved dehydron-residues in protein aggregation. A significant number of oncogenic mutants of p53 are reported to undergo loss of function through unfolding and aggregation. A molecular mechanism of destabilization, if established, can provide potential strategies to rescue such mutants.

Chapter 3: A combinatorial approach to probing the dynamics of aggregating mutants of p53 DNA Binding Domain reveal a novel ‘druggable’ pocket

The DNA Binding Domain (DBD) of the tumor suppressor p53 is the site of several “hot-spot” oncogenic mutations. A subset of these mutations is known to lower the melting temperature of the DBD that results in destabilization and unfolding. Unfolding leads to the exposure of a hydrophobic β -strand and nucleates aggregation. Subsequently, this results in cancer pathologies through loss of function, dominant negative and gain of function effects. Driven by the idea that knowledge about initial conformational changes causing unfolding in DBD is likely to suggest inhibitory strategies, in this chapter, long (totaling $\sim 30 \mu\text{s}$), all atom explicit solvent MD simulations of the wild type and numerous aggregating hot spot mutants of p53 DBD are performed. Dynamics reveal differential conformational sampling of a turn region (S6-S7) between the wild and mutants; this turn is contiguous to the known aggregation-prone strand, between the wild type and mutants. A network of interacting residues on the DBD that regulate the conformational state of this S6-S7 turn is identified, leading to the identification of potential structural and dynamical features that could destabilize the DBD and

mark the onset of aggregation in p53. Within these conformational changes in the DBD as seen in the simulations, a “druggable” cavity is identified.

3.1 Introduction

Cellular trauma triggers the stabilization of the transcription factor/tumor suppressor p53, which in turn, regulates the level of downstream genes involved in apoptosis, DNA repair, senescence and metabolism. Any perturbation in the p53 gene network causes abnormalities/disorders in these cellular pathways ([Vogelstein et al, 2000](#)). The gene encoding p53, TP53, is mutated in 50-70% cases of cancers ([Schuijjer & Berns, 2003](#)), thus making it a potential therapeutic target, with some molecules that modulate this pathway already in clinical trials ([Brown et al, 2011a](#); [Brown et al, 2009a](#)).

The IARC TP53 Mutation Database ([Olivier et al, 2002](#)) lists ~95% of these oncogenic mutations to be in the DNA binding domain (DBD) (alternatively referred to as core domain or p53C). The p53 DBD is structurally composed of a β -sandwich scaffold and a DNA-binding surface, including a loop-sheet-helix motif (LSH) and two loops (L2 and L3) tethered by a single Zinc atom. ([Figure 3.1](#)) and binds DNA promoters in a sequence specific manner, initiating transcription. p53 DBD oncogenic mutants have been classified into 2 groups based on their mode of action: the DNA contact mutants and the destabilizing (also referred as structural or aggregating) mutants. The DNA contact mutants lose their ability to interact with specific DNA promoters (*loss of function*) and sustain local structural changes

with negligible perturbations in the thermodynamic stability of the DBD (Joerger & Fersht, 2008). Some contact mutants oligomerize with wild type (henceforth to be referred to as WT in this chapter and the next chapter) p53 (when the mutation is in one allele) and inactivate the WT allele and also bind to the paralogs p63 and p73, exhibiting ‘*dominant negative*’ and ‘*gain of function*’ effects respectively (Brosh & Rotter, 2009) (Joerger et al, 2009). In contrast, the destabilizing mutants, which constitute 30% of all the oncogenic mutations in p53, are thermodynamically unstable at physiological temperature and undergo partial unfolding and/or aggregation. Such mutants also trigger the unfolding and co-aggregation of WT p53 (*dominant negative effect*) and the paralogs p63 and p73 (*gain of function effect*) (Xu et al, 2011). This weakens the apoptotic propensity and expression of p63 and p73 transcribed genes in cells (Gaiddon et al, 2001). Destabilizing mutants can be partly/fully functional at sub-physiological temperatures (Bullock & Fersht, 2001; Bullock et al, 1997; Bullock et al, 2000). For instance, the V143A mutant of DBD aggregates at physiological temperatures, but binds DNA non-specifically at sub-physiological temperatures. The aggregation kinetics of a destabilizing mutant Y220C (Wang & Fersht, 2012; Wilcken et al, 2012) suggests that the aggregation competent state results from a slow rate-determining first step followed by rapid polymerization of the unfolded DBD (Wilcken et al, 2012). Indeed, even the WT p53 DBD is inherently unstable, unfolds at temperatures slightly above physiological temperatures (~44°C) and has a low kinetic stability with a half-life of just ~9 minutes at physiological temperature (Ang et al, 2006; Bullock et al, 1997; Friedler et al, 2003).

Computational and experimental studies suggest that the strand S9 (residues 251-257) (Figure 3.1) of the DBD is involved in the nucleation process (Ghosh et al, 2014; Soragni et al, 2016; Xu et al, 2011), however a detailed molecular picture of this process remains undescribed.

Figure 3.1: DNA Binding Domain (DBD) with secondary structures labeled (PDB ID- 2XWR_A). Green sphere represents Zn and the residues (Cys176(SG), Cys238(SG), Cys242(SG) and His 179(ND1)) coordinated to Zn are shown in sticks. The aggregation prone region in p53 DBD is present in strand S9 and is colored in magenta. The N and C termini of the domain are labeled.

Elegant genetic experiments have identified several second site mutations that could revert the effects of destabilizing mutations and restabilize mutant p53 (Brachmann et al, 1998; Joerger et al, 2004; Joerger et al, 2005; Merabet et al, 2010; Nikolova et al, 2000; Otsuka et al, 2007). These also suggested that ligands could potentially be identified to stabilize mutant p53 into a WT conformation (Brown et al, 2011a; Brown et al, 2009a; Foster et al, 1999; Selivanova & Wiman, 2007; Wiman, 2010). PRIMA-1, a compound already in clinical trials, is thought to stabilize p53 by alkylating cysteine residues near the L1/S3 region of the DBD (Bykov et al, 2002a; Bykov et al, 2002b; Wassman et al, 2013). The same site was also shown to accommodate Stictic acid, which was shown to be more efficient than PRIMA-1 at thermally stabilizing p53 (Wassman et al, 2013). For another destabilizing mutant of p53 DBD, Y220C, a druggable pocket at the mutation site was targeted by small molecules, but these resulted only in marginal stabilization (Boeckler et al, 2008; Joerger et al, 2015; Liu et al, 2013; Wang & Fersht, 2015). In another approach, the DNA binding surface of DBD has been targeted in mutant p53 with the idea that the peptides (derived from a p53 binding protein- 53BP2 as it is called) will act as chaperones, stabilizing mutant p53 and these ligands would then be competed out in the presence of natural binding partners of p53 (Friedler et al, 2002; Issaeva et al, 2003). Additionally, some mutants are thought to unfold because they lose zinc (Meplan et al, 2000). ZMC1, a small molecule, acts as a zinc-metallochaperone that is thought to stabilize p53 by increasing the intracellular concentration of zinc, thereby enhancing zinc binding to DBD (Blanden et al, 2015; Yu et al, 2014). Recently, in a very promising study, a

peptide was designed to block aggregation by targeting the APR (residues 251-257) and shown to rescue the function of p53 in R248 and R175 mutants of p53 ([Soragni et al, 2016](#)). Regardless, there is a need to identify new ways to rescue the function of p53. Although a ‘global suppressor motif’ made up of N235, N239 and S240 has been proposed ([Baroni et al, 2004](#)), it was found to rescue only half of the most common p53 oncogenic mutations. This suggests that there likely are different pathways of destabilization of DBD. Hence, it is of interest to see if there is a common initial event that triggers the unfolding of the DBD and if so, can that be prevented.

With this aim, I study the dynamics, using atomistic computer simulations, of WT and several destabilizing mutants of p53 DBD. Studying the dynamics of the full-length p53 is infeasible because the atomic structure of the full-length protein is not available. However, the DNA Binding Domain (DBD) has been used as a reasonable model to recapitulate the properties (at least DNA binding) of full-length p53 ([Lukman et al, 2013](#)). Although the native functional state of p53 is known to be tetrameric, monomers and dimers have been observed in vivo ([Rajagopalan et al, 2011](#)) which exhibit very similar thermodynamic stabilities ([Lubin et al, 2010](#)). Structural features of DNA-binding have been shown to be independent of its oligomeric state ([Ho et al, 2006](#); [Kitayner et al, 2006](#); [Wang et al, 2007](#)). Structurally destabilized oncogenic mutants of p53 aggregate through an aggregation prone region (APR) that is located in the S9 strand of the DBD (residues 251-257) ([Ghosh et al, 2014](#); [Xu et al, 2011](#)). Additionally, the initiation of aggregation of mutant p53 is a unimolecular process, dependent on the

concentration of DBD alone, and invariable with respect to the full length or the oligomeric state of p53 (Wang & Fersht, 2012). Together, these make monomeric DBD a good model to understand the mechanism of aggregation of p53.

I perform extensive molecular dynamics (MD) simulations of the WT and the temperature sensitive V143A mutant of p53 DBD; the mutant is thermodynamically less stable than the WT by about ~3.7 kcal/mol (Joerger et al, 2006). This enables us to study the stability of V143A DBD at different temperatures including sub-physiological (298K), physiological (310K) and above-physiological (323K) temperatures. It is observed that the S6-S7 turn in DBD samples two different conformations, which seem to be associated with the thermodynamic stability of the protein at a particular temperature. The observations on the sampling of the S6-S7 turn are substantiated with several simulations (Table 3.1), including four other destabilizing mutants which are known to aggregate through strand S9 (Xu et al, 2011), these being E258V, R110L, R175H and R248Q. Subsequently, a common mechanism for the loss of function of a structurally destabilizing mutant form of the protein is hypothesized. Multiple simulations of p53 DBD variants and appropriate benchmarks of the observations against the available experimental data make this hypothesis quite robust. Finally, I identify a druggable pocket in the S6-S7 region of DBD and propose a novel strategy to rescue destabilizing mutants of p53 DBD using small molecules/peptides.

Sr no.	MD Simulations		Purpose	Temperature	Simulation length	Total simulation time
	Protein	Type				
1	p53 DBD	Wild (WT)	Preliminary simulations show differential conformational sampling of S6-S7 turn that seem to relate to the thermodynamic stability of DBD	298 K	500ns x 3	1.5 μ s
2	p53 DBD	WT		310 K	500ns x 3	1.5 μ s
3	p53 DBD	WT		323 K	500ns x 3	1.5 μ s
4	p53 DBD	V143A		298 K	500ns x 3	1.5 μ s
5	p53 DBD	V143A		310 K	500ns x 3	1.5 μ s
6	p53 DBD	V143A		323 K	500ns x 3	1.5 μ s
7	p53 DBD	E258V	<u>Hypothesis test 1</u> : Do the destabilizing mutants with experimentally identified Aggregation Prone Region (APR) conform to the above observed S6-S7 turn conformation for V143A?	310 K	500ns x 1	500 ns
8	p53 DBD	R110L		310 K	500ns x 1	500 ns
9	p53 DBD	R175H		310 K	500ns x 1	500 ns
10	p53 DBD	R248Q		310 K	500ns x 1	500 ns
11	p53 DBD	A143V (Back mutate V143A mutant at end of simulation to WT)	<u>Hypothesis test 2</u> : Does the mutant DBD revert to WT like conformation of the S6-S7 turn?	310 K	600ns x 1	600 ns
12	p53 DBD	V143A-N239Y	<u>Hypothesis test 3</u> : Do the secondary site suppressor mutations of	310 K	300ns x 1	300 ns
13	p53 DBD	V143A-N268D		310 K	300ns x 1	300 ns

14	p53 DBD	V143A-S240N	V143A mutant of p53 DBD sample a WT like conformation of S6-S7 turn?	310 K	300ns x 1	300 ns
15	p63 DBD	WT	<u>Hypothesis test 4</u> : Do the thermodynamically stable paralogs of p53 DBD sample WT like conformations of S6-S7 turn?	310 K	500ns x 1	500 ns
16	p73 DBD	WT		310 K	500ns x 1	500 ns
17	p53 DBD	WT (Umbrella sampling simulation)	What energy landscape characterizes the conformations of S6-S7 turn?	310 K	50ns/window	1150 ns
18	p53 DBD	V143A (Umbrella sampling simulation)		310 K	50ns/window	1150ns
19	p53 DBD	WT (Benzene mapping simulations)	<u>Possible biological implications of the conformation of S6-S7 turn</u> : Quantify the exposure of aggregation strand S9	310 K	20ns x 10	200 ns
20	p53 DBD	V143A (Benzene mapping simulations)		310 K	20ns x 10	200 ns
21	p53 DBD	E258V (Benzene mapping simulations)		310 K	20ns x 10	200 ns
22	p53 DBD	R110L (Benzene mapping simulations)		310 K	20ns x 10	200 ns
23	p53 DBD	R175H (Benzene mapping simulations)		310 K	20ns x 10	200 ns
24	p53 DBD	R248Q (Benzene mapping simulations)		310 K	20ns x 10	200 ns

25	p53 DBD	WT	p53 DBD with extended N-terminus (residues 91-293) is simulated to quantify its role in domain stability	310 K	500ns x 3	1.5 μ s
26	p53 DBD	R175H		310 K	500ns x 3	1.5 μ s
27	p53 DBD	R248Q		310 K	500ns x 3	1.5 μ s
28	p53 DBD	R174A-E180A (network residue mutation 1)	Test the hypothesis that the identified subset of interacting residues regulate the conformation of S6-S7 turn	310 K	300ns x 1	300 ns
29	p53 DBD	R174A (network residue mutation 2)		310 K	300ns x 1	300 ns
30	p53 DBD	R174A (network residue mutation 3)		310 K	300ns x 1	300 ns
31	p53 DBD	V143A-D207A	<u>Proposing D207A as a global secondary site suppressor mutation for the destabilizing mutants of p53</u> : Do these double mutants sample a WT like conformation of S6-S7 turn?	310 K	$\sim 1 \mu$ s x 3	3 μ s
32	p53 DBD	E258V-D207A		310 K	$\sim 1 \mu$ s x 1	1 μ s
33	p53 DBD	R110L-D207A		310 K	$\sim 1 \mu$ s x 1	1 μ s
34	p53 DBD	R175H-D207A		310 K	$\sim 1 \mu$ s x 1	1 μ s
35	p53 DBD	R248Q-D207A		310 K	$\sim 1 \mu$ s x 1	1 μ s
36	p53 DBD	Ligand docked onto V143A mutant of DBD	Examine the druggability of S6-S7 pocket using Dodecane-trimethylamine that is identified as a ligand from	310 K	100ns x 5	500 ns

			structure based pocket search			
--	--	--	-------------------------------------	--	--	--

Table 3.1: Number and types of the various MD simulations of p53 DBD and mutants and the paralogs p63 and p73 that are carried out in this study.

3.2 Methods

3.2.1 MD Simulation

Monomers of the WT and mutant p53 DNA Binding Domain (DBD) (residues 94-293) were simulated using the amber ff99SB force field ([Hornak et al, 2006](#)). The crystal structure of WT DBD (PDB ID 2AHI ([Kitayner et al, 2006](#)), chain A, 1.85 Å resolution) was used for the modeling studies. This chain has only one missing residue (S94) and only 3 residues (S95, K292, and G293) have one missing atom each. The missing coordinates of the atoms/residues of S94, S95 and G293 were grafted from PDB ID 3IGL ([Kitayner et al, 2010](#)) (1.8 Å resolution) and for K292 were taken from 1GZH chain A ([Derbyshire et al, 2002](#)) (2.6 Å). The Modloop program ([Fiser & Sali, 2003](#)) was used to refine the coordinates of these missing atoms/residues. For creating the mutant structures, the mutation was introduced in the WT DBD using the program PyMOL ([Schrodinger, 2010](#)) and the rotamers with minimal clashes were chosen.

PDB2PQR ([Dolinsky et al, 2007; Dolinsky et al, 2004](#)), MOLPROBITY ([Chen et al, 2010](#)) and WHATIF ([Vriend, 1990](#)) were used to generate the likely protonation states of amino acids. The final protonation states (Histidines with protonation on ND1 were residues 178, 214, 233, while those with protonation on NE2 were residues 115, 168, 179, 193) were chosen after visual inspection. ACE

and NME residues were used to cap the N-terminal and C-terminal residues respectively. Zinc is coordinated to 4 residues in the L2 and L3 loop, including C176(SG), C238(SG), C242(SG) and H179(ND1). The bonded model of Zinc ([Peters et al, 2010](#)) is used to covalently bond Zinc to these 4 residues as it preserves the tetrahedral coordination which has been reported to be crucial for the L2-L3 stability and also for the DBD interactions with DNA ([Bullock et al, 2000](#)). The structures were solvated in cubic boxes of TIP3P water molecules ([Jorgensen et al, 1983a](#)). The minimum distance between any protein atom and the edge of the water box was 12 Å. The systems were then neutralized by the addition of appropriate numbers of Na⁺ and Cl⁻ ions.

The SANDER module of Amber 12 is used (and later switched to Amber 14 for speed) to perform minimization and molecular dynamics (MD) simulation. Minimization was carried out using 1000 steps each of steepest descent followed by conjugate gradient methods. Subsequently, the systems were gradually heated to 298K, 310K & 323K. These temperatures were chosen because V143A is a temperature sensitive mutant; it is known to unfold and aggregate at physiological and higher temperatures, while known to retain a native fold that is capable of binding DNA non-specifically at sub-physiological temperatures. This was followed by a 2ns NPT equilibration and a subsequent production run in the NPT ensemble for 500 ns using a GPU implementation of PMEMD ([Gotz et al, 2012](#)). The SHAKE algorithm ([Ryckaert et al, 1977a](#)) was used to constrain bonds involving hydrogen atoms. Particle Mesh Ewald method ([Darden et al, 1993b](#)) has been used to calculate long range electrostatic interactions beyond a distance of

8Å. The Langevin thermostat is used to regulate the temperature ([Loncharich et al, 1992](#)). Temperature across the simulations fluctuates around the average (preset/target temperature) with a standard deviation of ~3K (data not shown). A weak coupling with a relaxation time of 1ps is used to maintain constant pressure (1.0 bar) in the simulations. An integration time step of 2 fs is applied; coordinates were saved at an interval of 50 ps.

Several variants of the DBD simulations (details in [Table 3.1](#)) are run for the purpose of this study. The crystal structures were used as the starting structures for the simulation of thermodynamically stable paralogs of p53 i.e. p63 (PDB ID 3QYM chain A, 3.2 Å resolution, ([Chen et al, 2011](#))) and p73 (PDB ID 3VD2 chain A, 4 Å resolution, ([Ethayathulla et al, 2012](#))). The goal, here, is not to investigate global unfolding but local conformational changes that could potentially destabilize the DBD and trigger subsequent unfolding.

3.2.2 Umbrella sampling simulations

The free energy associated with the conformational state of the S6-S7 turn of p53 DBD was computed using Umbrella sampling and weighted histogram analysis method (WHAM) ([Kumar, 1995](#)). The distance measure used to characterize the conformational state of S6-S7 turn (distance between CZ atom of R209 and backbone carbonyl O atom of D259) is used as a reaction coordinate. A biasing potential of 1 kcal/mol is then introduced along this reaction coordinate to considerably enhance the sampling of the phase space. A distance cut off of 1 Å and 40 Å defines the ends of the potential well. The distance is sampled using a flat bottom potential with equilibrium values ranging from 6 Å to 28 Å at intervals of 1

Å (thus resulting in 23 windows). Each window of the umbrella sampling MD is run for a minimum 50 ns. The distance distributions are used to derive the Potential of Mean Force (PMF) using Alan Grossfield's implementation of WHAM ([Grossfield, 2.0.7](#)).

3.2.3 Benzene mapping simulations

The difference in the exposure of APR, as measured using SASA, between the simulations of WT and the destabilizing mutants of p53 DBD appeared to be quite small and it was difficult to discriminate between them. This could well arise from the timescales of the current simulations being too short. Thus, fragment mapping simulations ([Foster et al, 2012](#); [Guvench & MacKerell, 2009](#); [Raman et al, 2011](#)) was used, which are a type of non-canonical MD where a protein is immersed in a water box at a defined concentration of fragment(s). These have been used to enhance local conformational sampling of the protein. Using benzenes as a probe in fragment mapping simulations have been used to map transiently exposed hydrophobic patches on the protein surface, including cryptic binding pockets ([Lama et al, 2015](#); [Tan et al, 2012](#); [Tan et al, 2015](#)). In order to quantify the exposure of the aggregation prone region that is largely hydrophobic (sequence- ILTIITL), benzene mapping simulations for the WT and five destabilizing mutants of p53 DBD including V143A, E258V, R110L, R175H and R248Q is carried out. A 0.2 M concentration of benzene was used and 10 simulations with random orientations of benzenes around the DBD were carried out for each mutant. Each system was allowed to evolve for 20 ns. Thus, the total simulation time was 200 ns each for the WT and destabilizing mutants of DBD.

Benzene density was contoured at two times the density at which no benzenes are visible in the bulk solvent.

3.2.4 MUTINF- Divergence calculations

To investigate differential conformational sampling between MD simulations of the WT and destabilizing mutants of p53 DBD, the MUTINF analysis package that computes differential conformational sampling (using Kullback-Leibler Divergence or KL-div) is used. KL-div ([McClendon et al, 2012](#)) outputs the differential sampling of internal coordinates of residues between two simulations and is useful for identifying residues that sample different orientations between the WT and the destabilizing mutants. The program splits a trajectory into a user defined number of windows (here 5 windows are used) over which the distribution of internal coordinates is considered for computing KL-div. Splitting the trajectory into multiple small windows eliminates the weak correlations (not biologically meaningful) between the residues. Both the side-chain and backbone dihedrals are included for performing these calculations. The method has been found to be immensely useful in studying allosteric effects in proteins ([McClendon et al, 2014](#); [Michel, 2014](#); [Wagner et al, 2016](#)).

3.2.5 Structure based pocket/cavity searching

Having identified a pocket that seemed to be common to the mutants, the next step involved finding ligands that would bind to this pocket and prevent unfolding of the DBD. For this it was decided to use a structure based approach. 101,609 proteins in the Protein Data Bank (PDB) are scanned to search for pockets similar to the one identified in the p53 DBD for any potential ligands. The CLICK

algorithm is used to align the pocket residues to PDB structures (Nguyen & Madhusudhan, 2011; Nguyen et al, 2011). CLICK superimposes pairs of protein structures based on similarity of local structural packing, and thus is capable of aligning structures with dissimilar topologies, conformations, or even molecular types (Nguyen & Verma, 2015). R82, E88, Q100, H101 (HIE), D115, D116, R117, F120, R121 and H122 (HID) were used to define the shape of the pocket that is identified in the S6-S7 region. These pocket residues are structurally aligned to the dataset proteins by matching cliques with similar features of Cartesian coordinates and solvent accessibility. In this study, C α and C β were chosen as the representative atoms. The root mean square (RMSD) deviation and structure overlap for each alignment between the pocket residues and dataset protein were monitored. A structure overlap cut-off of 95% and RMSD cut-off of 1.5Å was empirically chosen to determine the most accurate matches.

3.2.6 Docking

The final structure from the simulation of V143A of DBD was used for docking the compound that is identified through structure based pocket search, i.e., Dodecane trimethylamine, in the S6-S7 pocket. Docking was performed in AutoDock 4.2 using the AutoDockTools graphical user interface (Morris et al, 2009). The protein is kept rigid and the ligand is made flexible by specifying all the torsions involving heavy atoms as rotatable. The program adds hydrogens to the protein and ligand using the default protonation states of the atoms. A search space was designed in the X, Y and Z directions and was centered in the vicinity of the residues forming the pocket. 150 random states of the ligand were docked in the

defined space of the protein. A Lamarckian Genetic algorithm was used to search for optimal docking poses of the ligand with 25000000 energy evaluations for each starting state of the ligand. 30 docked conformations of the ligand in the pocket are generated and these conformations were subsequently ranked according to their binding energies. The ligand conformation with the most favorable binding energy is chosen to perform MD simulations of the DBD with the docked ligand in the S6-S7 pocket. The ligand Dodecane trimethylamine was parameterized using the Antechamber module (Wang et al, 2006) of AmberTools using the AM1-BCC charge model (Jakalian et al, 2002).

3.3 Results and Discussion

3.3.1 Differential conformational sampling of S6-S7 turn in the WT and destabilizing mutants of p53 DBD

The S6-S7 turn (residues 208-213) of the p53 DBD links the S6 and S7 β -strands (Figure 3.1). Simulations of the WT and V143A mutant of DBD at 310K show distinct conformations of the S6-S7 turn (Figure 3.2); which are quantified by measuring the distance between the CZ atom of R209 on the S6-S7 turn and the backbone carbonyl oxygen of D259 on the S9-S10 turn. The rationale behind choosing this distance measure is that these atoms appear to be interacting predominantly in the WT DBD and not in the mutant DBD.

Figure 3.2: Dominant conformational states of S6-S7 turn in p53 DBD. Figure shows structures of WT (blue) and V143A (red) DBD from MD simulations (at 310K) superimposed to highlight differential conformational sampling of the S6-S7 turn in these two forms of the domain. Atoms, which are labeled on the S6-S7 turn and S9-S10 turn, are used to quantify the conformational sampling of this turn by measuring the distance between them. WT samples a ‘closed’ state whereas V143A DBD samples an ‘open’ state of the S6-S7 turn.

This distance distribution across the simulations of the WT and V143A mutant of DBD ([Figure 3.3A](#)), at 310K (physiological temperature) peaks at ~5-7 Å (referred to as the ‘closed’ state of the S6-S7 turn) and ~23 Å (referred to as the ‘open’ state of the S6-S7 turn) respectively. However, at 298K (sub-physiological temperature), both the WT & the mutant form sample the ‘closed’ conformations of the S6-S7 turn. Contrasting to this, at 323K (above physiological temperature), the WT DBD seems to lose the ‘closed’ conformation while the mutant DBD samples the ‘open’ conformation. Existing experimental data suggests that at 298K, both the WT and the V143A mutant of DBD are thermodynamically stable but the latter loses DNA binding specificity; at 310K, only the WT DBD is stable; and at 323K, even the WT DBD is known to lose its stability ([Ang et al, 2006](#); [Bullock et al, 1997](#); [Friedler et al, 2003](#)). A hypothesis, is thus formulated, that the

S6-S7 turn samples 2 dominant conformations - '*open*' and '*closed*', that are predominant in the destabilizing (e.g. V143A) mutant(s) and WT DBD, respectively suggesting that the conformational state of the S6-S7 turn relates to the thermodynamic stability of the DBD. I now computationally probe this hypothesis.

To ascertain that the differences in the conformations of the S6-S7 turn actually arise due to the mutation, the final structure from the MD trajectory of the V143A mutant at 310K is back-mutated A143 to V143 (corresponding to the WT) and now an independent (reassigned velocities) simulation of this DBD is run. The conformation of the S6-S7 turn in this WT DBD shifts from the '*open*' towards the '*closed*' state (Figure 3.3A). Though the conformational transition to the '*closed*' state is not smooth and linear, with clear changes only apparent in one of the 3 replicates of simulations carried out, the perturbation in the '*open*' state is quite apparent. Moreover, the gradual transition from the '*open*' to the '*closed*' states could imply

Figure 3.3: Quantifying the conformational state of S6-S7 turn of p53 DBD across simulations of DBD variants by measuring the distance between CZ atom of Arg209 (S6-S7 turn) and backbone carbonyl oxygen atom of Asp259 (S9-S10 turn). (A) Distance distribution for WT and V143A DBD at different simulation temperatures. Final frame of V143A DBD simulation at 310K is ‘back-mutated’ to the WT sequence i.e. A143V, and simulated for 600ns (as shown in flowchart); (B) Distance distribution for secondary site suppressor mutations of V143A DBD and; (C) additional destabilizing mutants of p53 DBD- E258V, R110L, R175H and R248Q. Simulations suggest that a thermodynamically stable DBD (e.g. WT) is likely to sample a ‘closed’ state whereas a destabilized DBD (e.g. the destabilizing mutants) is likely to sample an ‘open’ state. Also, shown is the structure of the p53 DBD. Residues undergoing a destabilizing mutation (only the ones studied here), are shown as spheres.

Additionally, Brachmann et al. have identified three reversal or secondary site suppressor mutations for the V143A mutant of DBD ([Brachmann et al, 1998](#)) including N239Y, S240N and N268D. Of these, the N239Y and N268D point

mutations in the V143A DBD were found to restore p53 activity at 37⁰ C in yeast while the S240N was found to restore p53 activity at 30⁰ C. This agrees well with human reporter gene assays which showed increased transcriptional activity for the V143A-N239 and V143A-N268D double mutants of p53. Using this data to benchmark the calculations (characterization of the S6-S7 turn and its correlation to the stability of DBD); the following reversal mutants are simulated: V143A-N239Y, V143A-S240N, V143A-N268D at 310K from the crystal structure (PDB ID 2AHI ([Kitayner et al, 2006](#))), chain A as the starting conformation. Interestingly, the V143A-N239Y and V143A-N268D simulations of DBD largely sample the WT like '*closed*' conformation of the S6-S7 turn while the V143A-S240N simulation of DBD samples the mutant-type like '*open*' conformation ([Figure 3.3B](#)), which correlates very well with the experimentally known activity/stability of these reversal mutants. There was no point in running this last simulation at 30⁰ C, the temperature at which V143A-S240N is active, because V143A at this temperature samples the '*closed*' state ([Figure 3.3A](#)).

Could the '*open*' state of the S6-S7 turn be a general feature of the destabilizing mutants of the DBD? In order to test this hypothesis, four aggregating hot spot mutants of p53 DBD- E258V, R110L, R175H and R248Q ([Figure 3.3C](#)) are simulated. These are specifically chosen since the Aggregation Prone Region (APR) in these mutants is known ([Xu et al, 2011](#)), which is not the case for the temperature sensitive V143A DBD. E258V (extent of destabilization of this mutant is not known) is in the immediate vicinity of the APR- strand S9, while R110L (extent of destabilization of this mutant is not known) is part of the stand S1.

R175H (destabilized by ~ 3 kcal/mol (Bullock et al, 1997)) and R248Q (destabilized by 2 kcal/mol (Bullock et al, 1997)) are parts of the loops L2 and L3 respectively. These mutations are present on very different regions of the DBD and are biochemically very diverse i.e. V143A (large to small residue mutation), E258V (negatively charged to hydrophobic residue mutation), R110L (positively charged to hydrophobic residue mutation), R175H (mutation retaining polarity), R248Q (positively charge to a polar residue). These four mutants of the p53 DBD, too, sample an '*open*' conformation of the S6-S7 turn that appears to be very stable across the simulation (Figure 3.3C). This observation, thus, allows us to generalize the hypothesis: the '*open*' conformation of the S6-S7 turn could indeed suggest a destabilizing (or a destabilized) state of the DBD, while the '*closed*' conformation of the S6-S7 turn represents the stable state. These observations are also in accord with a recent computational study where it was shown that the R248Q destabilizing mutant of the DBD samples a solvent exposed S6-S7 turn while the WT samples a more buried state (Ng et al, 2015). The '*open*' and the '*closed*' conformations of the S6-S7 turn in the simulations refer to the solvent exposed and buried forms respectively.

Figure 3.4: WT p53 DBD in cartoon representation is shown on left. APR (strand S9) is colored in red. Figure on right shows distribution of the distance between CA atoms of Arg209 and Asp259, measured across simulations of WT and destabilizing mutants (V143A, E258V, R110L, R175H and R248Q) of p53 DBD as well as of WT p63 and p73 DBD. This distance measure is used to characterize the conformational state of S6-S7 turn in DBD. Shaded areas in the plot correspond to the conformational states of S6-S7 turn- '*closed*' (cyan), '*open*' (orange), '*far-open*' (pink).

I further characterize the conformational states of the S6-S7 turn in the more thermostable homologs of p53, i.e. p63 and p73 (Klein et al, 2001). Previously, the distance between the CZ atom of R209 on the S6-S7 turn and the backbone carbonyl oxygen of D259 on the S9-S10 turn is used to characterize the '*open*' and '*closed*' conformational states of the S6-S7 turn in the p53 DBD. However, the same metric cannot be used to characterize the conformational states of the S6-S7 turn in the DBDs of p63 and p73 because of differences in the side chains in these paralogs: R209 in the p53 corresponds to P279 and P229 in p63 and p73 respectively. So, the Cα-Cα backbone distance between residues at positions 209 and 259 (of p53 DBD) and the corresponding residues in the p63 and p73 is used (Figure 3.4). This backbone distance distribution is bimodal for the WT p53

DBD as well as for the p63 and p73 DBD whereas it is trimodal for the destabilizing mutants of p53 DBD (V143A, E258V, R110L, R175H and R248Q). The additional peak that is sampled almost exclusively in the destabilizing mutants is referred to as '*far-open*' conformation (pink shaded area in Figure 3.4 at a distance of ~ 25 Å). MD simulations of the WT p53 DBD, p63 and p73 DBD do not sample the '*far-open*' state within the timescales of these simulations.

3.3.2 Biological/Functional implications of the '*open*' conformation of the S6-S7 turn of the DBD?

Figure 3.5: Benzene density (black mesh) near the APR (strand S9) in WT and destabilizing mutants of p53 DBD is highlighted in blue shades. APR is shown in red.

The residues 251-257 in the S9 β -strand of the p53 DBD that act as APR (Xu et al, 2011), are largely buried in the native fold and must get exposed in order to nucleate aggregation (Wang & Fersht, 2012). Based on the idea that quantifying the exposure of this APR can presumably help us gauge the extent of destabilization of the DBD, the exposure of the strand S9 is measure using Solvent Accessible Surface Area (SASA) of strand S9. However, SASA did not report noticeable differences in the exposure of this APR between the WT and destabilizing mutants. So, benzene mapping is used as an enhanced sampling method. Previously, benzene mapping simulations have been used to identify cryptic hydrophobic pockets in the binding sites of proteins and a suitable concentration of benzene in the water box maps to the transiently exposed hydrophobic patches on the protein surface (Lama et al, 2015; Tan et al, 2012; Tan et al, 2015), thus enhancing the sampling of such regions in the MD simulations. Benzene mapping simulations of the p53 DBD at 0.2 M concentration of the benzene (Figure 3.5) shows distinct benzene density near the strand S9 in the destabilizing mutants that contrasts with the minimal benzene density at the strand S9 seen in the WT DBD. Prominent densities at the APR in the destabilizing mutants suggest a higher propensity for the APR to be exposed and hence likely to aggregate.

Figure 3.6: Interactions that stabilize the ‘closed’ (A) and ‘open’ (B) states of S6-S7 turn in WT and V143A mutant form p53 DBD respectively. Residue 143 is shown in red sticks. The ‘closed’ state is populated in simulations of the WT DBD while the ‘open’ state is largely populated in the simulations of destabilizing mutants.

Moreover, differences in the interactions that stabilize the ‘closed’ and ‘open’ states of the S6-S7 turn of the DBD reveal possible mechanisms of destabilization of the mutant forms of the protein (Figure 3.6). A particularly interesting interaction occurs between the side chain of R209 with the backbone carbonyl oxygen of D259 and the side chain of E258 in the ‘closed’ state of the S6-S7 turn (largely observed in the WT DBD). In contrast, R209 interacts with D207 in the ‘open’ state of the S6-S7 turn (largely observed in the destabilizing mutants of DBD). Although it may seem that since R209 is involved in a salt bridge in both the WT (R209-E258) and mutant (R209-D207) conformations, and it will not contribute significantly to the stability of DBD, a recent computational and experimental study suggested that Glu-Arg salt bridges may be stronger than Asp-

Arg salt bridges by up to ~ 5.9 kJ/mol (~ 1.4 kcal/mol) (White et al, 2013). Additionally, one of the destabilizing mutants (also simulated here) that is known to aggregate is E258V. This is quite interesting as it suggests that the R209-E258 salt bridge is lost in this mutant and that the loss of this interaction might well underlie the loss of stability in the E258V mutant of the p53 DBD. This strongly implies the importance of the R209-E258 salt bridge interaction in stabilizing the DBD.

Figure 3.7: Antibody binding epitopes are mapped on to the p53 DBD. Antibodies recognizing the WT p53 DBD: PAb 246: residues 201-212 (dark green); PAb 1620: residues 201-212 & 145-157 (light green). Antibodies recognizing mutant p53: PAb240: residues 213-217 (magenta), DO11: residues 181-190 (orange), DO12: residues 256-270 (red)

A good test of the conformational changes in the DBD is the use of antibodies that have been developed to bind to specific epitopes on the p53 DBD and report on its WT and mutant states. Antibodies PAb246 and PAb1620 bind to the native state of p53 DBD; while antibodies PAb240, DO11 and DO12 bind to

the denatured states of the protein (Wang et al, 2001). Figure 3.7 shows the epitopes of these antibodies mapped onto the p53 DBD. The solvent accessible surface area for these epitopes is computed in the MD simulations of the WT and mutant variants of DBD (Figure 3.8). No major differences in the exposure of these epitopes between the WT and mutant forms of the protein were observed, thereby suggesting that such conformational changes likely require much longer simulations.

Figure 3.8: SASA of the antibody binding epitopes of the WT (black) and the destabilizing mutants (red) including V143A, E258V, R110L, R175H and R248Q of p53 DBD in the MD simulations. Antibodies recognizing the native p53 DBD: (A) PAb 246, (B) PAb 1620; Antibodies recognizing the denatured p53 DBD: (C) PAb240, (D) DO11, (E) DO12.

A recent computational study reported that the DNA bound state of DBD is likely to assume a '*closed*' conformation for the S6-S7 turn [\[Just accepted\]](#)

manuscript- Matteo Lambrughi et al. (2016), DNA-binding protects p53 from interactions with cofactors involved in transcription-independent functions, NAR]. Given that DNA binding stabilizes p53 (Silva et al, 2010); it implies that the 'closed' conformational state of the S6-S7 turn of p53 DBD represents a low energy state of DBD. Similarly, the 'open' state of the S6-S7 turn of DBD might represent an energetically unfavorable state. Additionally, negative regulators of p53 transcriptional activity such as Aurora-A kinase, Ku70 were predicted to bind to the 'open' state of the S6-S7 turn of DBD [Just accepted manuscript- Matteo Lambrughi et al. (2016), DNA-binding protects p53 from interactions with cofactors involved in transcription-independent functions, NAR]. Finally, phosphorylation of S215 (present in strand S7) of p53 DBD is known to destabilize the protein (Cai et al, 2012; Fraser et al, 2010; Liu et al, 2004; Warnock et al, 2011). It was shown that the p53 DBD with phosphorylation on S215 exclusively samples an 'open' state of the S6-S7 turn in MD simulations. Together, these studies provide compelling evidence that strongly associates the 'open' state of the S6-S7 turn of DBD with the loss of native conformation and onset of destabilization in the p53 DBD.

Figure 3.9: Final frames from the 500ns triplicate simulations of the p53 DBD (residues 91-293) are shown. The N-terminus is shown in ribbon representation and the rest of the protein is shown in surface. It is apparent that the N-terminus in the WT DBD samples a crystal-like native conformation whereas the N-terminus in the destabilizing mutants of the DBD- R175H and R248Q is displaced from its native position into a non-native conformation.

I next examine the N-terminus of the p53 DBD as it wraps the Aggregation Prone Region (APR) or strand S9, which is buried in the WT conformation. It is clear that the N-terminus must be displaced from its native position for the strand S9 to be exposed to solvent ([Wang & Fersht, 2012](#)). Fersht et al. have also shown that the extended N-terminus of the p53 DBD (residues 91-293) stabilizes the DBD by ~2kcal/mol ([Natan et al, 2011](#)); all the simulations reported here so far have been carried out on residues 94-293 of the DBD. Motivated by this, the conformation of the N-terminus in the simulations of the WT and two of the most common destabilizing mutants of p53 DBD i.e. R248Q and R175H, with an extended N-terminus (i.e. up to residue 91) is investigated. It is observed that the N-terminus appears to get destabilized and displaced from its position (in the crystal structure) into a non-native conformation in the mutants ([Figure 3.9](#)). In contrast, the N-terminus in the simulations of the WT DBD samples a crystal

structure-like native conformation. The simulations with the extended N-terminus too show that the destabilizing mutants appear to sample the ‘*open*’ state of the S6-S7 turn of the DBD with a greater probability (Figure 3.10).

Figure 3.10: Distribution of the distance between CZ atom of R209 (S6-S7 turn) and backbone carbonyl oxygen atom of D259 (S9-S10 turn) for the WT (blue), R175H and R248Q mutants of p53 DBD (red).

Together, the analysis strongly suggests that the ‘*open*’ conformation of the S6-S7 turn of p53 DBD (predominantly observed in the destabilizing mutants) represents a destabilized state of the protein and might be an initial event in the onset of unfolding and aggregation in the p53 DBD.

3.3.3 A network of interacting residues that regulates the conformation of the S6-S7 turn

Simulations of the p53 DBD suggests that the S6-S7 turn assumes '*open*' and '*closed*' states in the destabilizing mutants and WT forms respectively. I next proceeded to identify any path describing the '*closed*' to '*open*' state conformational change, that could link, changes in the mutation site to the effector site (the S6-S7 turn where the conformational changes are observed). For this, Kullback-Leibler divergence (KL-div) calculations are performed on the MD ensembles of the WT and destabilizing mutants of the p53 DBD using the program MUTINF developed by Christopher L. McClendon et al ([McClendon et al, 2009](#); [McClendon et al, 2012](#)). KL-div characterizes the differential sampling of residues between two MD trajectories.

Figure 3.11: Structures of p53 DBD colored according to divergence value derived from KI-div computations across simulations of wild and respective mutant DBD. Mutation sites are shown with a red arrow.

I compare the distribution of internal coordinates of the residues between the MD trajectories of the WT and the destabilizing mutant forms of the DBDs: V143A, E258V, R110L, R175H and R248Q ([Figure 3.11](#)). The residues with a high degree of divergence (between the WT and mutant DBD) are located in loops L2, L3 and strands S6, S7. This correlates well with the available experimental data on the NMR of the destabilizing mutants R248Q and V143A, which suggests that residues in L2, L3 and S7 are the ones affected between the WT and the destabilizing mutants ([Wong et al, 1999](#)).

The residues coordinated to the Zinc ion and other vicinal residues have the largest values of divergence. Experimental data suggesting that the loss of Zinc ion from the DBD is a major factor for the destabilization and subsequent aggregation is only available for the R175H mutant ([Blanden et al, 2015](#)); in any case, the Zinc remains covalently linked in these simulations and hence I cannot comment on this aspect of destabilization.

Residues (Crystal no.)	WT- V143A mutant (310K)	WT-E258V mutant (310K)	WT-R110L mutant (310K)	WT-R175H mutant (310K)	WT-R248Q mutant (310K)
Ser 94					
Ser 95					
Ser 96					
Pro 98					
Ser 99					
Gln 100					
Lys 101					
Thr 102					
Arg 110					

Leu 111					
Gly 112					
Phe 113					
Leu 114					
His 115					
Ser 116					
Gly 117					
Thr 118					
Ala 119					
Lys 120					
Ser 121					
Val 122					
Thr 123					
Cys 124					
Thr 125					
Tyr 126					
Ser 127					
Leu 130					
Met 133					
Phe 134					
Cys 135					
Gln 136					
Leu 137					
Lys 139					
Thr 140					
Cys 141					
Pro 142					
Val 143					
Gln 144					
Trp 146					
Arg 156					
Arg 158					
Met 160					
Ile 162					
Tyr 163					
Lys 164					
Gln 165					
Ser 166					
Gln 167					
His 168					
Met 169					

Thr 170					
Glu 171					
Val 173					
Arg 174					
Arg 175					
Cys 176					
Phe 177					
His 178					
His 179					
Glu 180					
Arg 181					
Cys 182					
Ser 183					
Asp 184					
Ser 185					
Asp 186					
Gly 187					
Leu 188					
Ala 189					
Pro 190					
Pro 191					
Gln 192					
His 193					
Leu 194					
Ile 195					
Arg 196					
Val 197					
Glu 198					
Asn 200					
Tyr 205					
Leu 206					
Asp 207					
Asp 208					
Arg 209					
Asn 210					
Thr 211					
Arg 213					
His 214					
Ser 215					
Val 216					
Val 217					

Tyr 220					
Glu 224					
Val 225					
Gly 226					
Ser 227					
Asp 228					
Cys 229					
Thr 230					
Ile 232					
His 233					
Tyr 234					
Asn 235					
Tyr 236					
Met 237					
Cys 238					
Asn 239					
Ser 240					
Cys 242					
Met 243					
Gly 244					
Gly 245					
Met 246					
Arg 248					
Arg 249					
Ile 251					
Ile 255					
Glu 258					
Arg 267					
Glu 271					
Arg 273					
Val 274					
Cys 275					
Glu 285					
Glu 286					
Asn 288					
Leu 289					
Arg 290					

Table 3.2: KL-div values are shown for different destabilizing mutants. Entry is colored green if the residue exhibits divergence value of >0.2 (0.2 is an arbitrary cut off chosen to filter sites with minimal divergence) between WT and mutant simulations.

A tabulation of the divergence calculations for several aggregating mutants (Table 3.2) shows the residues that consistently sample different dihedral between the WT and the aggregating mutants. Visualization of the dynamics of these residues allowed us to identify a subset of residues which appears to be associated with the ‘opening’ of the S6-S7 turn in the destabilizing mutants of p53 DBD i.e. E258V, R175H, R110L and R248Q. These residues are R174, E180, Q192, D207, R209 and H214. Figure 3.12 shows the time series plots for the distance and the dihedral reaction coordinates involving this subset of residues for the V143A DBD. It is quite apparent how the perturbations in one reaction coordinate trigger a downstream change in another. To begin with, residues R174-E180-Q192 appear to form an interacting triad. During the course of the simulation, some unidentified perturbation destabilizes this triad (Plot A in Figure 3.12). The loss of interaction between Q192 and R174 is followed by a change in the χ_1 dihedral angles of Q192, H214 and D207 (Plot B in Figure 3.12). This results in the ‘opening’ of the S6-S7 turn as can be seen by the distance between CZ of R209 and the backbone carbonyl O of D259. Finally, this change displaces the N-terminus (measured by the V172 – D259 distance) which then adopts a non-native conformation (Plot C in Figure 3.12). A very recent computational work [Just accepted manuscript- Matteo Lambrughi et al. (2016), DNA-binding protects p53 from interactions with cofactors involved in transcription-independent functions, NAR] reported the existence of allostery between the DNA binding region (Loop L1) and the S6-S7 turn of the DBD. The simulations and the KL-div computations not only confirm

their observations but also identify a probable route along which the changes in conformations of residues regulate the conformational state of the S6-S7 turn.

Figure 3.12: A change in the interaction pattern of a subset of residues across MD simulations is associated with the ‘opening’ of the S6-S7 strand in the destabilizing mutants of p53 DBD. Structure of V143A DBD is shown on the left. Plots (A), (B) and (C) on the right show that how a perturbation in the reaction coordinate (involving the residues labelled in the structure on the left) (dihedral/distance) trigger a change in another. Plot (A) shows that the triad of three interacting residues (R174, E180 and Q192) is destabilized. Plot (B) records ensuing changes in the side chain dihedrals of the vicinal residues (Q192, D207, H214). Subsequently the S6-S7 strand assumes an ‘open’ conformation (Plot (C)) displacing the N-terminus into a non-native conformation (as shown in the structure on the left).

Figure 3.13: S6-S7 turn in R174A-E180A double mutant of p53 DBD changes its conformation from '*closed*' state (blue shaded region) to an '*open*' state (red shaded region) in its MD simulation.

To test the hypothesis that a stable R174-E180-Q192 interaction triad is vital for the '*closed*' conformation of the S6-S7 turn, a double mutation (R174A-E180A) is introduced in the WT p53 DBD and performed a 500ns NPT simulation. The interaction triad is no longer stable and Q192 loses interactions with residues A180 and A174. Subsequently, the S6-S7 turn assumes an '*open*' conformation (Figure 3.13). Furthermore, the loop L3 is observed to assume a non-native conformation in this double mutant of DBD. However, single mutation simulations of R174A and E180A sample a '*closed*' conformation of the S6-S7 turn (data not shown) thereby suggesting that the loss of Q192 interactions with both R174 and E180A might be crucial to trigger the changes leading to the loss of the '*closed*' conformation.

Figure 3.14: A region in the crystal structure of WT p53 DBD that was resolved with an extended N-terminus (residue number 91 onwards) is shown. Interaction triad of R174-E180-192 might be crucial for a stable cation- π interaction between W91 and R174 that is known to stabilize p53 DBD by ~2 kcal/mol.

Several lines of evidence underscore the significance of these residues which are involved in this intricate network of interactions. R174 interacts with an N-terminal residue W91 and is known to stabilize p53 DBD by ~2 kcal/mol ([Natan et al, 2011](#)); indeed, a stable and less flexible R174 might be crucial to maintaining this interaction with W91 ([Figure 3.14](#)). A very exciting piece of evidence for the importance of these ‘network-residues’ comes from an examination of the DBD of p53 of the subterranean blind mole rat (*Spalax*) that leads a hypoxic life and is not prone to cancer. Striking differences between the *Spalax* and human p53 map to residues 174 and 209, which form two ends of the interaction network ([Figure 3.12](#)), identified here. Both 174 and 209 are Arg in humans but Lys in *Spalax*. A double mutant of Arg at these two positions in *Spalax* p53 is known to be five times more potent in promoting growth arrest ([Ashur-Fabian et al, 2004](#); [Pagano et](#)

[al, 2013](#)) underscoring the significance of Arg at these two positions. Additionally, R174 and R209 have been directly implicated in cancer: R174L point mutations in human p53 has been noted in breast, colon, lung and liver cancers, while an R209 mutation (to K among others) often exists in several cancers like skin, colon and oesophagus. Finally, these network residues are quite conserved across other species ([Figure 3.15](#)). Since, these network residues are not directly involved in any intermolecular interactions of p53 DBD (to the best of my knowledge); it is likely that they are critical to the conformational regulation of the S6-S7 turn and hence, to the structural viability of DBD.

Figure 3.15: Multiple sequence alignment of p53 from different organisms. Also included in the alignment are sequences of human p63 and p73. The alignment is adopted from the IARC p53 mutation database (Olivier et al, 2002) and edited to highlight the conservation of (network) residues that seem to regulate the conformation of S6-S7 turn in human p53 DBD.

3.3.4 Energetic barrier between the 'open' and 'closed' conformational states of S6-S7 turn

Figure 3.16: Potential of mean force across the reaction coordinate (distance between CZ of R209 and backbone carbonyl oxygen of D259) that is used to characterize the conformation of S6-S7 turn of p53 DBD. The WT DBD (in blue ribbon) and V143A mutant DBD (in red ribbon) that dominantly sample the 'closed' and 'open' states of the S6-S7 turn are shown. The S6-S7 turn and the S9-S10 turn in these structures are shown in green surface while the APR (strand S9) is shown in yellow surface.

I next investigate the nature of any energy barriers between the ‘*open*’ and ‘*closed*’ states of the S6-S7 turn of the WT and V143A mutant of p53 DBD. Umbrella sampling is performed using the distance between the CZ of R209 and the backbone carbonyl O of D259 as the reaction coordinate (this is one of the major interactions characterizing the ‘*closed*’ conformation of the S6-S7 turn). In [figure 3.16](#), it is seen that the WT and mutant DBD follow very different trends suggesting that the WT DBD is likely to sample a ‘*closed*’ conformation while mutant DBD is likely to sample an ‘*open*’ conformation of the S6-S7 turn.

3.3.5 D207A as a global suppressor mutation for the destabilizing mutants of p53 DBD

Figure 3.17: Schematic diagram showing the idea behind proposing D207A as a global site suppressor mutation. The WT DBD largely samples a ‘*closed*’ state (R209 interacts with E258 and D259 on the S9-S10 turn residues) while a destabilizing mutant of DBD samples an ‘*open*’ state (R209 interacts with D207 on the S6 strand). D207A mutation will disrupt the R209-D207 interaction and presumably drive the system towards the ‘*closed*’ state.

Identification and analysis of the network residues that regulate the S6-S7 turn conformation allows us to propose and computationally test D207A as a global suppressor mutation for the destabilizing mutants of p53. It is observed that, in the simulations of WT, the side chain of R209 interacts with backbone carbonyl oxygen of D259 on the S9-S10 turn and this appears to be an important interaction stabilizing the ‘*closed*’ state of the S6-S7 turn. Contrary to this, R209 loses this interaction in the ‘*open*’ state of the S6-S7 turn of the mutant DBD and instead interacts with D207. Based on the idea that disrupting the R209-D207 interaction in the mutant DBD might shift the conformation of S6-S7 turn towards the WT-like ‘*closed*’ state, Aspartate is mutated to Alanine at position 207 in the destabilizing mutants of DBD (Figure 3.17).

Figure 3.18: Hypothesizing D207A as a proposed global suppressor mutation for the destabilizing mutants of p53. Distance distribution between CZ of R209 and backbone carbonyl oxygen of D259 for different variants of p53 DBD are shown. This distance measure is used to characterize the conformation of the S6-S7 turn across simulations. Highlighted are the distances that correspond approximately to the ‘closed’ and ‘open’ state of the S6-S7 turn.

To investigate this, the following double mutants of p53 DBD: V143A-D207A, E258V- D207A, R110L- D207A, R175H- D207A and R248Q- D207A are simulated to $\sim 1 \mu\text{s}$ each. Two replicates of V143A-D207A and one replicate of R248Q-D207A distinctly sample the ‘closed’ state of the S6-S7 turn (Figure 3.18).

R209 is unable to engage in a charge-charge interaction with Alanine at position 207 and lacks a stable interaction partner near the S6-S7 region. Instead R209 largely interacts with backbone carbonyl oxygen of D259, a feature characteristic of the WT DBD. Other simulations sample an intermediate conformation (a broad distribution with a peak at a distance of $\sim 16\text{\AA}$). However, the shift from the '*open*' conformation is quite obvious. Thus, extending these simulations further might sample stable '*closed*' states of the S6-S7 turn for these double mutants. It can be speculated that the '*open*' state of the S6-S7 turn represents a destabilized DBD state, and D207A could act as a secondary site suppressor mutation for a set of destabilizing mutants of p53 DBD.

3.3.6 Identification and Characterization of a pocket that is predominantly formed in the destabilizing mutants

Figure 3.19: Structures of (A) WT and (B) V143A mutant of DBD from MD simulations. The S6-S7 region is shown in surface. The pocket formed in the mutant is shown in dotted line. (C) Residues used to quantify the pocket are shown in p53 DBD. Two dihedral angles (arrows) and a distance measure (dotted circle), shown in the figure on the right, are used to define a geometric criteria to characterize the pocket.

Taking all our observations, I next ask if it is possible to identify a cavity through which the conformation of the S6-S7 turn may be regulated. A pocket is found that is predominantly formed when the S6-S7 turn occupies the ‘open’ conformation (Figure 3.19). In contrast, this pocket is largely nonexistent during the simulations of the WT DBD. Table 3.3 shows the occupancy of the pocket in the simulations of the WT and the destabilizing mutants of DBD studied. The largest occupancy of the open pocket is seen for the R248Q destabilizing mutant. The amino acids lining this pocket are: R174, E192, D207, D208, R209, N210, T211, F212, R213 and H214. The pocket is hydrophobic at the base while the walls of the pocket have charged moieties.

WT			V143A			E258V	R110L	R175H	R248Q
298K	310K	323K	298K	310K	323K	(310K)	(310K)	(310K)	(310K)
6.4	23.3	11.3	12.5	33.3	31.1	62.6	35.1	29.2	82.4

Table 3.3: Quantifying the ‘open’ state of the S6-S7 pocket across the MD simulations of WT and destabilizing mutants of p53 DBD using a geometric criterion. The pocket is not formed (‘closed’) if it satisfies the mathematical expression below and formed otherwise -

$[(-60^\circ > \text{Chi1 of H214} > -120^\circ) \text{ OR } ((-180 < \text{Chi1 of Q192} < -120) \text{ or } (180 > \text{Chi1 of Q192} > 120)) \text{ OR } (\text{distance between CZ of R209 \& CG of D207} > 6 \text{ \AA})]$.

Numbers represent the open state of the pocket expressed as percentage of the total simulation time of the trajectory.

3.3.7 Is the S6-S7 pocket of p53 DBD druggable and functional?

It is well recognized that stabilizing the destabilizing mutants of p53 is a potential approach to treat cancer (Brown et al, 2011a; Brown et al, 2009a). I wondered if the S6-S7 pocket observed in the simulations of destabilizing mutants

of p53 DBD has the potential to be a druggable site that can, upon occupation by a suitable ligand, stabilize the DBD in its WT conformation. There are two features in favor of this cavity's potential as a druggable site. A yet unreported dehydron is identified in the S6-S7 turn (D208 - R213) in the simulations (which is absent in the X-ray crystal structures of DBD) (Figure 3.20) (dehydrons are backbone hydrogen bonds in proteins that are prone to hydration (Fernandez & Berry, 2002), are concentrated at protein-protein/ligand interfaces (Fernandez, 2004a; Fernandez et al, 2003; Fernandez & Scheraga, 2003; Fernandez & Scott, 2003) and their presence has proved immensely useful in the structure guided design of inhibitors (Fernandez & Crespo, 2008; Fernandez et al, 2007); see Chapter 2). This suggests that a small molecule could potentially bind to the S6-S7 pocket. In addition, Φ -value analysis (suggestive of the extent of local unfolding) of the p53 DBD (Wang & Fersht, 2015) suggested that residues, V216 and V217, residues that are present at the base of this pocket, exhibit the highest values of Φ_u (Φ for unfolding) implying significant local unfolding of the molecule. This further suggests that these two residues (as well as the vicinal residues) may be responsive to compounds designed to stabilize the DBD and inhibit unfolding and aggregation. Therefore, I proceeded to explore the possibility of finding a ligand that could fit into this pocket.

Figure 3.20: Dehydrons (green dashes) mapped onto the final frame from the simulation of V143A destabilizing mutant of the p53 DBD. Two new dehydrons (D208 - R213) in the S6-S7 turn of the domain are identified and highlighted in pink circle.

3.3.8 Search for compounds that could dock into the S6-S7 pocket of p53 DBD

Figure 3.21: The final frame of the V143A mutant type DBD is taken as a representative to highlight the S6-S7 pocket. The residues lining the pocket are indicated.

I scanned the entire Protein Data Bank (PDB) to search for similar (structurally and/or chemically) pockets with the hope that one of them might already have been crystallized with a suitable ligand bound. The pocket is defined by residues R174, E180, Q192, H193 (HIE), D207, D208, R209, F212, R213 and H214 (HID) ([Figure 3.21](#)). Using the program CLICK (see [Methods section 3.2](#)) 3

structures of Beta-lactoglobulin (PDB IDs 4IB7, 4IBA and 4KII) with similar pockets are identified that had the ligands Dodecane trimethylamine, Dodecyl sulfate and [N,N-di(pyridin-2-yl-kappaN)dodecanamide]rhodium bound. Because all the ligands were very similar (Figure 3.22), Dodecane trimethylamine was chosen for further investigation.

Figure 3.22: Structure based search of the S6-S7 pocket in p53 DBD identified small molecules that are bound to a similar pocket in Beta-lactoglobulin. (A) dodecane trimethylamine (B) Dodecyl sulfate (C) [N,N-di(pyridin-2-yl-kappaN)dodecanamide]rhodium.

I next docked Dodecane trimethylamine into the S6-S7 pocket of V143A DBD using Autodock (Morris et al, 2009) and subjected the complex with DBD to

100ns MD simulations. A ligand was considered dissociated if the distance between the centre of mass (COM) of the His214 (at the base of the pocket) and that of the ligand increases by 5 Å compared to that at start of the simulation. It is important to note that the part of the ligand, Dodecane trimethylamine, that is present inside the pocket, is very hydrophobic. Despite lacking any specific electrostatic or hydrogen bonding interactions with the protein, the van der Waals interactions hold the ligand in the pocket in four of the five replicates. This likely implies the potential of the pocket to bind to drug-like compounds. Additionally, when the ligand is bound to the pocket, the S6-S7 turn, however, samples an '*open*' conformation, which is characteristic of the mutant form of DBD ([Figure 3.23](#)). The ligand is then removed from the structure at the end of the simulation and then allowed the apo DBD to evolve for another 100ns. The fluctuations in S6-S7 are significantly less ($\sigma = 1.11$) in the ligand bound state compared to when the ligand is removed ($\sigma = 2.57$), thereby seemingly locking the S6-S7 in the '*open*' state, compared to the subsequent apo DBD simulations. Though it may appear that the ligand has stabilized the mutant-like '*open*' state of the S6-S7 turn of DBD, which contradicts the idea of stabilizing the WT-like '*closed*' state, it is not surprising since the DBD is in an '*open*' state at the start of the simulation.

Figure 3.23: (A) Dodecane-trimethylamine is identified from PDB as bound to a pocket (similar to that of S6-S7 pocket) in Beta-lactoglobulin. (B) Dodecane-trimethylamine is subsequently docked in the S6-S7 pocket of the V143A mutant of p53 DBD and simulated. (C) Conformation of the S6-S7 turn when the ligand is bound and dissociated to/from the pocket.

In summary, the simulations of the ligand docked into the S6-S7 pocket suggest that a drug like molecule can fit into the pocket. Moreover, this pocket could indeed be functional as suggested by the presence of dehydrons in the S6-S7 turn and the Φ_u values of residues in the pocket available from a recent study. Experimental studies on this pocket will be required to validate these findings.

3.3.9 Testable hypothesis to probe loss of function in destabilizing mutants of p53 DBD

In summary, an exhaustive set of simulations probing the dynamics of the S6-S7 loop as a reporter for the instability of the DBD of p53 suggest some testable hypotheses:

- (a) Given the finding that the interacting triad R174-E180-Q192 is broken as the S6-S7 loop assumes the “*open*” state, a double alanine mutation R174A/E180A should result in lowering of the melting temperature of the DBD as it will sample a distinct ‘*open*’ state of the S6-S7 turn (Figure 3.13).
- (b) Figure 3.15 suggests that Histidine at position 214 is completely conserved in all the orthologs of p53. However, the thermostable paralogs of human p53- the p63 and the p73 have a Glutamine at this position. Hence, the H214Q mutant may result in an increase in the melting temperature of p53. In addition, H214 is present at the base of the S6-S7 pocket identified in this study. The H214Q single mutant of p53 has been shown to be more potent at inducing apoptosis than the WT DBD and is regarded as a ‘super p53’ (Kakudo et al, 2005). It is possible that it may be a powerful second site suppressor.
- (c) Finally, exposing mutant p53 DBD to varying concentrations of dodecane trimethylamine or derivatives to see if such molecules enhance the thermodynamic stability of the mutants. The hypothesis is that this compound can bind to the S6-S7 pocket. Moreover, recent experimental data on the Φ -value analysis of p53 DBD suggest that the pocket is receptive to compounds that are aimed at stabilizing the DBD (Wang & Fersht, 2015).

I hope that the experiments performed to test these hypotheses will begin to provide answers to the following questions:

- 1] What is the likely unfolding pathway for p53 DBD and the initial events triggering the pathway? Does it vary across the mutants of p53?
- 2] What is the role of the S6-S7 turn dynamics in the unfolding pathway?
- 3] Is the novel pocket in the S6-S7 region druggable and functional? Can a molecule bind to the pocket and stabilize the DBD?

3.4 Conclusion

A subset of mutations in TP53 is known to destabilize the protein, cause unfolding and aggregation and is linked to cancer pathologies. Although, most of these mutations are monoallelic and missense in nature, a dominant negative effect deactivates the WT allele ([Brosh & Rotter, 2009](#)). It is well recognized that stabilizing mutant p53 to reactivate its WT function can allow us to specifically target cancer/tumor cells ([Rangel et al, 2014](#); [Wang & Fersht, 2015](#)). However, the diverse nature of these mutations in the TP53 gene makes restoring normal levels of the native p53 in cells a very challenging task. Partial solutions to this problem may emerge from knowledge of the unfolding and aggregation pathways that characterize the dynamics of p53 and its mutants.

The DNA Binding Domain (DBD) of p53 houses most of the high frequency oncogenic mutations as well as the Aggregation Prone Region (APR) and suffices as a model for the study of aggregation in p53. Extensive MD

simulations of the WT and numerous destabilizing mutants of the p53 DBD carried out in this study revealed differential sampling of the S6-S7 turn between the WT and the destabilizing mutants. It is clear that the WT and mutant DBD largely sample ‘*closed*’ and ‘*open*’ conformations of the S6-S7 turn *respectively*. This observation was benchmarked against the available experimental data and through several control simulations. Rigorous free energy calculations using the Potential of Mean Force (PMF) were carried out and further support these observations. Additionally, differential sampling of the side chain and backbone dihedral angles between WT and mutant DBD revealed a network of interacting residues that governs the conformational states of the S6-S7 turn. These residues belong to the loop L2, helix H1 and β -strands- S6 and S7. Thus, a probable mechanism of the loss of function in the destabilizing mutants of the p53 DBD is hypothesized: the ‘*open*’ conformational state of the S6-S7 turn probably triggers events that are associated with activating the unfolding pathway of DBD resulting in the onset of loss of stability in the DBD. Further, it is shown that incorporating a secondary mutation D207A in the destabilizing mutants of DBD appears to induce these double mutants to sample a WT-like ‘*closed*’ conformation of the S6-S7 turn. Thus, D207A is proposed as a global secondary site suppressor mutation for restabilizing the destabilizing mutants of DBD.

Finally, a pocket that is formed near the S6-S7 region in the mutant p53 DBD is identified and characterized. Previous Φ -value analysis studies and an analysis of dehydrons in the ensemble structures (performed in this study) provide compelling evidence that the pocket is druggable and functional. A search of the

PDB database for pockets similar to the one identified in mutant DBD here, lead to the observation of a compound Dodecane trimethylamine, bound to Beta-lactoglobulin which shares a similar pocket. This ligand was found to dock and remain stable in MD simulations of the complex with mutant DBD.

Taken together, the investigations of the dynamics of WT and destabilizing mutants of p53 DBD provide atomistic insights into probable initial events that result in the onset of aggregation and also offer routes to the development of novel restabilizers of mutant p53.

Preface to Chapter 4

In chapters 2 and 3, a hydration metric to quantify the aggregation prone regions in proteins is proposed and tested. In the next chapter, I develop a software package of programs for efficient characterization of hydration properties in structural ensemble of biomolecules. Along with the standard computations (solvent density, residence time, distance of closest water and favorable/unfavorable solute-solvent interactions), 3 novel algorithms, to compute buried waters in proteins, to identify multi-water bridges in proteins and to compute the displaceability of waters at the protein-ligand interfaces, are embedded. While buried waters provide critical insights into the thermostability of proteins, evaluating the likelihood of favorably displacing interfacial waters is a routine chore performed during drug optimization.

Chapter 4: ‘JAL’: A software package for efficient characterization of hydration properties in structural ensembles of biomolecules

Water interactions with proteins are critical for modulating protein stability and protein-protein/ligand interactions. Explicit solvent Molecular Dynamics (MD) simulation are routinely used to probe such interactions. However, detailed analysis of the hydration patterns seen in the MD simulations can be both time consuming and complex to interpret. There are no general purpose publicly available packages that enable comprehensive and rigorous computation of hydration properties, particularly for buried waters and interfacial waters in an MD ensemble. Here, I present a free software package (‘JAL’) consisting of programs that compute hydration properties of a protein, nucleic acid or in general any molecule from its MD trajectory; these include solvent density, residence time, buried waters, favorable/unfavorable solvent interactions and closest-water distance measures. The utility of ‘JAL’ is demonstrated through analysis of two proteins, p53 DNA Binding Domain (DBD) and the translational initiation factor eIF4E. A study of buried waters in the p53 DBD revealed a hitherto unappreciated buried water molecule in the loop L3 that regulates the conformation of this loop and the stability of the domain. In eIF4E, I demonstrate how the computations of

residence times can be used to identify water bridges in ligand binding sites. Additionally, a novel knowledge-based approach (also embedded in ‘JAL’) that evaluates the likelihood of favorably displacing an interfacial water molecule by mining the structural dataset of proteins is proposed. Using the examples of scytalone dehydratase and EGFR kinase for which experimental data is available, very precise insights are provided not only about which interfacial waters to displace but also towards the design of fragments that can be used to optimally design a ligand. ‘JAL’ is hosted at [‘mspc.bii.a-star.edu.sg/minhn/jal’](http://mspc.bii.a-star.edu.sg/minhn/jal).

4.1 Introduction

Optimal levels of protein hydration are critical for maintaining the native states of proteins and in mediating their interactions with other molecules. A detailed picture of protein hydration can provide insights into the associated thermodynamics and kinetics (Ball, 2008a; Ball, 2008b; Csermely, 2001; Levy & Onuchic, 2004). However, protein hydration is extremely sensitive to surface chemistry and topology which have very diverse characteristics, thus making the study of hydration very challenging. Several experimental (Ball, 2003) approaches have been used to study biomolecular hydration. For example, multiple, independently determined, high resolution crystal structures of proteins in the Protein Data Bank (PDB) (Thanki et al, 1988) have lead to the identification of consensus hydration sites (Savage & Wlodawer, 1986). Nevertheless, crystal structures correspond to an average and are rarely representative of the biological conformation and, while B-values have provided some guidance on flexibility, a

more dynamic picture is necessary for understanding biological function (Madhusudhan & Vishveshwara, 2001) (Donald Hamelberg, 2008). This need has been complemented by developments in experimental techniques such as Nuclear Magnetic Resonance (NMR) (Otting et al, 1991), femtosecond spectroscopy (Zhong et al, 2011), vibrational spectroscopy (Panman et al, 2013), ultrafast optical Kerr-Effect spectroscopy (Hunt et al, 2007), 4D Ultrafast Electron Crystallography (Yang & Zewail, 2009) and more recently by advances in computational techniques such as MD simulations (Dodson et al, 2008; Helms & Wade, 1995; Helms & Wade, 1998; Lazaridis & Karplus, 1997; Wade et al, 1991). MD provides a detailed spatio-temporal picture of conformational variability and associated dynamic interactions, including water interactions.

A range of computational algorithms have been developed in the past to predict water locations around a protein structure. Very early on, GRID (Goodford, 1985) was the method of choice; this involved placing a water molecule around the protein and evaluating its interaction potential to identify hydration sites. Soon after, AQUARIUS was developed using a knowledge-based approach, utilizing amino acid specific hydration patterns from structures in the PDB (Thanki et al, 1988) to compute solvent structures around proteins. SuperStar (Verdonk et al, 1999) creates solvent propensity maps in protein cavities by using existing knowledge about structures from the PDB (Berman et al, 2000) and CSD (Cambridge Structural Database) (Allen, 2002). More recently, a semi-heuristic approach to predict solvent structure was developed that involves the iterative placement and computation of water propensities at protein sites (Setny &

Zacharias, 2010). HyPred (Virtanen et al, 2010) is a method that predicts solvent density using radial distribution functions for a set of atom types around a constrained protein in an all atom explicit solvent MD simulation. Mobywat (Jeszenoi et al, 2015) uses MD simulations to compute hydration patterns based on water dynamics. Some algorithms are specifically designed to only predict buried waters in proteins; one such method uses the FOLD-X force field (Schymkowitz et al, 2005) while another uses a semi-explicit, discrete solvent model (Setny, 2015). In addition, some algorithms are available as plugins to commonly used molecular visualization software packages. For instance, WATsite (Hu & Lill, 2014) and PyWATER (Patel et al, 2014) are plugins to the Pymol visualizer (Schrodinger, 2010) and predict invariant waters as well as hydration sites at binding interfaces based on analyses of superimposed structures respectively. Some other efforts involve setting up publicly available web resources for different aspects of the analysis of water interactions. For instance, 3DSS (Sumathi et al, 2006) is a web server for superimposition of multiple structures of a protein and computation of invariant water molecules. Similarly, Water Analysis Package (WAP) (S. Praveen, 2008) and Protein Structure Analysis Package (PSAP) (B. Balamurugan, 2007) are web servers that compute and analyze water interactions with proteins in a single input structure. A variety of programs for computing solvent properties in an ensemble of conformations are embedded in the standard trajectory analysis modules of the MD simulation packages CHARMM (Brooks et al, 2009), Amber (D.A. Case, 2015) and Gromacs (Pronk et al, 2013).

Is there a way to test the validity of these methods? One could of course embark on a set of elaborate and finely controlled biophysical experiments to try and test hypotheses regarding hydration and yet this is not really practical. There is however one particular avenue that lends itself to experimental tests. If the hydration patterns extracted from such analyses have rigour, then a question can be asked: given a water molecule in a binding site that this analysis has helped identify, can it be determined whether this water is displaceable while designing drugs to fit into the binding site? Small modifications to a ligand affecting the hydration in the binding cavity have been shown to have dramatic effects on the binding affinities of the modified ligands ([Michel et al, 2009a](#)). It is clear that hydration is associated with enthalpic as well as entropic factors and the commonly used enthalpic measure by itself is not sufficient to capture hydration properties rigorously. The traditional approaches of rigorous desolvation free energy calculations to quantify the effects of displacing such water molecules have been quite powerful. For example, the binding energies of interfacial water computed using the sophisticated and yet computationally demanding double decoupling ([Barillari et al, 2007](#)) method suggested that on average the displaceable waters were bound strongly by 2.5 kcal/mol; however, it is not clear whether this method can easily discriminate between the conserved and displaceable waters. There are several programs that are ongoing to resolve these issues. Programs such as WaterMap ([Abel et al, 2008](#); [Young et al, 2007](#)), SZMAP ([Bayden et al, 2015](#)), GRID/CRY probe ([Goodford, 1985](#)), IFST ([Huggins & Payne, 2013](#); [Lazaridis, 1998](#)) and GIST ([Nguyen et al, 2012](#)), Grand Canonical Monte Carlo ([Ross et al,](#)

2015), 3D-RISM (Beglov, 1997; Kovalenko, 1998), JAWS (Michel et al, 2009b) are rigorous and fairly accurate and have been used to great effect, although most of them currently suffer from being computationally costly. To complement these efforts, rapid machine learning (ML) based approaches, such as Consolv (Raymer et al, 1997), WaterScore (Garcia-Sosa et al, 2003), Waterdock (Ross et al, 2012) have been developed to balance accuracy and speed. The drawback of these ML methods is that the distribution of the features (e.g. B-value, number of solute-solvent hydrogen bonds, the density, hydrophilicity of the vicinal protein atoms, solvent contact surface area and total hydrogen bond energy, etc.), that are used to distinguish conserved from displaceable waters, overlap to a large extent between these two classes of interfacial waters (Garcia-Sosa et al, 2003; Raymer et al, 1997). Additionally, given that the displaceability of a water is very intimately coupled to the local chemical environment, a general equation to discriminate waters seems unlikely. These factors limit the maximum possible accuracy that can be achieved using the ML approaches. Furthermore, the dependence on the temperature factor, as in the case of Consolv (Raymer et al, 1997), renders it unusable for predicting hydration sites. Finally, it appears that the training and testing datasets used in these ML approaches are not optimal: in general, a water molecule is considered conserved if it is present in the apo and in the ligand bound forms of the protein or is considered displaceable otherwise. However, there might be some waters that are categorized as conserved (according to this definition) but which can actually be displaced with an optimal ligand modulation.

There are 4 factors that limit a more regular yet rigorous water analysis. The software/packages available are either not free, or are computationally intensive, or have limited accuracy, or else are not comprehensive in that the properties they provide are limited. In an attempt to overcome these limitations, here, I present a software package ('JAL') that can complement these various approaches. 'JAL' consists of reliable and user-friendly python programs that enable the computation of solvation properties (solvent density, residence times, buried waters, favorable/unfavorable solvent interactions with solute, closest-water distance measures) in a structural ensemble generated from the MD simulations. Additionally, a novel knowledge-based algorithm is presented to address what interfacial water molecules are to be displaced or retained. As has rightly been pointed out, computing the probability of favorably displacing an interfacial water is a regression problem and not a classification problem ([Ross et al, 2012](#)), and here I attempt to provide an accurate, albeit qualitative and relative estimate, of displaceability of interfacial waters. The approach uses a structure-based search to mine the PDB for chemical environments that are similar to that around the interfacial water molecule whose displaceability is of interest. The utility of 'JAL' is demonstrated by using examples of the p53 DBD, eIF4E, Scytalone dehydratase and EGFR kinase.

4.2 Methods

Figure 4.1: The flowchart and hydration metrics embedded in the software package ‘JAL’; ‘COM’ refers to the Center of Mass.

4.2.1 Hydration properties

The hydration programs developed here enable a rigorous and rapid characterization of several hydration properties in MD simulation trajectories using

the metrics described (Figure 4.1). The program takes as input a single file in PDB format consisting of frames extracted from an MD trajectory. The *mdtraj* module (McGibbon et al, 2015) of python can be used to convert trajectories in diverse formats to the required PDB format. The input file is first preprocessed ('preprocess.py') by removing the hydrogen atoms and saving the coordinates of solute (user defined or all atoms based on the user input) along with the water molecules in the 1st solvation shell which is defined as all solvent molecules within 3.5 Å (or as specified by the user) of any solute heavy atom. This reduces the trajectory size, speeding up subsequent computations. Solvent properties (solvent density, residence time, favorable-unfavorable solute-solvent interactions, closest water distance) are computed by tracking water molecules within a sphere of 3.5 Å from the reference (non-hydrogen protein) atoms in each snapshot of the input structural ensemble. This approach of computing the solvent properties is different from a grid based approach. A grid based approach requires creating a grid over the water box and then counting the number of solvent molecules within each grid box (usually of 0.5 Å dimension); I use a protein centric approach where the protein (non-hydrogen) atoms that interact with the solvent (water oxygen) are monitored. Although, this approach is computationally more expensive than the grid based approach, this method provides insightful representations of the data as discussed in the results section. This also makes sense, in that the solvation around a protein is a function of the chemistry and topology of the protein surface. Representing solvent density as grid boxes in 3D (as represented for example, by the '*grid*' function in Amber's ptraj module) obscures the identity of the solute atoms that are

critical for the respective solvent densities around the solute; current approach does not suffer from this. These hydration programs produce 2 kinds of output for analysis - graphical and textual. The detailed nature of the output and the associated analyses will be discussed in the results section.

The hydration programs embedded in the ‘JAL’ package compute the following solvation properties:

1. Solvent density: The number of times a water oxygen is observed within a distance of 3.5 Å from the reference atom. The reference atom can be input by the user or can be all the non-hydrogen solute atoms. It is measured as percent solvation across the ensemble and is computed using the following formulation-

$$S = \sum_{t=0}^{100ns} \gamma(W(t), W(t + \Delta t)) \quad \dots (4.1)$$

‘W’ is the any water oxygen (irrespective of the ID) at a given instance of time and ‘ Δt ’ is the time interval between two successive snapshots of the MD trajectory. Here, Δt is 100 ps. Value of ‘ γ ’ is 1 if $W(t) = W(t + \Delta t)$ and 0 otherwise.

2. Residence time: This measure involves tracking the identity of water molecules within a 3.5 Å sphere around the reference atom. The program calculates 3 values for the residence times of water - total, average and maximum residence times. It is calculated as suggested previously ([De Simone et al, 2005](#)) and is an estimate of how stable/trapped a water is at a particular site.

$$RT = \sum_{t=0}^{100ns} \gamma(W(t), W(t + \Delta t)) \quad \dots (4.2)$$

‘ W ’ is the ID of water oxygen at a given instance of time and ‘ Δt ’ is the time interval between two successive snapshots of the MD trajectory. Here, Δt is 100 ps. Value of ‘ γ ’ is 1 if $W(t) = W(t + \Delta t)$ and 0 otherwise.

3. Distance of closest water: It computes the distance between a solute atom and the closest solvent molecule. The metric estimates the accessibility of a site on the solute to the water and hence provides a likely estimate of the strength of the interaction.
4. Buried waters: Some water molecules in the vicinity of a protein interact very strongly with the protein residues and hence reside longer before diffusing into the bulk solvent ([Levitt & Park, 1993](#)). Such waters are referred to as bound waters. Bound waters make up to 20-60% by weight of a hydrated protein ([Csermely, 2001](#)). The core of the protein is usually well packed but can occasionally contain cavities. Hydration of the cavity depends on the cavity volume and residues lining the cavity. These waters, which are usually found in the protein cavities and core, have little solvent accessibility and are referred to as buried waters. Buried waters are computed using a novel methodology based on iterative calculations of Solvent Accessible Surface Area (SASA) of water molecules and deleting those waters which have SASA more than zero after every iteration (‘*evaporation*’). Ultimately waters left with SASA of zero are the ones which are present in the protein interior, and are thus deemed to be buried. SASA is computed using the rolling ball algorithm ([Shrake & Rupley, 1973](#)). The algorithm is shown below ([Algorithm 4.1](#)). The calculations of buried waters are benchmarked using bovine trypsin inhibitor (PDB ID 4PTI,

resolution 1.5 Å, (Marquart, 1983)). The program correctly identifies the 4 buried waters (Figure 4.2); their B-values are very low (labeled in the figure) and they engage in a strong interaction network with the protein.

Algorithm 4.1: Computing buried waters in an MD trajectory of a protein

Input: MD trajectory of a protein

Output: buried waters ($SASA \leftarrow 0$)

```
1  For each frame  $f$  in the input MD trajectory
2      Compute SASA using the ‘Rolling Ball Algorithm’
3       $fs \leftarrow \text{size}(\text{frame})$ 
4       $temp\_file \leftarrow ''$ 
5      For each non-hydrogen atom  $nh$  in  $f$ 
6          If  $nh == \text{protein atom } pa$ :
7               $temp\_file.write(pa)$ 
8          If  $nh == \text{water oxygen atom } wo$ :
9              If  $SASA(wo) > 0$ :
10                 continue
11             Else:
12                  $temp\_file.write(wo)$ 
13      End For
14      Open ( $temp\_file$ )
15      If  $fs == \text{size}(temp\_file)$ :
16          Output buried waters  $bwo$  in frame  $f$ 
17      Else:
18          Go to line 5
19 End For
```

Figure 4.2: Program to identify buried waters is benchmarked against bovine pancreatic trypsin inhibitor (PDB ID- 4PTI chain A, 1.5 Å resolution (Marquart, 1983)). Protein is shown in transparent surface colored by atom types. Program correctly identifies the four buried waters (labeled and colored in black). B-values of these waters are shown in brackets.

5. Role of water in protein stabilization/destabilization: Hydration of the backbone charges in loops/turns/coils in proteins contributes to structural destabilization, while hydration of backbone charges in helices/sheets is classified as stabilizing (Park & Saven, 2005). The interaction valencies of protein backbone charges in loops/turns/coils are usually unsatisfied but are satisfied for the helices/sheets. Thus, hydration of the former would fulfill the interaction valencies and imply stabilization of the structure, while hydration of the latter would likely perturb the secondary structure and induce destabilization. Based on this theory, the programs track the solvent

interactions with protein backbone atoms (using the same interaction cut off of 3.5 Å) and quantitatively characterize the stability of regions in the protein structure; this is output as percent stabilization/destabilization of each residue; the algorithm is outlined below (Algorithm 4.2).

Algorithm 4.2: Computing favorable/unfavorable solute/solvent interactions

Input: MD trajectory of a protein

Output: favorable/unfavorable solute/solvent interactions

```

1      Compute the secondary structure of the 1st frame of the input MD
trajectory
2       $fav, unfav \leftarrow 0$ 
3      For each frame  $f$  in the input MD trajectory
4          For each backbone atom – N and O  $bck\_no$  in protein
5              If  $sec\_str(residue\ i) == [\alpha\text{-helix} \mid \beta\text{-strand}]$ :
6                   $unfav \leftarrow unfav + 1$ 
7              Else if  $sec\_str(residue\ i) == [loop \mid turn \mid coil]$ :
8                   $fav \leftarrow fav + 1$ 
9          End For
10     End For
11      $fav\_count \leftarrow (fav / (fav + unfav)) * 100$ 
12      $unfav\_count \leftarrow (unfav / (fav + unfav)) * 100$ 
13     Print  $fav\_count, unfav\_count$ , regions of protein contributing to
 $fav\_count$  and  $unfav\_count$ 

```

6. Displacing waters using a novel structure based search: A critical question that still eludes a robust solution is that given a water molecule in a binding site,

should it be displaced or not, when modeling the binding of a ligand. Attempts have been made based on rigorous computations of desolvation free energies (Abel et al, 2008; Bayden et al, 2015; Beglov, 1997; Goodford, 1985; Huggins & Payne, 2013; Kovalenko, 1998; Lazaridis, 1998; Michel et al, 2009b; Nguyen et al, 2012; Ross et al, 2015; Young et al, 2007). Here, I propose a somewhat different approach which is guided by the available structural data in the PDB database. The topology independent CLICK method (Nguyen & Madhusudhan, 2011; Nguyen & Verma, 2015) is used to mine the protein structures in the PDB for a match to the exact and immediate local environment around a water molecule that one is interested in displacing. If similar chemical environments are found, then it is likely that the water is important and should not be displaced. On the other hand, if the position of the water has been found to be replaced by some protein atom or heteroatom, then it is likely that the water can be replaced and the identity of the atom or heteroatom can be used to modulate and optimize a ligand to displace waters. Given that a water molecule can engage in a maximum of four hydrogen bonds at any given instance of time, a local environment is defined around the water molecule (to be displaced) as the four closest atoms that lie within a distance of 3.5 Å from it. It is observed that this cut-off distance yielded results with 100% structure overlap (SO). SO is defined as the percentage of the representative atoms in the structure that are within 3.5 Å of the corresponding atoms in the superimposed structure (Nguyen & Madhusudhan, 2011). This immediate and exact local milieu (defined using 3.5 Å cut off) represents the chemistry and topology

surrounding the water. Additionally, the identities of atoms constituting the local environment were matched to the exact atomic identities from other structures in the database. Such an atomic identity based search helps maintain the chemical environment. For instance, an 'NE1' nitrogen atom (present only in Tryptophan) is matched exclusively to 'NE1' atoms. If on the other hand, the local environment involves hetero atoms, then the exact atom type is ignored and only the elements are matched (a nitrogen atom is matched to any nitrogen). While this current study examines the complete structural database (PDB), a non-redundant dataset of PDB (NR95) containing unique structures (based on sequence identity cut off of 95%) totaling 32,700 is also provided for download that will enable a faster search. Similarly, NR96 (33,513 structures), NR97 (34,810 structures), NR98 (37,512 structures) and NR99 (45,717 structures) databases are made available for similar searches. The algorithm followed is outlined below ([Algorithm 4.3](#)).

Algorithm 4.3: Computing displaceability of interfacial waters

Input: Input coordinates of interfacial water and surrounding atoms within 3.5 Å of it

Output: Percent occupancy of interfacial water with different atom types

```
1      Input the interfacial water  $c$  and the 4 closest non-hydrogen atoms  
       $cm$  within 3.5 Å of it  
2      For each structure  $s$  in the PDB database  
3          Run CLICK algorithm to find the site structurally similar to  
       $cm$  using Structural Overlap ( $SO$ ) and  $RMSD$   
4          If  $SO == 100\%$  and  $RMSD < RMSD\_cutoff$ :  
5               $List\_of\_Hits.append(s)$   
6      End For  
7      For each atom  $i$  in the  $List\_of\_Hits$   
8          If distance ( $i, c$ ) < 1 Å:  
9              Save atom_type  $at$   
10         Else:  
11             continue  
12     Print atom_types  $at$ 
```

4.2.2 MD simulations

I performed all-atom explicit solvent MD simulations of the WT and the V143A mutant of the p53 DBD (for the WT, chain A of PDB ID 2AHI resolved at 1.85 Å, (Kitayner et al, 2006))) and of the translational initiation factor eIF4E (chain B of PDB ID 2W97 resolved at 2.29 Å, (Brown et al, 2009b)) in their apo states. Mutants were generated using the mutatgenesis plugin of PyMOL

(Schrodinger, 2010). Simulations were carried out using the Amber ff99SB force field (Hornak et al, 2006). The N and C termini of the proteins were capped with ACE and NME moieties respectively. All the crystallographic waters were retained. Proteins were solvated in a cubic water box with a minimum separation of 10 Å between the protein atoms and the edge of the water box. TIP3P (Jorgensen et al, 1983a) water model was used to model water molecules. This was followed by neutralization of the systems with the addition of an appropriate number of counter ions. The solute-solvent systems were subjected to an energy minimization involving 1000 steps each of steepest descent followed by an equal number of conjugate gradient steps. After this, the systems were heated to 310 K in an NVT ensemble followed by 2ns of equilibration in the NPT ensemble. Finally, a 100 ns production run in NPT was carried out for each system using the GPU implementation of PMEMD (Gotz et al, 2012). A cut off of 8 Å was used to evaluate the Lennard-Jones (LJ) and short-range electrostatic interactions, while the long-range electrostatic interactions were evaluated using the particle-mesh Ewald (PME) method with a grid spacing of 1 Å and a fourth-order interpolation (Darden et al, 1993b). Langevin thermostat (Loncharich et al, 1992) was used to regulate the temperature. A constant pressure (1.0 bar) was maintained during the simulations using a weak coupling with a relaxation time of 1 ps. SHAKE algorithm (Ryckaert et al, 1977a) was used to constrain the bond lengths of all covalent bonds involving the hydrogen atoms, thus enabling an integration time step of 2 fs; coordinates were saved at an interval of 5 ps. Hydration analysis is carried out on 1000 evenly spaced snapshots extracted from the 100 ns MD

trajectories; thus the time interval between consecutive trajectory snapshots is 100ps; this can be changed by the user. The appropriate values for these parameters are dependent on the protein system being studied and the purpose of these calculations. Simulations were performed in triplicates.

4.2.3 Buried water clusters

PDB- ID_CHAIN- ID	Resolution (Å)	PDB- ID_CHAIN- ID	Resolution (Å)	PDB- ID_CHAIN- ID	Resolution (Å)
1KZY_A	2.5	2BIM_B	1.98	3D0A_C	1.8
1KZY_B	2.5	2J1W_A	1.8	3D0A_D	1.8
1TSR_A	2.2	2J1W_B	1.8	3KMD_A	2.15
1TSR_B	2.2	2J1X_A	1.65	3KMD_B	2.15
1TSR_C	2.2	2J1X_B	1.65	3KMD_C	2.15
1TUP_A	2.2	2J1Z_A	1.8	3KMD_D	2.15
1TUP_B	2.2	2J1Z_B	1.8	3Q01_A	2.1
1TUP_C	2.2	2J20_A	1.8	3Q01_B	2.1
1UOL_A	1.9	2J20_B	1.8	3Q05_A	2.4
1UOL_B	1.9	2OCJ_A	2.05	3Q05_B	2.4
1YCS_A	2.2	2OCJ_B	2.05	3Q05_C	2.4
2AC0_A	1.8	2OCJ_C	2.05	3Q05_D	2.4
2AC0_B	1.8	2OCJ_D	2.05	3ZME_A	1.35
2AC0_C	1.8	2PCX_A	1.54	3ZME_B	1.35
2AC0_D	1.8	2VUK_A	1.5	4AGL_A	1.7
2ADY_A	2.5	2VUK_B	1.5	4AGL_B	1.7
2ADY_B	2.5	2WGX_A	1.75	4AGM_A	1.52
2AHI_A	1.85	2WGX_B	1.75	4AGM_B	1.52
2AHI_B	1.85	2X0U_A	1.6	4AGN_A	1.6
2AHI_C	1.85	2X0U_B	1.6	4AGN_B	1.6
2AHI_D	1.85	2X0W_A	2.1	4AGO_A	1.45
2ATA_A	2.2	2X0W_B	2.1	4AGO_B	1.45
2ATA_B	2.2	2XWR_A (TEMPLATE)	1.68	4AGP_A	1.5
2ATA_C	2.2	2XWR_B	1.68	4AGP_B	1.5
2ATA_D	2.2	3D0A_A	1.8	4AGQ_A	1.42
2BIM_A	1.98	3D0A_B	1.8	4AGQ_B	1.42
4HJE_A	1.91	4IBT_C	1.7	4KVP_B	1.5

4HJE_B	1.91	4IBT_D	1.7	4KVP_C	1.5
4HJE_C	1.91	4IBU_A	1.7	4KVP_D	1.5
4HJE_D	1.91	4IBU_B	1.7	4LO9_A	2.5
4IBQ_A	1.8	4IBU_C	1.7	4LO9_B	2.5
4IBQ_B	1.8	4IBU_D	1.7	4LO9_C	2.5
4IBQ_C	1.8	4IBY_A	1.45	4LO9_D	2.5
4IBQ_D	1.8	4IBY_B	1.45	4LOE_A	1.85
4IBS_A	1.78	4IBZ_A	1.92	4LOE_B	1.85
4IBS_B	1.78	4IBZ_B	1.92	4LOE_C	1.85
4IBS_C	1.78	4IBZ_C	1.92	4LOE_D	1.85
4IBS_D	1.78	4IBZ_D	1.92	4LOF_A	2
4IBT_A	1.7	4IJT_A	1.78	4MZI_A	1.25
4IBT_B	1.7	4KVP_A	1.5		

Table 4.1: Crystal structures of p53 DBD used for superimposition. Structures were superimposed onto 2XWR chain A since it is solved with an extended N-terminus (residues 91 onwards, refer [section 3.3.2](#)).

To demonstrate the utility of the program to compute buried waters, high quality crystal structures (resolution ≤ 2.5 Å) of p53 DBD from the PDB were used. Multiple chains of the DBD within a PDB entry were considered as separate structures. This resulted in a total of 119 unique structures ([Table 4.1](#)). Buried waters were computed in each of the structures. All the non-buried waters were removed. These structures of the p53 DBD with buried waters retained were then superimposed using the ‘align’ function in PyMOL ([Schrodinger, 2010](#)). The structure of the p53 DBD with an extended N-terminus resolved was considered as the template (chain A of PDB ID 2XWR, resolved at 1.68 Å, ([Natan et al, 2011](#))) and all the other chains were superimposed on it. It is observed that many structures of p53 DBD are well-hydrated but vary in the position of buried waters and therefore a cluster of crystallographic buried water molecules is identified as the one which contains water molecules in at least 50% of the superimposed

structures (Figure 4.3A). Protein structures being dynamic in MD simulations, the buried water clusters in the MD simulations are identified as the ones which are present in at least 5% of the simulation (Figure 4.3C); this was determined by the minimum time required to reproduce the crystallographically observed water cluster.

4.3 Results and Discussion:

4.3.1 Case study 1: The role of buried waters in the stability of the p53 DNA Binding Domain

To evaluate the utility of identifying buried waters in an MD trajectory using ‘JAL’, the role of buried waters in the DNA Binding domain (DBD) of p53 was investigated. The DBD interacts with DNA and other proteins, and modulates the transcription of downstream genes involved in regulating cell cycle processes. The DBD is often mutated in tumor cells, thus, making p53 an important tumor suppressor (Brown et al, 2011a; Brown et al, 2009a; Schuijjer & Berns, 2003; Vogelstein et al, 2000). A subset of these mutations is known to lower the melting temperature of the protein causing subsequent unfolding and aggregation (Xu et al, 2011). Although, there have been many studies on p53 DBD, few have concentrated on the role of hydration, specifically of any buried waters. Experiments report that ligand (DNA) binding & dehydration in p53 should stabilize the native folded state, although the underlying molecular mechanisms remain unknown (Silva et al, 2010). Reports suggest that the hydration shell, which has been decomposed into “tight” and “dynamical” water molecules, makes

the p53 DBD structure more labile (Xu et al, 2011). The marginal stability of the DBD at physiological temperatures makes it a good model to study the role of buried waters. Here, I investigate the role of buried waters in the stability of the WT and the V143A mutant of p53 DBD. V143A is one of the high frequency mutations observed in tumors that are known to induce destabilization and unfolding in the DBD under physiological conditions.

Figure 4.3: (A) Clusters of buried waters (colored according to the B-value: a color gradient of blue-white-red denotes low to high B-values) present in at least 50% of the 119 superimposed crystal structures of p53 DBD. (B) Distribution of the number of buried waters in simulations (triplicates combined) of the WT and V143A mutant form. (C) Clusters of buried waters, shown as cyan (for WT) and pink (for V143A) spheres, as present in at least 5% of the MD simulations of WT and V143A (blue and red cartoon) p53 DBD respectively. Mutation site (V143A) is shown with yellow sticks. Strand colored in green represents the Aggregation Prone Region (APR). A buried water cluster marked with dashed line is the prominent difference between the two forms of the protein since it is present almost exclusively in WT.

Most of the buried water clusters in the crystal structures of the p53 DBD are present around loops L2, L3 and strand S7 and have low B-values, thereby suggesting that they are tightly bound (Figure 4.3A). The MD simulations of p53 DBD carried out here suggest that the V143A mutant undergoes a reduction in the number of buried waters as compared to the WT DBD (Figure 4.3B). It can be seen that the positions of buried water clusters in the simulations of WT and V143A DBD are quite different (Figure 4.3C). The positions of buried waters in the MD simulations of the WT DBD, but not the V143A mutant form of DBD, correspond well to the positions of the crystallographic buried waters. It appears that in the V143A mutant of DBD there are fewer buried water clusters around the loops L2 and L3 as compared to WT DBD (shown with dashed line in Figure 4.3C). Therefore, I investigate if the dehydration of the loops L2 and L3 imply loss of stability in the DBD. The importance of buried waters for the stability of protein structures has been abundantly suggested in the literature. For instance, it has been suggested that buried waters fulfill the unsatisfied interaction valencies of the protein atoms through water mediated interaction networks (Park & Saven, 2005). Alternatively, hydration of a hydrophobic cavity in the protein core may prevent its collapse and help a protein retain its native structure (Das et al, 2013). Moreover, protein hydration may also aid protein flexibility and hence contribute to entropic stabilization (Csermely, 2001; Fischer & Verma, 1999).

Figure 4.4: Buried waters (green spheres) in a snapshot from simulations of (A) WT and (B) the V143A mutant form of DBD. It appears that the native conformation of loop L3 is stabilized by the buried water in WT, while this water is lost when the loop L3 assumes a non-native conformation in one of the three replicates of the V143A mutant. Arrows in (A) and (B) point to loop L3. (C) Crystallographic structures of p53 DBD from PDB confirm the presence of this water molecule. A cluster of water molecules (red spheres) at this position is present in 100 out of a total of 119 chains of p53 DBD that were superimposed (Structure with PDB ID- 2XWR_A is shown). The average B-values of water molecules contributing to these clusters are shown in brackets. Lines in cyan represent hydrogen bonds. Loops L2 and L3 are shown in pink and green respectively.

A closer inspection of the positions of the buried waters in the DBD reveals a differential conformational sampling of the loop L3 between the two forms of the protein (Figure 4.4A and 4.4B). The loop L3 in V143A samples a non-native like

conformation in one of the three replicates. It appears that the crystal-like native conformation of loop L3 in the WT DBD is stabilized by a buried water molecule, which is lost when the mutant DBD assumes a non-native conformation. This buried water makes interactions with the backbone of loop L2 (oxygen of R174) and loop L3 (nitrogen of M246), as well as the side chain of C238, which is also coordinated with Zinc (Figure 4.4C). The total residence time of a water molecule at this position is 73.33 (± 9.6) ns for the WT, while it is only 34 (± 20.5) ns for V143A across triplicates of MD simulations, each with a total sampling time of 100ns. Water at this position is also conserved in the crystal structures. Out of the 119 chains (Table 4.1) of p53 DBD that were superimposed, buried water at this position in loop L3 is present in 100 structures (Figure 4.4C). The average B-value of the waters contributing to this cluster is $20 \pm 8 \text{ \AA}^2$, implying that this water is indeed tightly bound to the protein (Eric Martz, 2014).

Figure 4.5: RMSF from the MD simulations (in triplicates) of the WT with charges on atoms (Arg174(O), Met246(N) and Cys238(SG)) interacting with the buried water in loop L3 ON and OFF. The perturbations in the conformations of loops (A) L2 and (B) L3, when the charges are switched OFF are quite dramatic as compared to when the charges are left ON.

I ran another set of simulations of WT and V143A by removing all the crystal waters. In this case, a water molecule from the bulk solvent is observed to occupy this position almost instantaneously, clearly suggesting that this site prefers to be hydrated. Additionally, a control simulation of WT DBD is run by removing the charges on the 3 atoms (R174(O), M246(N) and C238(SG)) that were observed to form hydrogen bonds with this water. The rationale behind this was that

removing the charges on these atoms would displace the water from its position and this would allow us to examine the effect of this buried water on the conformation of loop L3. Simulations show that switching off the charges indeed displaces this water and has a drastic effect on the conformations of loops L2 and L3 which is quantified using the RMSF of residues in these loops ([Figure 4.5A and 4.5B](#)). When this water molecule is displaced, it increases the flexibility in the region of the loop L2 that follows helix H1, while decreases the flexibility in the region of L2 that precedes helix H1 as compared to the WT DBD. Similarly, displacing this water also increases the flexibility of the residues in loop L3.

Figure 4.6: Density of the Zinc atom, shown as a black grid, across triplicate simulations of p53 DBD with charges on atoms (R174(O), M246(N) and C238(SG)) interacting with the buried water in L3, ON (A, B, C) and OFF (D, E, F). Zinc is shown as sphere and colored green and orange in these simulations respectively. It can be observed that the density of Zinc in D, E and F is a little more diffused as compared to A, B and C.

Given the fact that most of the hot spot mutations are located in loops L2 and L3 ([Brown et al, 2009a](#); [Bullock & Fersht, 2001](#); [Joerger & Fersht, 2008](#)), the

change in the dynamics of these regions resulting from the loss of the buried water molecule could have implications in tumorigenesis. The increased flexibility also correlates well with the available experimental data on the NMR of the destabilizing mutants including, V143A, which suggests that residues in the loops L2 and L3 regions are affected differently between the WT and the destabilizing mutants (Wong et al, 1999). However, the role of this buried water molecule in the conformation of loop L3 and hence in the stability of p53 DBD has not been reported and needs to be critically examined. One of the ways in which the buried water molecule in L3 might be related to the structure and function of p53 DBD is through regulation of zinc binding. Zinc is coordinated to residues in the loops L2 and L3 through C176, H179, C238 and C242. Zinc binding is known to stabilize the DBD and the loss of zinc is reported to be a major factor in the destabilization of p53 DBD (Blanden et al, 2015; Meplan et al, 2000; Yu et al, 2014). Simulations suggest that the loss of this water molecule enhances the flexibility of the zinc binding region. The density of zinc across simulations is computed to measure the perturbations in the position of zinc (Figure 4.6). It can be observed that when the buried water molecule is displaced (by switching off the charges on the atoms coordinated to the buried water molecule in loop L3), the zinc density is a little more diffuse (Figure 4.6 D, E, F) as compared to the case when the buried water molecule is retained (Figure 4.6 A, B, C). A diffused zinc density could suggest that zinc binding is weakened, which in turn could decrease the stability of the p53 DBD. However, the models include zinc covalently bonded (rigorous non-bonded models for zinc coordination are not yet available) and hence I can only speculate.

4.3.2 Case study 2: Residence times of waters at the eIF4E-mRNA cap binding interface reveals a water bridge

Figure 4.7: mRNA cap binding site of eIF4E is colored according to (A) Solvent density (eqn. 4.1) and (B, C) residence times (eqn. 4.2) with a color gradient from blue to red regions representing higher to lower values of these respective hydration properties. W1 and W2 (green spheres) positions show markedly high residence times (distinct blue patches on the surface). Residues shown in sticks are W166 and N155 that form the 2 ends of the water bridge. (C) mRNA cap, m7GTP, is shown for reference.

Residence times of water often provides crucial insights into protein structure and function (Makarov et al, 2000; Prakash et al, 2012). The residence

times of waters in eIF4E, a translation initiation factor which binds to the 5' mRNA cap (m⁷GTP) and initiates a cascade of events that lead to the loading of the mRNA complex onto the ribosome, were examined in detail. The mRNA cap binding interface of eIF4E in the cap-unbound state is very well hydrated (Brown et al, 2007; Brown et al, 2009b). Recently, an intricate network of 14 conserved water molecules was identified that mediates the interactions between eIF4E and mRNA cap [manuscript under review]. Central to this network is a water bridge involving two water molecules that link the side chains of W166 and N155. A subset of these waters was also reported to be present in the cap-unbound state. This water bridge has been hypothesized to be a crucial recognition motif for binding of the mRNA cap. The solvent density and the residence times of water around eIF4E are computed by tracking water interactions with all the respective non-hydrogen protein atoms within a sphere of radius 3.5 Å. Figures 4.7A and 4.7B show solvent density and residence time values mapped onto the protein. Solvent density clearly shows that the mRNA cap binding site of eIF4E is indeed well hydrated. However, it is apparent that not many sites in the binding pocket have high residence times. Although the mRNA binding site of eIF4E is well solvated, it appears that most of the solvent molecules are rapidly exchanging with bulk solvent. But two sites in the binding pocket with extremely high residence times (seen as blue patches on the surface) are seen. The two sites with high residence times correspond to the waters involved in forming the water bridge (labeled as W1 and W2 in figure 4.7B). This water bridge motif appears to be important for mRNA cap recognition by eIF4E (Figure 4.7C) and might play a

crucial role in the kinetics and thermodynamics of the process. This example demonstrates how mapping the residence times hold the potential to help identify water bridges in proteins. The ‘*hbond*’ function in the cpptraj module of AmberTools15 ([D.A. Case, 2015](#)) yields information that needs to be comprehensively integrated to extract such details, while ‘JAL’ provides this in a straightforward manner.

4.3.3 Case study 3: Structure based search to displace interfacial waters in Scytalone dehydratase, EGFR kinase and eIF4E

Displacing waters from a protein-ligand binding pocket has a significant impact on the ligand binding affinity. Several rigorous free energy computations have been proposed for computing the associated energetics ([Abel et al, 2008](#); [Bayden et al, 2015](#); [Beglov, 1997](#); [Goodford, 1985](#); [Huggins & Payne, 2013](#); [Kovalenko, 1998](#); [Lazaridis, 1998](#); [Michel et al, 2009b](#); [Nguyen et al, 2012](#); [Ross et al, 2015](#); [Young et al, 2007](#)). The possibility of using a structure based search to gauge the feasibility of favorably displacing waters in ligand binding pockets of proteins (detailed algorithm and protocol in [Methods section 4.2](#)) is explored now. For this purpose, the ligand binding pockets of Scytalone dehydratase ([PDB ID- 5STD, 1.95 Å, \(Wawrzak et al, 1999\)](#)) and EGFR kinase ([PDB ID- 4HJO, 2.75 Å, \(Park et al, 2012\)](#)) are explored together with evaluation of the displaceability of the interfacial waters W537 (has a B-value of 14.45 Å) in Scytalone dehydratase and W1104 (has a B-value of 42.31 Å) in EGFR kinase respectively. The rationale behind choosing these proteins and these particular waters is the availability of experimental ligand binding affinity data, with the ligands displacing these

particular water molecules, to benchmark the present calculations. In Scytalone dehydratase, the ligand 3-cyanocinnoline used to displace W537 is an inhibitor designed to reduce fungal infections in rice plants ([Chen et al, 1998](#)); in EGFR kinase, the ligand 3-cyanoquinoline used to displace water W1104 is an anticancer agent ([Wissner et al, 2000](#)). Despite the fact that both ligands were modified with the same chemical group, a cyano group ($-C\equiv N$), to displace W537 and W1104, the observed effects were found to be very different. In Scytalone dehydratase, displacing W537 by the cyano group from the ligand lead to an increase in the binding affinity by 30-fold ([Chen et al, 1998](#)), while in the EGFR kinase, displacing W1104 by the cyano group lead to a decrease in the binding affinity by 3-fold ([Wissner et al, 2000](#)). Rigorous free energy calculations using Free Energy Perturbations (FEP) were reported to explain how displacing these interfacial waters with the same ligand functional groups can have a very different effect on the binding affinity ([Michel et al, 2009a](#)). This demonstrates that the binding affinity change upon displacing a water molecule not only depends on the functional group that displaces it but also on the chemical environment the water molecule is in. This motivated a structure based search of the local chemical environment of the interfacial water molecule that is to be displaced.

Figure 4.8: Protocol of structure based search used to displace interfacial waters (Wat537 and Wat1104) from the ligand binding sites of (A) Scytalone dehydratase (PDB ID- 5STD, 1.95 Å, (Wawrzak et al, 1999)) and (B) EGFR kinase (PDB ID- 4HJO, 2.75 Å, (Park et al, 2012)). It can be broadly categorized into 3steps- defining a local chemical environment within a distance of 3.5 Å around the interfacial water molecule, using the 4 closest atoms within this distance cut off for a structure based search across the entire PDB and setting up a screening criteria (of structural overlap and RMSD) to filter out the top hits. The hits are then analyzed to identify the atoms from the structures that overlap with the interfacial water molecules in the query structure.

The 4 closest non-hydrogen atoms within a sphere of 3.5 Å from the interfacial water molecules W537 and W1104, to define their local chemical environments (Figure 4.8), were chosen. The entire PDB (Berman et al, 2000) was then searched for structures that consisted of similar chemical environments. A

screening criterion is then set to filter the hits for 100% structural overlap and an arbitrary RMSD cut off (Figure 4.8). The absolute values of the structural overlap and the RMSD are tuned to extract a sufficient number of hits that have the same chemistry and topology as the region surrounding the query interfacial water (here W537 and W1104). This resulted in 4346 and 4037 hits for Scytalone dehydratase and EGFR kinase respectively. Subsequently, the percentage occupancy of the different atom types (from the structures in the hits) that overlap with W537 and W1104 were calculated (Table 4.2); atom types of amino acid atoms as defined previously were used (Wendy D. Cornell 1995).

The data suggests that the chance of observing a water molecule overlapping with the position of W537 (in Scytalone dehydratase) and W1104 (in EGFR kinase) is ~6% and ~4% respectively. A low occupancy of water suggests that a water molecule may not be energetically favorable in the chemical environment that these waters are in and so displacing them with a ligand group is a promising option.

This poses a question: Can the percentage of observing different atom types that overlap with W537 and W1104 in these searches provide insights into which ligand groups are likely to favorably displace these interfacial waters? The chance of observing a void (no atom found to overlap), given the local chemical environment that W537 and W1104 are in, is quite high at 91% and 72% respectively. A very high occupancy of void suggests that the chemical group from the ligand designed to favorably displace W537 and W1104 should have an atom type that is not present in proteins. This indeed agrees well with experiments

demonstrating that a nitrogen atom from a cyano group (an atom type not present in proteins) can displace these waters. Furthermore, comparing the occupancy of different atom types across W537 and W1104 hits suggests that there is a relatively higher chance of observing any sp^3 hybridized carbon atom (CT) and an sp^2 hybridized nitrogen atom in amides (N) at the position of W1104 (13.5% and 2% respectively) as compared to W537 (0.1% and 0% respectively). Given that displacing W1104 with the nitrogen atom of the cyano group is mildly destabilizing (3-fold decrease in the binding affinity), displacing W1104 with ‘CT’ or ‘N’ atom type from the ligand might be worth investigating. Additionally, a structure based search of the local chemical environments of W537 and W1104 also reported hits where ligand atoms (non-protein hetero atoms) were found to overlap with the positions of the waters ([Table 4.2](#)). Thus, the structure based search can also identify fragments with hetero atoms from matching hits that may potentially displace waters W537 and W1107.

Despite its potential, how feasible is the structure based search of the local chemical environment against the entire PDB? A structure based search of the same closest 4 atoms (within 3.5 Å of W537 and W1104) was carried out against a smaller unique dataset of proteins from the PDB, the NR95 dataset (see [Methods section 4.2](#)). The NR95 dataset is almost 1/9th the size of entire PDB with a maximum sequence identity of 95% between any two proteins within the dataset. NR95 was primarily designed to enable a faster search even on a local computer. The occupancies of atom types that overlap with W537 and W1104 as computed from the search across NR95 is compared to the occupancies computed from the

search against the entire PDB (Table 4.2). The search of the local chemical environments (closest 4 atoms within 3.5 Å of W537 and W1104) against NR95 produced a total of 391 and 421 hits for Scytalone dehydratase and EGFR kinase with an RMSD cut off of ≤ 0.3 and ≤ 0.45 respectively. Thus, the occupancy of overlapping atom types from database searches across PDB and NR95 indeed match very well (correlation coefficient (r) of 0.99 for both W537 and W1104) and thus, should enable similar interpretations about the displaceability of the interfacial water molecules, but at a faster rate.

Element	Atom types	Scytalone dehydratase (W537)		EGFR kinase (W1104)	
		Entire PDB	NR95	Entire PDB	NR95
Carbon	CT	0.1	0	13.5	13.6
	C	0	0	0.5	0.9
	CC	0	0	0	0
	CW	0	0	0	0
	CR	0	0	0	0
	CA	0	0	0	0
	C*	0	0	0	0
	CN	0	0	0	0
	CB	0	0	0	0
Hydrogen	H	0	0	0	0.2
	H1	0.1	0.2	0	0.2
	HC	0	0	0.1	0.2
	H4	0	0	0	0
	H5	0	0	0	0
	HO	0	0	0	0
	HS	0	0	0	0
	HP	0	0	0	0
	HA	0	0.2	0	0
Nitrogen	N	0	0	2	1.6
	NA	0	0	0	0
	NB	0.1	0	0	0
	N2	0.1	0.2	0.2	0
	N3	0	0	0	0
	O	1	0.2	5.1	5.6

Oxygen	O2	0.1	0.2	0	0
	OH	0	0	0.3	0.7
Sulphur	SH	0	0	0	0
	S	0	0	0	0
Water (O/H)	OW	6.6	4	4.7	2.1
	HW	0.1	0	0	0.2
	Hetero atom (ligand)	0.1	0	0	0
	Void	91.1	92	72.8	74.4

Table 4.2: Percentage occupancy of atom types that overlap with the interfacial water molecules in Scytalone dehydratase (water W537) and EGFR kinase (water W1104). This is computed from the hits obtained after a structure based search, across PDB and a non redundant subset of proteins from the PDB that is referred to as NR95 (see Methods [section 4.2](#)), for motifs that resemble the local chemical environments of these 2 interfacial waters.

Figure 4.9: Structure based search used to evaluate the displaceability of W1 in the mRNA cap binding interface of eIF4E. The motifs similar to that defined by the 4 closest atoms within a distance of 3.5 Å around W1, are searched for, across the PDB.

An additional structure based search against the entire PDB is performed for a water molecule W1 that is present in the mRNA cap binding site of the human translation initiation factor eIF4E (Figure 4.7) [manuscript under review]. W1 is extremely conserved across the crystal structures of the homologs and also in the MD simulations of eIF4E, as characterized using solvent density and residence times. It is part of the water bridge motif that was suggested to act as a recognition motif for the mRNA cap binding. Given that W1 is crucial for the binding of mRNA cap, it is likely to be energetically favorable and less likely to be favorably displaced by a ligand. This structure-based search was designed to act as a control: while W537 (in Scytalone dehydratase) and W1104 (EGFR kinase) were likely to be favorably displaced, W1 (in eIF4E) is much less likely to be favorably displaced. A structure based search of motifs similar to the local chemical environment of W1 in the entire PDB resulted in 599 hits (Figure 4.9). Analysis of these hits shows that there is a significant chance (~18%) of observing a water molecule overlapping with the position of W1 (Table 4.3). This contrasts well with a lowwater occupancy at positions W537 (~6%) and W1104 (~4%) in Scytalone dehydratase and EGFR kinase respectively. Therefore, a structure-based search clearly has the potential to provide qualitative discrimination amongst different displaceabilities of interfacial waters.

Element	Atom types	eIF4E
Nitrogen	N2	4
Water (O/H)	OW	18.1
	Void	77.4

Table 4.3: Percentage occupancy of atom types that overlap with the interfacial water molecule (W1) in eIF4E is shown. This is computed from the hits obtained after a structure based search, across the entire PDB, for motifs that are similar to the local chemical environment of W1.

The structure based search, however, has a limitation of being sensitive to the coordinates to be searched for against the database. Small deviations in the coordinates of the interfacial water or its local chemical environment can affect the output. Therefore, it is important to limit the search to high resolution crystal structures or high quality models. In summary, this approach of using a structure based search to displace interfacial waters holds tremendous promise as it can be integrated into ligand design protocols.

4.3.4 Performance assessment

	p53 peptide (120 atoms, 1000 frames)		eIF4E (1500 atoms, 1000 frames)		p53 DBD (1579 atoms, 1000 frames)	
	CPU time	RAM	CPU time	RAM	CPU time	RAM
Preprocessing	9 min	~ 16 MB	6.01 hrs	~ 16 MB	7.32 hrs	~ 16 MB
Solvation, Residence time, Exchange frequency, Favorable/Unfavorable interactions	31 min	~ 16 MB	3 days 8 hrs 50 min	~ 16 MB	3 days 16 hrs 30 min	~ 16 MB

Table 4.4: Computational cost associated with the hydration calculations. The frames refer to snapshots taken from MD simulations.

Table 4.4 shows the computational cost associated with the hydration analysis and how it varies with the system size. Analysis was done for entire protein chains. The optimal trajectory input size and spacing between frames depends on the questions being asked. Here, the analysis is performed on 1000 evenly spaced frames across a simulation time of 100 ns; thus the time interval between consecutive frames is 100ps. Several measures were taken to optimize the speed of the calculations. For instance, a preprocessing step is introduced to reduce

the size of the system. It trims the hydrogens and water molecules beyond the first solvation shell. Tracking water molecules for solvent density/residence time involves measuring distances between solute atoms and all solvent atoms which is extremely time consuming. This problem is partly overcome by sorting the solvent molecules to be searched according to the X, Y and Z coordinates and then searching only for waters that are in the vicinity of the solute atoms. Further optimization of the programs and the updates will be made available at 'mspc.bii.a-star.edu.sg/minhn/jal'. The program to identify buried waters in an MD trajectory was quite efficient and turns out to be largely independent of the protein size, requiring a computing time of ~1 second per snapshot. On average, the program to perform a structure based search to displace waters in an input single crystal/MD structure took 6 and 24 hours for NR95 and the full PDB datasets respectively, on a cluster of 80 CPU nodes.

4.4 Conclusion

With the surge in computational power, MD simulations are increasingly being used to simulate biomolecular processes (Dodson et al, 2008). Subsequently, powerful analysis programs are required to process the generated data. Motivated by the scarcity of programs that allow efficient computation of hydration properties, particularly residence time and buried water calculations in an MD trajectory, a software suite of programs ('JAL') is developed that allows the computation of solvation properties in a conformational ensemble. The programs are interactive and take a MD trajectory as input from the user, in addition to the

optional parameters required for the calculation. More emphasis is given on making the program user friendly and the computations fast and accurate. The calculation of several hydration properties have been benchmarked against the available experimental data. Finally, to the best of my knowledge, this is the first program that allows the computation of buried waters in an MD trajectory and uses a novel “evaporation” technique. Textual output is available in a format that can be easily examined visually (through plots or in 3D through PyMOL) and also for easy and rigorous analysis of the solvation properties.

The utility of ‘JAL’ is demonstrated by using it to enquire into hydration in the proteins p53 DBD, eIF4E, Scytalone dehydratase and EGFR kinase; the findings also yielded useful biological insights and novel directions to be pursued. The study of buried water molecules in MD simulations of WT and V143A mutant form of p53 DBD provides interesting insights on the possible role of water in structural stabilization. A buried water molecule that bridges the loops L2 and L3 in the WT DBD is lost in the V143A oncogenic mutant form as the loop L3 assumes a non-native conformation. Calculations suggest that the loss of this buried water in the mutant could likely affect zinc coordination to p53; the latter is known to have a role in p53 stabilization. Interrogation of the hydration dynamics of the mRNA cap binding interface of eIF4E demonstrates that the waters in the mRNA cap binding site having high residence times can form long-lived water bridges that connect sites on proteins and can contribute to stability and recognition. A highlight of ‘JAL’ is a novel algorithm that evaluates the likelihood of displacing interfacial waters during drug optimization. Using examples of

Scytalone dehydratase, EGFR kinase and eIF4E, 'JAL' mines the PDB for a similarity-based search that not only provides an insight into which waters at an interface can be favorably displaced by a designed ligand, but also suggests ligand modifications that can favorably displace these waters. The approach is rapid and can complement the more expensive but rigorous free energy based methods. Finally, based on the computation of solvation and residence time of waters, it will be interesting to investigate if sites with high solvation and low residence times are more amenable to be favorably displaced than sites with high solvation and high residence time.

Preface to Chapter 5

In the next chapter, the waters at the nucleotide binding interface of eIF4E are probed using the programs in 'JAL' package with the purpose of gaining an understanding of the role of water in the binding process. eIF4E is a translation initiation factor and an established therapeutic target for cancer. Available crystallographic data for the human eIF4E suggests the extensive presence of waters at the interface of protein and mRNA cap. A systematic characterization of these interfacial waters can guide the design of novel inhibitors of eIF4E.

Chapter 5: Systematic data mining reveals a water-bridge that mediates recognition of mRNA cap in eIF4E

Ligand binding pockets in proteins contain water molecules which are known to play an important role in mediating/modulating protein-ligand interactions. The 5' mRNA cap binding interface of the translation initiation factor protein eIF4E represents one such interface. Analysis of the available crystallographic data of eIF4E yields several structurally conserved waters in the binding pocket, most of which also persist in MD simulations of the protein. These conserved waters are involved in the formation of an intricate hydrogen-bond network between the cap and the protein. A subset of the water positions appears to be conserved even in the absence of the ligand, and two particular positions in the cleft of the binding pocket show a high degree of occupancy both in the crystal and simulated structures. They are involved in the formation of a water-mediated solvent bridge connecting two residues which are spatially constrained by a set of evolutionarily conserved residues and which act as pillars supporting the bridge. Together with the bound waters, these residues form a scaffold in the binding pocket that appears to be part of a recognition module for the mRNA cap. It can be postulated that these two structural waters are critical and additionally recruit other waters, thereby facilitating an efficient interaction between the molecules. A new crystal structure of the ligand free protein reported here further supports the

observation made from the comparative structural and simulation studies. These findings provide new insights towards our current understanding of the mRNA cap-binding interface in eIF4E and open fresh directions towards designing and screening of compounds targeting eIF4E for the development of therapeutics.

5.1 Introduction

Water is the most abundant component of the cellular environment and strongly influences the behavior of macromolecules (Ball, 2008b; Zhong et al, 2008). It is known to have critical roles in their structure, dynamics and functional properties (Bagchi, 2013; Hummer & Tokmakoff, 2014). These observations have been underscored by the presence of conserved water molecules at specific spatial locations in different proteins (Bottoms et al, 2002; Kanaujia & Sekar, 2009; Knight et al, 2009; Likic et al, 2000; Milano et al, 2015; Vijayan et al, 2010). Indeed, water has often been referred to as the 21st amino acid, filling voids to stabilize packing in protein structures and thereby mediating their structure-activity relationships (Carugo, 2015). This has resulted in attempts to understand the role of “biological water” i.e. water molecules/shells associated with protein surfaces/binding sites in facilitating molecular interactions and catalysis (Ahmed et al, 2011; Barillari et al, 2007; Raschke, 2006). A natural consequence of this is the recognition and rational exploitation of protein-ligand binding cavity hydration for engineering compounds during lead optimization in drug design and development (Michel et al, 2009a; Wong & Lightstone, 2011).

The eukaryotic translation initiation factor protein eIF4E is a cap-binding protein that recognizes mRNA molecules through the cap structure (7-methyl-guanisine triphosphate) present at the 5' end of the mRNA (Brown et al, 2007). This complex forms higher order states through interactions with other protein partners, eventually culminating in the ribosome complex, resulting in translation (Jia et al, 2012; Rhoads, 2009). It is therefore an essential component for all proteins that get translated through the cap-dependent pathway (Merrick, 2004). Three-dimensional structures of eIF4E in complex with different cap mimics have provided a detailed understanding of the nature of the interactions between the protein and the ligand (Figure 5.1) (Brown et al, 2009b). The cap binding interface of eIF4E has been well characterized with a defined region that acts as a site of docking for the cap. The residues that mediate the interactions include hydrophobic (W56, W102 and W166) and hydrophilic (E103, R112, R157 and K162) amino acids. The tryptophan residues along with E103 are involved in recognition of the N⁷ substituted guanosine base of the cap whereas the basic residues interact with the phosphate group (Figure 5.1).

Figure 5.1: Crystal structure of m^7GTP bound to eIF4E. The structure with PDB entry 2W97 (chain A, 2.3 Å (Brown et al, 2009b)) is taken as a representative to indicate the binding mode of the mRNA cap analogue m^7GTP (7-methylguanosine triphosphate) with eIF4E at the cap-binding interface. The residues recognized to be involved in the recognition are indicated. The m^7GTP cap analogue is shown in stick representation.

There have been reports in the literature about the presence of explicit water molecules in the cap-binding interface that mediate the interactions between eIF4E and the mRNA cap ligands (Brown et al, 2007; Marcotrigiano et al, 1997; Niedzwiecka et al, 2002; Tomoo et al, 2003); however a detailed and comprehensive study on their precise role is lacking. This knowledge becomes more critical because eIF4E is an oncogene which is known to be overexpressed in various malignancies (Carroll & Borden, 2013; Dua et al, 2001; Mamane et al, 2004) and hence its inhibition using oligonucleotides (Bhattacharyya et al, 2014;

Thumma et al, 2015), small organic molecules (Chen et al, 2012; Ghosh et al, 2005) and peptides (Brown et al, 2011b; Lama et al, 2013) is being pursued as an anticancer therapeutic opportunity (Borden, 2011; Hsieh & Ruggero, 2010). To date, while structural data on the complexes between eIF4E and cap analogues have been used to guide the development of promising lead compounds targeting the cap binding interface (Chen et al, 2012), inclusion of interfacial water molecules has not been incorporated in the process.

The present study investigates the hydration profile of the cap-binding interface of eIF4E through a combination of comparative structural analysis and detailed molecular dynamics (MD) simulations. These studies have led to the identification of a set of invariant water positions and their intricate network with the protein and the cap analogues. Detailed analysis has established the presence of two spatially conserved water positions that act as a solvent bridge in facilitating the formation of a structural network in the cap-binding interface of eIF4E, suggesting a possible role for this structural module in facilitating efficient recognition between eIF4E and the cap analogues. The study motivated the resolution of a new crystal structure of apo eIF4E, which also revealed the presence of this water bridge and the associated local scaffold. These observations collectively indicate the identification of a water mediated recognition module involved in the fundamental mechanism of eIF4E-mRNA cap recognition and which could be exploited in the design and development of lead compounds against the oncogenic protein.

5.2 Methods

5.2.1 Analysis of available crystallographic data on eIF4E

Sr. No.	PDB ID	Structure	Resolution (Å)	Number of waters	References
<i>Cap-bound</i>					
1	2W97:A	eIF4E-eIF4G1-m ⁷ GTP	2.3	31	(Brown et al, 2009b)
2	1WKW	eIF4E-4EBP1-m ⁷ GpppA	2.1	22	(Tomoo et al, 2005)
3	1L8B	eIF4E-m ⁷ GTP	1.8	26 - 30	(Niedzwiecka et al, 2002)
4	1EJ1	eIF4E-m ⁷ GDP	2.2	17 - 26	(Marcotrigiano et al, 1997)
5	1EJ4	eIF4E-4EBP1-m ⁷ GDP	2.3	14	(Marcotrigiano et al, 1999)
6	2V8W	eIF4E-4EBP1-m ⁷ GTP	2.3	32 - 38	(Brown et al, 2007)
7	2V8X	eIF4E-4EBP1-Bn ⁽⁷⁾ GMP	2.3	39 - 40	(Brown et al, 2007)
8	2V8Y	eIF4E-4EBP1-FBn ⁽⁷⁾ GMP	2.3	30 - 35	(Brown et al, 2007)
9	2IDV	eIF4E-m ⁷ GDP	2.3	32	(Monzingo et al, 2007)
10	2WMC	eIF4E-m ⁷ GTP	2.2	13 - 29	(Ashby et al, 2011)
11	3AM7	eIF4E-4EBP2-m ⁷ GTP	2.2	26	-
12	3HXG	eIF4E-4EBP1-m ⁷ GpppG	2.1	34	(Liu et al, 2009)

13	3HXI	eIF4E-4EBP1- m ^(2,2,7) GpppG	2.1	24	(Liu et al, 2009)
14	4TPW	eIF4E-m ⁷ GTP	1.5	37 - 49	(Papadopoulos et al, 2014)
15	4TQB	eIF4E-m ⁷ GTP	1.6	49 - 54	(Papadopoulos et al, 2014)
16	4TQC	eIF4E-m ⁷ GTP	1.9	46 - 49	(Papadopoulos et al, 2014)
17	4AZA	eIF4E-eIG4G1- m ⁷ GTP	2.2	26 - 32	(Zhou et al, 2012)
18	4UEC	eIF4E-eIF4G- m ⁷ GTP	2.4	11	(Peter et al, 2015a)
19	5BXV	eIF4E-4EBP1- m ⁷ GTP	2.1	28	(Sekiya et al, 2015)
20	5ABU	eIF4E-Mextli- m ⁷ GTP	2.2	16	(Peter et al, 2015b)
21	5ABX	eIF4E3-Mextli- m ⁷ GTP	1.7	21	(Peter et al, 2015b)
Cap-free					
1	2W97:B ^d	eIF4E	2.3	23	(Brown et al, 2009b)
2	2IDR	eIF4E	1.9	33-35	(Monzingo et al, 2007)
3	3TF2	eIF4E	2.1	10 - 14	(Siddiqui et al, 2012)
4	3U7X	eIF4E-4EBP1	2.1	13 -14	(Siddiqui et al, 2012)
5	4UED	eIF4E-4EBP1	1.8	21	(Peter et al, 2015a)
6	4UE9	eIF4E-4ET	2.2	15	(Peter et al, 2015a)
7	4UE8	eIF4E-Thor	1.1	31	(Peter et al, 2015a)

8	5ABY	eIF4E-Mextli	2.0	22 - 31	(Peter et al, 2015b)
9	5ABI	eIF4E	2.0	23	-

Table 5.1: Crystal structures considered for the analysis of structurally conserved waters in the cap-binding interface of eIF4E. Crystal structures (available as of 1st October 2015) were downloaded from the PDB. They were classified into cap-bound and cap-free structures based on the presence/absence of the cap analogue in the cap-binding interface. eIF4E structures are solved in apo, binary or ternary states. The peptides and the cap-analogue in these structures are shown. The water molecules were calculated within a distance of 15Å radius from the center of mass of m⁷GMP (7 methyl guanosine triphosphate) which is the minimum common moiety in all the cap analogues. See Methods (section 5.2) for details about the calculation. For structures with multiple chains, the range of resolved waters is indicated. See Table 5.1 for details on the exact number of waters of every chain. Only structures with more than ten resolved waters in the cavity were considered for the analysis. For 2W97, chain A and chain B are considered in two states separately as the structure was solved in a dimeric state with and without the cap analogue.

Sr. No.	PDB ID: Chain	No. of waters	Invariant water positions													
			W 1	W 2	W 3	W 4	W 5	W 6	W 7	W 8	W 9	W1 0	W1 1	W1 2	W1 3	W14
1	2W97: A	31														
2	3AM7	26														
3	2V8W: A	38														
4	2V8W: E	32														
5	2V8X: A	39														

23	5BXV: C	28														
24	1EJ4	14														
25	4UEC: C	11														
26	5ABU	16														
27	3HXG	34														
28	3HXI	24														
29	5ABX	21														
30	2IDV	32														
31	2WMC :A	29														
32	2WMC :B	20														
33	2WMC :C	24														
34	2WMC :D	24														
35	2WMC :E	21														
36	2WMC :F	13														
37	2WMC :G	19														
38	2WMC :H	18														
% Occupancy			97	97	87	76	63	55	63	74	74	82	66	61	68	63

Table 5.2: Crystal structures of cap-bound eIF4E and the structurally conserved waters in the cap-binding interface. Crystal structures (available as of 1st October

2015) were downloaded from the PDB and those with the presence of a cap analogue at the cap-binding interface were categorized as cap-bound structures. Individual chains of each structure were considered exclusively for the analysis. Water molecules were identified within a distance of 15Å radius from the center of mass of m⁷GMP (7 methyl guanosine triphosphate) which is the minimum common moiety in the cap analogues. See Method ([section 5.2](#)) for details about the calculation. Only structures with more than ten resolved waters in the interface region were considered for the analysis. Significant water clusters (%Occupancy > 50) observed are examined. Red and green background indicates present and absence of water molecules respectively in the individual structures.

Sr. No.	PDB ID: Chain	No. of waters	Invariant water positions														
			W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	
1	2W97:B	23	Green	Green	Green	Green	Red	Green	Green	Red	Green	Red	Red	Red	Green	Green	
2	3TF2:A	14	Green	Green	Green	Red	Red	Green	Green	Red	Green	Red	Red	Red	Red	Green	
3	3TF2:B	12	Green	Green	Red	Green	Red	Green	Green	Red	Green	Red	Red	Red	Green	Green	
4	3U7X:A	14	Green	Green	Green	Green	Red	Green	Green	Red	Red	Red	Green	Red	Green	Green	
5	3U7X:B	13	Green	Green	Green	Green	Green	Green	Green	Red	Red	Red	Green	Red	Green	Green	
6	5ABI	23	Green	Green	Green	Green	Green	Red	Red	Red	Green	Red	Red	Green	Red	Green	
7	4UED	21	Green	Green	Green	Green	Green	Red	Red	Red	Red	Green	Red	Red	Red	Green	
8	4UE8	31	Green	Green	Green	Green	Green	Red	Red	Green	Red	Green	Green	Red	Green	Green	
9	4UE9	15	Green	Green	Green	Red	Red	Red	Red	Red	Red	Green	Red	Green	Green	Green	
10	5ABY:A	31	Green	Green	Green	Green	Green	Red	Red	Red	Red	Red	Red	Green	Red	Red	
11	5ABY:C	22	Green	Green	Red	Green	Red	Red	Red	Green	Red	Red	Red	Green	Red	Red	
12	5ABY:E	26	Green	Green	Red	Green	Green	Red	Red	Red	Red	Red	Red	Green	Red	Red	
13	2IDR:A	35	Green	Red	Green	Green	Red	Red	Red	Green	Red	Red	Red	Red	Green	Red	
14	2IDR:B	33	Green	Green	Red	Red	Red	Red	Red	Red	Red	Red	Red	Green	Green	Red	

% Occupancy	100	93	71	79	43	36	36	21	29	21	21	43	57	64
-------------	-----	----	----	----	----	----	----	----	----	----	----	----	----	----

Table 5.3: Crystal structures of cap-free eIF4E and the structurally conserved waters in the cap-binding interface. Crystal structures (available as of 1st October 2015) were downloaded from the PDB and those without cap analogue at the cap-binding interface were categorized as cap-free structures. Individual chain of each structure was considered exclusively for the analysis. The coordinates of the center-of-mass of the m⁷GMP cap-analog was used as a reference to identify water molecules within a distance of 15Å radius. See Method ([section 5.2](#)) for details about the calculation. Only structures with more than ten resolved waters in the interface region were considered for the analysis. Water clusters observed and analyzed for occupancy. Red and green background indicates present and absence of water molecules respectively in the individual structures.

Crystal structures of eIF4E were retrieved from the PDB. The resolution of the structures ranged from 1.1 Å to 2.3 Å. For the structures with more than one molecule of eIF4E in the unit cell, each chain was considered separately, resulting in a total of 75 unique structures. The number of crystallographically resolved water molecules in the cap-binding interface was evaluated (as described below) for all these structures ([Figure 5.2](#)). It is observed that most of the structures are well-hydrated and hence a criterion of considering only those structures with more than 10 waters in the interfacial region was set for this study. These structures were then classified into cap-bound and cap-free states depending on the presence/absence of the cap analogue in the cap-binding interface ([Tables 5.1, 5.2 and 5.3](#)).

Figure 5.2: Water molecules around the cap binding interface of eIF4E structures.

Each chain of the different X-Ray resolved structures of eIF4E was considered as unique for the analysis (See Table 5.1, 5.2 and 5.3). The waters molecules were identified within a sphere of 35Å around the cap binding interface (See Methods (section 5.2) on details about the calculation). The water-structure plot is shown for both the cap-bound and cap-free structural states of eIF4E.

5.2.2 Structural alignment and detection of water clusters

Figure 5.3: Crystal structure of eIF4E with PDB entry 2W97. This structure of eIF4E which is solved as a dimer in the unit cell is taken as a representative to highlight the cap-bound (orange color, chain A) and cap-free structure states (gray color, chain B) of eIF4E (See [Table 1](#)). The crystal structure had only one cap molecule and it was found bound to chain A.

To study the hydration patterns in the cap-binding interface, the crystal structures of eIF4E are grouped into cap-bound states and cap-free states ([Tables 5.2 and 5.3](#)). The crystal structure with PDB ID 2W97 ([Brown et al, 2009b](#)) contains two molecules in the unit cell: one contained a cap analogue (m7GTP) bound while the other molecule had no ligand at the cap-binding interface ([Figure 5.3](#)). This structure thus served as a natural reference because the same unit cell contained a cap-bound and a cap-free eIF4E on to which the remaining structures could thus be superimposed separately. Structural superimposition was carried out using the “align” function in the PyMOL molecular visualization software ([DeLano, 2009](#)).

The cap analogue in the different structures of eIF4E consists of an N7 substituted guanosine base and a sugar with an attached monophosphate, diphosphate, or triphosphate moiety (Table 5.1). Some of the structures also contain a nucleic acid attached to the cap. The smallest structural motif common to all the structures is the base, sugar and the monophosphate (m7GMP). Thus for each structure, all water molecules which are within a radius of 15Å from the center-of-mass of the m7GMP moiety were first identified. Following this, a sphere of radius 3.5 Å was constructed centered at each water oxygen atom and any water molecule that lay within this sphere was selected. Consequently, if a particular sphere contained waters that were seen in at least 50% of the crystal structures, it was considered to be significant and classified as belonging to a water cluster. The same procedure was used for identifying water clusters in the cap-free structures of eIF4E. In this case, the coordinates of the center-of-mass of the cap-analog (obtained after superimposition with the cap bound eIF4E) were used as a reference to initially identify water molecules within a 15Å radius.

5.2.3 System preparation for simulation

The protein was modeled using the all-atom ff99SB force field parameters (Hornak et al, 2006). The parameters for the ligand m7GTP cap analogue were derived using the Gaff Force Field51 parameters (Wang et al, 2004) through the Antechamber module in AMBER12 (Case et al, 2012). For the ligand, RESP (Restricted Electrostatic Potential) based atomic charges corresponding to a net charge of “-3e” were derived using a combination of the building block fragments from F-90 R.E.D.D.B database (Dupradeau et al, 2008) and the R.E.D. server

([Vanquelef et al, 2011](#)) by employing the RESP-A1A (HF/6-31G*) charge model and the Gaussian_2009_C.01 quantum _mechanics program. The atomic coordinates of the initial structures for both the cap-bound and cap-free states of the protein were obtained from the crystal structure 2W97, using chains A and B respectively ([Brown et al, 2009b](#)) ([Figure 5.3](#)). All the crystallographically resolved water molecules were retained for the simulations. The structures were placed in the centre of a cuboid box ensuring a minimum distance of 12 Å between any protein atom and the edge of the box. TIP3P water model ([Jorgensen et al, 1983b](#)) was used and the systems were neutralized by adding appropriate number of ions ranging from 0 to 4 using the TLEAP module of AMBER12 ([Case et al, 2012](#)).

5.2.4 MD simulations

MD Simulations were carried out with AMBER12 ([Case et al, 2012](#)) using PMEMD. The solvated systems were energy minimized using steepest descent and conjugate gradient algorithms, heated to 300K under NVT conditions, equilibrated for 500 ps and finally subjected to 100 ns of production dynamics using the NPT ensemble. Each system was simulated in triplicate with different random seeds employed for the generation of velocities during equilibration. Temperature was regulated using langevin dynamics ([Loncharich et al, 1992](#); [Pastor et al, 1988](#)) with a collision frequency of 1.0 ps⁻¹ and the pressure was maintained at 1atm using weak-coupling ([Berendsen et al, 1984](#)) with a pressure relaxation time of 1ps. Periodic boundary conditions in the x, y and z directions were appropriately applied. The Particle Mesh Ewald (PME) method ([Darden et al, 1993a](#)) was used

for treating the long-range electrostatic interactions. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm (Ryckaert et al, 1977b), enabling a time step of 2fs.

5.2.5 Water occupancy maps and total residence times

Occupancy maps of water in the MD simulated trajectories were generated with the grid command in the PTRAJ module of AMBER12 (Case et al, 2012) with grid cells of size 0.5 Å X 0.5 Å X 0.5 Å. The number of waters within each grid cell was counted and the data was analyzed using the volume viewer menu in the Chimera visualization software (Pettersen et al, 2004). The cutoff isovalue was adjusted to “x” (x= 1.5, 2.5 or 3.5) times the peak value, which is the mode of the isovalue distribution. It is an arbitrary criterion that allows for removal of weak binding sites and hence filters the map to examine only the relatively more strongly bound water positions. Such an approach has previously been used to examine occupancy maps of solvent from MD simulated trajectories (Lama et al, 2015; Tan, 2015).

Total residence time (TRT): The total water residence time at a particular site was computed as a summation over the trajectory of the number of consecutive structures which do not undergo desolvation or replacement of a given water molecule.

$$TRT = \sum_{t=0}^{100ns} \gamma(W(t), W(t + \Delta t)) \quad \dots(5.1)$$

'W' is the ID of water oxygen at a given instance of time and ' Δt ' is the time interval between two successive snapshots of the MD trajectory. Here, Δt is 100 ps. Value of ' γ ' is 1 if $W(t) = W(t + \Delta t)$ and 0 otherwise.

5.2.6 Design and Construction of Surface Entropically Reduced eIF4E (23-217) for Crystallization (work described in the following sections on crystallization were carried out by our collaborators Dr Chris Brown and his team in the p53lab, A*STAR)

Crystallization involves the formation of intermolecular contacts, which allows the arrangement of protein molecules into a crystalline lattice. Immobilization of side chains with high conformational entropy at the point of crystal contacts is energetically unfavorable and is expected to impede crystallization. Therefore the replacement of solvent-exposed amino acid sequence motifs with high conformational entropy with residues of lower conformational entropy may permit the formation of thermodynamically favorable crystal contacts. This concept has been demonstrated by Derewenda et al ([Cooper et al, 2007](#)) to improve crystallization of several different proteins, in which clusters of two to three amino acids with high conformational entropy, such as Lys, Glu, and Gln, are replaced with Ala. The SERp server was used to identify the following residues to be mutated to Ala: E105, K106 and K108. In addition the long flexible N-terminal of eIF4E, which has not been observed in the electron density of any of the uploaded PDB entries, was truncated to generate the construct eIF4E (23-217) to enhance its crystallization. The Stratagene mutagenesis kit (Agilent) was used to

introduce the alanine mutations and the N-terminal truncation into the eIF4E sequence.

5.2.7 Protein expression and purification

The surface entropically reduced construct of the truncated human eIF4E (sER-eIF4E23-217) was expressed and purified identically to procedures described previously for full length human eIF4E ([Brown et al, 2011b](#); [Brown et al, 2007](#); [Brown et al, 2009b](#)).

5.2.8 Crystallization

The sER-eIF4E 23-217 construct was crystallized by vapor diffusion using the sitting drop method. Crystallization drops contained 125 μ M of eIF4E. Sitting drops were set up in 48 well Intelli-Plates (Hampton research) with 1 μ l of the protein sample mixed with 1 μ l of the mother-well solution. Crystals grew over a period of one week in 18-26% of Polyethylene glycol methyl ether 5000, 0.01-0.02M Ammonium Sulphate and 100mM Bis-Tris at pH 5.5. For X-ray data collection at 100 K, crystals were transferred to an equivalent mother liquor solution containing 20% (v/v) glycerol and then flash frozen in liquid nitrogen.

5.2.9 Data collection and refinement

The data was collected at the NSRCC beamline BL13B1 in Taiwan using a CCD detector. The crystal diffracted to a resolution of 1.97 Å and was integrated and scaled using HKL2000. The initial phases of the crystallised eIF4E were solved by molecular replacement with the program PHASER ([McCoy et al, 2007](#)) using the cap-free human eIF4E structure found in the asymmetric unit with the bound peptide removed (PDB accession number: 2W97) as a search model. The

starting models were subjected to rigid body refinement and followed by iterative cycles of manual model building in Coot and restrained refinement in Refmac 6.5.([Murshudov et al, 1997](#)) Models were validated using PROCHECK ([Laskowski et al, 1993](#)) and the MOLPROBITY webserver ([Chen et al, 2010](#)). Final models were analyzed using PYMOL (Schrödinger). [Table S1](#) lists the data collection and refinement statistics. The eIF4E complex structure has been deposited in the PDB under the submission code 5ABI.

5.3 Results

5.3.1 Characterization of structural water molecules at the cap-binding interface

Figure 5.4: Water clusters, occupancy map and B-factors in the protein eIF4E. The identified water clusters from comparative structure analysis and the water occupancy map from MD simulations around the cap-binding interface for the (A) cap-bound and (B) cap-free states of the protein. The waters positions from W1 to W14 in the cap-bound state and its subset in the cap-free state of the protein are indicated. The reference structure for both states is taken from PDB ID 2W97. The water occupancy map is shown at an isocutoff value of 3.5 times the peak value which is the mode of the isovalue distribution. (C) Average B-factor values of the waters molecules which are part of the clusters present in the cap-bound and cap-free states.

Analysis of waters around the cap-binding interface in the cap-bound state of the protein led to the identification of 14 significant clusters which occupy distinct spatial positions around the cap. These positions are termed W1 to W14 (Tables 5.2 and 5.3; Figure 5.4A). Of the 14 waters clusters identified, W1 and W2 show maximum occupancy, being conserved in more than 90% of the structures. There are 6 distinct clusters in the cap-free states of the protein (W1, W2, W3, W4, W13 and W14) (Tables 5.2 and 5.3; Figure 5.4B) which are a subset of the groups observed in the cap-bound states. The elevated number of water clusters in the cap-bound states indicates that water retention around the binding interface is higher in the presence of the cap; this may reflect a cap dependent ordering within the cap-binding interface. Interestingly, W1 and W2 have high occupancy of more than 90% in the cap-free states of the protein also. The average crystallographic B-factors of the different water clusters were computed from the crystal structures (Figure 5.4C) for both the cap-bound and cap-free states. The corresponding waters positions in the cap-bound states are more ordered as compared to the ones in the cap-free state. Further, W1 and W2 waters have the lowest B-factor values when compared to other waters within the two different states which suggest that the waters at these positions are structured and relatively less mobile.

Figure 5.5: Water mediated hydrogen-bond network. The potential hydrogen-bond network in the cap-binding interface involving the water clusters for (A) cap-bound and (B) cap-free states. The representative structure for both states is taken from 2W97. UCSF chimera visualization software package was used. The hydrogen-bond was defined for the heavy atoms of the residues in the protein, m⁷GTP cap analogue and the oxygen atoms of the water molecules. Hydrogen atoms were not added explicitly in the calculations. Residues in the cap-binding interface which are possibly involved in the formation of the interaction network between the water molecules and ligand are indicated.

Solvation features around the cap-binding interface were further investigated by analysis of ensembles of structures generated from the MD simulations in terms of the water occupancy profiles (Figures 5.4 and 5.6). The

occupancy map was investigated at three different threshold cutoff values (Figures 5.6). Most of the water clusters that were observed in the crystal structures are also seen be occupied in the simulations at lower threshold. Although, as the threshold was increased, only the waters at W1 and W2 seemed to be distinctly occupied which indicates the relatively higher degree of ordered character of the waters at these positions. This corroborates the observations from the crystal structure analysis that these two water positions are less dynamic and undergo relatively less exchange with the bulk solvent.

Figure 5.6: Water clusters and occupancy maps in the protein eIF4E. The identified water clusters from comparative structure analysis and the water occupancy map from MD simulations around the cap-binding interface for the (A, C, E) cap-bound and (B, D, F) cap-free states of the protein. The waters from W1 to W14 in the cap-bound state and its subset in the cap-free state of the protein are indicated. The reference structure for both states is taken from PDB ID 2W97. The water

occupancy map is shown at three different isocutoff values of 1.5 (A, B), 2.5 (C, D) and 3.5 (E, F) times the peak value which is the mode of the isovalue distribution. W1 and W2 water positions show significant occupancy as compared to other positions even at greater cutoff values.

5.3.2 Water mediated interactions between the protein and cap

The conservation of waters at specific locations in the different structures of eIF4E indicates that they may play particular role in influencing the binding of the mRNA cap. Analysis of the interactions formed by these waters ([Figure 5.5A](#)) shows potential hydrogen-bond networks bridging the nitrogen in the indole group of W166, waters at position W1 and W2 and the side-chain carbonyl oxygen of N155. The water at W1 also interacts with the nitrogen in the guanidinium group of R112 and with the water at W3. Such a favorable interaction environment could be a primary reason for the stability and high occupancy of the water at W1. Basic residues R112 and K162 are involved in water-mediated interactions with the negatively charged phosphate moiety of the cap via water molecules at positions W4, W5, W6 and W7. This is the electrostatic component of molecular recognition that had earlier been identified as important for steering the ligands towards the cap-binding interface ([Blachut-Okrasinska et al, 2007](#); [Niedzwiecka et al, 2002](#)). The strategic ordering of waters at positions W4, W5, W6 and W7 likely ensure an efficient hydrophilic interaction after ligand-complexation associated desolvation of the region. Waters at positions W8 and W9 form a network between the indole of W102, which is one of the two tryptophan residues that sandwiches the guanosine ring, and the side-chain of residues H200 and T203. Structurally conserved water at position W3 bridges W10, W11 and W12, which together with

W13, W14 and the hydroxyl of S92 form a closed caged network. The water at position W12 links the cap and protein via the side-chain of residue D90, phosphate of the cap and the side-chain indole of W56 which is the other tryptophan residue that interacts with the guanosine.

Thus, it seems that the spatial arrangement of the structural waters from W1 to W14 form an intricate interaction network that facilitates binding between the protein and the cap. A similar analysis of the cap-free state and its comparison with the cap-bound form shows that the network involving the subset of conserved waters at position W1, W2, W3, W4, W13 and W14 is identical (Figure 5.5B). The presence of this sub-network suggests that it could act as a scaffold facilitating the formation of a higher order network as observed in the cap-bound state. Water at W1 is involved in the formation of similar interactions as seen in the bound states involving waters at W2 and W3 which could be important as it is observed to link most of the other waters in the complex. Comparatively, it appears that the set of connections involving residue W166, waters at positions W1, W2 and residue N155 lie at the core of this network and the waters at W1 and W2 act as linkers in bridging the residues.

5.3.3 W1 and W2 waters are held by interactions with residues W166 and N155

Figure 5.7: Water occupancy map in MD simulations of wild and mutant eIF4E. Occupancy map highlighting the waters at positions W1 and W2 from MD simulations of the protein in the cap-bound and cap-free states for (A) wild type, (B) N155L, W166F double mutant, (C) W166F and (D) N155L single mutants. The map is shown at an isocutoff value of 3.5 times the peak value which is the mode of the isovalue distribution. Both or either W1 and W2 waters are present in the wild type or single point mutants whereas they are lost in the double mutant. The arrow in (D) indicates the new water position that appears in the W166F mutant simulation.

The presence of a subset of spatially similar structural waters and their interaction network in the cap-free state of the protein suggests that they are an integral part of the binding interface and hence likely influence the recognition of the ligand in the cap-binding interface. This seems to be especially true for waters at W1 and W2. Analysis of the hydration properties of waters at these positions in terms of their solvation profiles and total residence times ([Table 5.4](#)) suggests that both the water positions are very well hydrated throughout the simulation period

which is in accord with their high occupancies in the crystal structures (Table 5.2, 5.3 and 5.4). However, the total residence times of both the waters varies between the cap-bound (86 ns and 95 ns) and cap-free states (65 ns and 37 ns). The relatively lower value in the cap-free state suggests that they are more dynamic and undergo exchange with the bulk solvent in the absence of the cap. The average crystallographic B-factors of both these waters are also lower in the cap-bound state as compared to the cap-free state (Figure 5.4C). This analysis indicates that waters at positions W1 and W2 are retained in their spatial location in the vicinity of residues N155 and W166 for significant periods of time and which keeps the region solvated between the two different states of the protein (Figure 5.4A).

System	Solvation (%)				Total Residence Time (ns)			
	Cap-free		Cap-bound		Cap-free		Cap-bound	
	W1	W2	W1	W2	W1	W2	W1	W2
Wild	84	97	90	99	65	37	86	95
N155L - W166 F	4	41	13	45	0	2	4	9
N155L	78	42	92	49	52	1	78	7
W166 F	4	97	18	93	0	40	3	59

Table 5.4: Hydration properties of water at positions W1 and W2. Structures extracted from the 100 ns of the simulated trajectory at equal intervals for each system was analyzed. Sites W1 and W2 were considered solvated in a structure if a water molecule was found within a sphere of radius 1.7 Å from a reference protein atom which is NE1:W166, CZ:F166 for W1 and OD1:N155, CD1:L155 for W2. The total residence times of water molecules at sites W1 and W2 was measured with the equation as described in the Methods (section 5.2).

Next, the influence of residues N155 and W166 on the stability of these waters was examined by generating *in-silico* mutants at positions 155 and 166, and evaluating the hydration profiles from the MD simulations of these mutants (Table 5.4). N155 was mutated to leucine and W166 to phenylalanine, both mutations designed to abrogate the side-chain hydrogen-bonds with the respective waters, without significantly affecting the van der Waals contacts with the surroundings. For the double mutant N155L_W166F, the water properties at W1 and W2 were adversely affected, with the region getting dehydrated, as observed with low solvation values and absence of water occupancy at these positions (Figures 5.7A and 5.7B). In case of the single point mutation N155L, the position W2 is poorly hydrated and the water undergoes rapid exchange, as indicated by the low total residence times (Table 5.4). However, the properties of W1 remain similar to those in the wild protein suggesting that the water at this position is independent of the changes at W2 (Figure 5.7C). For the other point mutation W166F, the position W1 gets desolvated as observed in the double mutant. However, a new water density in the region appears which is clearly shifted from the position where W1 would be present in the wild type protein, and this new position is closer to W2 (Figure 5.7D). One of the primary reasons for the occurrence of this new water position could be a stable hydrogen-bond interaction it forms with W2 in the absence of interactions due to the W166F mutation. This indicates that W166 is extremely important for holding the water at W1 at its precise location; however alternate water positions can be created when W166 is mutated, as the water readjusts in the new micro environment. Similarly as observed in the previous point mutation, W2

is unaffected by the change at W1 which again highlights the independent nature of occupancy of water at this position. Taken together, these *in silico* mutational experiments indicate that tryptophan and asparagine residues are both important for retaining the two respective water molecules at specific locations in the binding interface and thereby form the water-bridge connecting their side-chains. This bridge could be the basic structural motif that would act as a scaffold for the formation of the water mediated interactions between the mRNA cap and eIF4E.

5.3.4 Preformed docking site at the cap binding interface

Figure 5.8: Residue fluctuations, spatial constraint on N155 and W166 and residue conservation. (A) RMSF (Root mean square fluctuation) of all heavy atoms from the MD simulations for the selected set of residues in the cap-binding interface compared between the two structural states. (B) End state structure comparison of the selected residues between cap-bound (PDB ID: 2W97, Chain A, residues colored in grey) and cap-free states (PDB ID: 2W97, Chain B, residues colored in

orange). (C) The potential hydrogen-bond interaction network involving residues N155, W166 and the waters at position W1 and W2 is indicated. Also, the interacting hubs formed by H200, W166, F94 and D90, R157 and N155 are shown. The reference structure for the figure is 2W97. (D) The non-redundant sets of eIF4E proteins from different organisms whose structures have been solved were selected for the alignment which was performed using the program MUSTANG. The PDB entry of each structure and the respective organism are shown. Residue number is shown with reference to human eIF4E (Uniprot ID: P06730). The structurally equivalent residues which are involved in the recognition of the cap analogue are highlighted in red and those residues which are part of the water bridge recognition module as identified in this study are shown in green. See Figure 5.9 for complete alignment.

		40	50	60	70
2W97: Human	:	PEH.....YI.....KEPLQNRWALWFF.KND...K...S...KT	QANLRLISKFDTVE		
1L8B: Mouse	:	KEPLQNRWALWFF.KND...K...S...KT	QANLRLISKFDTVE		
4UEC: Fruitfly	:	PHM.....KEPLHNVWTLWYL.END...R...S...KS	EDHQNEITSFDTVE		
1AP8: Yeast	:	...MSVEEVSKKFEENVSVDDTTATPKTVL.SDSAEFDVKEPLNTKWTWYTKPAVD...K...S...ES	SDLLRPVTSFTQVE		
2WMC: Pea	:	PELLENSWTFWFD.TPA.AKSKQ....AA	GSSMRPIYTFSTVE		
2IDV: Wheat	:	AHPLENANTFWFD.NPQ.....GKSQRQA	GSTIHPHTFTSTVE		
3M93: PigRoundworm	:	M.....REPLQCHWALWYL.KAD...R...S...KD	EDCLKQVAVFDTVE		
3HXG: BloodFluke	:	EF.....PEPLQDSWSYLYF.QFR...K...A...LD	DECLKVAFTSTIE		
5ABX: Roundworm	:	T.....REPLQNRWALWYL.KAD...R...N...KE	EDCLKHVSIFDTVE		

		80	90	100	110	120	130	140
2W97: Human	:	DFWALYNHI.QLSSNLMFGCDYSLPK.DGIEPM	DEKN.KRGG	WLITLTKQQR..RSDLD	RFWLETLCLIGESFD.D.YSD			
1L8B: Mouse	:	DFWALYNHI.QLSSNLMFGCDYSLPK.DGIEPM	DEKN.KRGG	WLITLTKQQR..RSDLD	RFWLETLCLIGESFD.D.YSD			
4UEC: Fruitfly	:	DFWSLYNHI.KPPSEIKLGSBYSLPK.KNIRPM	DAAN.KQGG	NVITLTK...S.SKTDL	DNWLVLCLIGEAFTD...HSD			
1AP8: Yeast	:	EFWAIIQ.NIPEPHELPLKSDYHVRND.VRPE	DEAN.AKGG	WSFQLRG...K...GADID	ELWLRTLLAVIGEFTD.E.DDS			
2WMC: Pea	:	EFWSIYNNI.BHPGKLAVGADFYCPK.HKIEPK	DDPIC.ANGG	WTANYPK...G...KSDTS	WLYTLLAMIGEQFD...EGD			
2IDV: Wheat	:	DFWGLYNNI.BNPSKLNVGABFECFK.NKIEPK	DDPIS.ANGG	NTISCGR...G...KSDTF	WLHTLLAMIGEQFD...FGD			
3M93: PigRoundworm	:	DFWSLYNHI.QAASGLTWGSDYLYPK.EGIKPM	DDENN.VKGG	NLVVVVKQKR..AQLD	DEYLELLHMAIGEQFE.D.NGE			
3HXG: BloodFluke	:	DFWSVLTH.TVRPREITYGKDLVMPK.SDIMPK	DDPNK.ENG	GWLINVT.A.RQ...DVD	FLWDELLHLLIGSDWTDDEDR			
5ABX: Roundworm	:	DFWSLYNHI.QSAGGLNWGSDYLYPK.EGIKPM	DDVNN.VQGG	NLVVVVD.T..Q...L	DEYLELLHMAIVGEQFD.E.YGD			

		150	160	170	180	190	200	210
2W97: Human	:	DVCGAVVNVAKGDIAI	TTECENREAVTHIGRVYKERLGLP.P.KIVIGYQS	ADTATKSG.S.TTKNRFVV....				
1L8B: Mouse	:	DVCGAVVNVAKGDIAI	TTECENRDVTHIGRVYKERLGLP.P.KIVIGYQS	ADTATK.....KNRF..VV..				
4UEC: Fruitfly	:	QICGAVINIGSKNSISIT	TADGNNEEAALIEGKLRDALRG.R.NNSLQYQL	KDTMV.....				
1AP8: Yeast	:	QINGVVLISIKGGNFALTKS.EDKEPLLRIGGKFKQVKKLTDDGHE	..FFPSSANG...R..EPQPSIT..L...					
2WMC: Pea	:	EICGAVVNVGGRAGISIT	TKNASNEAAQVSIGKQWKEFLDYN.E.THG..FIF	DDAR...R.NAKNKYV..V...				
2IDV: Wheat	:	EICGAVVNVQKQZVAI	TKNAANEAAQVSIGKQWKEFLDYK.D.SIG..FIVE	DAKR...SDKGPKNRYT..V...				
3M93: PigRoundworm	:	YICGAVVNVQKGDVSL	TRDSLKDDVNLRIQILKAKLEIP.D.TEPIRYE	VT.....GSMVKPRIV..IPSK				
3HXG: BloodFluke	:	QICGAVVNVSRGSLSV	LTSDNEETILSIGRRIKERLELE.D.TIY..FQPVSD	QRSGTRGSDICTGKYEI.....				
5ABX: Roundworm	:	YICGAVVNVQKGDVSL	TRDATRDDVNLRIQVQLKQLSIP.D.TEILRYE	VDSSA.....KPRIC..L...				

Figure 5.9. Structure based sequence alignment of eIF4E proteins. The non-redundant sets of eIF4E proteins from different organisms whose structures have been solved were selected for the alignment which was performed using the program MUSTANG (Konagurthu et al, 2006). The PDB entry of each structure and the respective organism are shown. Residue number is shown with reference to human eIF4E (Uniprot ID: P06730). The structurally equivalent residues which are involved in the recognition of the cap analogue are highlighted in red with black background and those residues which are part of the water bridge recognition module as identified in this study are shown in green with black background.

I next examined if the spatial location of N155 and W166 residues themselves will have an influence on the interactions and hence retention of the water molecules at W1 and W2 positions. For this, the dynamic property of residues in the vicinity of the cap-binding interface is analyzed by comparing their fluctuations in the ensemble of structures generated from the MD simulations of the protein in the cap-bound and cap-free states (Figure 5.8A). Interestingly, a clear distinction was observed for a subset of residues in which D90, F94, R112, N155, R157, K162, W166 and H200 show a similar degree of flexibility between the two states whereas W56, W102, and E103 have reasonably high fluctuations. Crystal structure comparisons between the two structural states also highlight this difference (Figure 5.8B). Based on the observations, these residues in the binding interface can be categorized into two sub-groups comprising those that are dynamic and others which are relatively rigid, irrespective of whether the ligand is bound to the protein or not. It is noted that the amino acids which have a preformed spatial arrangement are those which are part of the water network with waters W1 and W2. N155 makes favorable electrostatic interactions with D90 and R157 whereas W166 is involved in hydrophobic stacking contacts with F94 and H200. This group of residues forms two separate hubs comprising of “D90, R157, N155” and “W166, H200, F94” which appear to act as ‘pillars’ in supporting the water bridge network (Figure 5.8C). Further, as stated previously, the critically important residues in the cap binding interface of eIF4E are W56, W102, E103, R112, K162 and R157 (Figure 5.1). This set of residues, along with those forming the solvent bridge and the pillars, are also evolutionarily conserved or at least their chemical

potential is preserved (Figure 5.8D). Hence this could likely compose the basic structural scaffold involved during the docking of the mRNA cap and hence form the nucleus of the recognition process.

Figure 5.10: Population distribution and water occupancy maps of wild and mutant eIF4E. The population distribution for residues (A) N155 and (C) W166 was measured in terms of the χ^2 side-chain torsion angles (measured along CA, CB, CG and CD) in MD simulations of both wild type and mutant states of eIF4E. N155 undergoes significant change in its spatial position between the wild type and mutant state simulations whereas W166 is relatively stable. Water occupancy map highlighting the presence / absence of waters at position W1 and W2 of (B) D90A_R157A and (D) F94A_H200A mutant proteins. The map is displayed at an isocutoff value of 2.5 times the peak value which is the mode of the isovalue distribution.

The influence of the residues (D90, F94, R157 and H200) that support W166 and N155 in forming the stable solvent bridge was also evaluated by *in silico* Ala mutations and MD simulations. The side-chain orientation of N155 is dramatically affected by D90A, R157A single or double point mutations as can be seen by the conformational distributions of the χ^2 angle of the residue (Figures 5.10A and 5.11A). This change in the spatial orientation of N155 results in the loss of water at W2 without affecting the occupancy at W1 (Figures 5.10B and 5.11B). Thus, the hydrophilic interaction network formed between residues D90-N155-R157 appears to be critical in order to maintain the distinct positioning of N155 and its interactions with water at W1. In contrast, the side-chain orientation of W166 does not change appreciably with F94A, H200A single or double mutations and consequently waters are detected at both W1 and W2 in these mutants (Figures 5.10C, 5.11C, 5.10D and 5.11D). This suggests that the hydrophobic packing formed by residues F94-W166-H200 is not primarily important to preserve the structural positioning of W166. However there are variations in the fluctuations of W166 between the wild type and mutant structures, which indicate that F94 and H200 do effect the optimal arrangement of the residue and hence appear to be part of a wider scaffold for supporting W166. Thus the supporting residues (D90, F94, R157 and H200) are important in order to anchor N155 and W166 in their positions albeit to different degrees; consequently, they result in the retention of the water bridge and subsequently in the maintenance of the local structural framework. Further, the evolutionary conservation of this water mediated building

block across different organisms indicates its significance in the structure-activity property of eIF4E.

Figure 5.11: Population distribution and water occupancy maps of wild type and mutant eIF4E. The population distributions for residues (A) N155 and (C) W166 were measured in terms of the χ^2 side-chain torsion angle (measured along CA, CB, CG and CD) in the MD simulations of both wild and mutant states of eIF4E. N155 undergoes significant change in its spatial position between the wild type and mutant state simulations whereas W166 is relatively stable. Water occupancy maps highlighting the presence / absence of waters at position W1 and W2 of (B) D90A and R157A and (D) F94A and H200A mutant proteins. The map is displayed at an isocutoff value of 2.5 times the peak value which is the mode of the isovalue distribution.

5.3.5 Crystal structure of apo eIF4E supports the presence of the water-bridge recognition module

Figure 5.12: Crystal structure of Apo eIF4E and water bridge recognition module at the cap binding interface of eIF4E (A) Cartoon view of the structure is shown. The secondary structure elements are labeled. (B) Structural superimposition of the three known Apo structures of eIF4E (PDB IDs: 5BEA, 3TF2 and 2IDR). The loop regions which show maximum variation among the three structures are shown in orange color and labeled as loops connecting the respective secondary structure elements. The three critical cap binding residues which are part of these regions are also shown. (C) Electron density of selected residues and water molecules present in the cap binding interface of the resolved eIF4E Apo structure. (D) Surface representation of the Apo eIF4E structure. The crystal waters as resolved in the structure are shown in red colored spheres whereas the two structurally conserved waters at positions W1 and W2 are shown in magenta color. The interactions involving residue W166, waters W1, W2 and residue N155 which form the water bridge, the hydrophilic (D90, N155 and R157) and hydrophobic (H200, W166 and F94) interaction network that together could act as preformed recognition module, is represented. The other critical residues which are recognized to be involved in the binding of the cap are shown.

Unit cell dimensions (Å)	a = 121.695, b= 128.024, c = 38.469 $\alpha = 90^\circ, \beta = 90^\circ, \gamma = 90^\circ$
Resolution (Å)	20.00-1.97, (2.04-1.97)
Space group	C 2 2 2 ₁
Temp (K)	100
Redundancy	4.5, (4.5)
Unique reflections	21660, (2145)
R Sym (%)	0.093, (0.324)
I/sigmaI	15.2, (4.4)
R factor (%)	23.05
R free (%)	25.42
RMS bonds (Å)	0.0064
RMS angle (°)	1.228
% Completeness	99.7, (100)
Wilson B-Factor (Å ²)	15.3
Average B Value (Å ²)	
Chain A	19.87
Water	25.75
Number of Solvent Molecules	109
Ramachandran Data	
Favored Region (%)	96.6%
Residues in Allowed Region (%)	2.8%
Residues in Outlier Region (%)	0.6%

Table 5.5: Crystallographic data collection and refinement statistics.

Apo eIF4E was successfully crystallized in the space group C2221 with one copy in the asymmetric unit in the absence of cap-analogue at an atomic resolution of 2.0 Å (Table 5.5). Besides the structure reported here, there are only two additional apo structures (structures without peptide or m⁷GTP cap bound) of eIF4E (PDB IDs: 3TF2 and 2IDR, Table 1) known, which is striking considering the relatively higher number of binary and ternary complex state structures of the protein (Table 5.1). The overall fold is consistent with other existing structures of eIF4E comprising of eight-stranded anti-parallel β sheet and four alpha helices (Figure 5.12A). Comparative analysis among the three apo structures shows that the regions connecting beta strands 1-2, 3-4 and 7-8 show considerable variations

(Figure 5.12B). These regions are in the cap-binding interface of the protein and include W56, W102 and E103 residues which are critical for cap binding. A considerable amount of resolved water molecules that are well defined in the 2Fo-Fc electron density map are seen in the structure and more significantly, the cap-binding interface has a total of 21 distinct crystal waters. The hydration profile at the interface region was analyzed and it was exciting to observe clear electron density for both the waters at W1 and W2 (Figure 5.12C). They are involved in the formation of the hydrogen-bond network with W166 and N155 and hence structure the water-bridge. This structure is included as one of the representatives for the cap-free data set in the crystal structure analysis (PDB ID: 5ABI, Table 5.1).

The 2Fo-Fc electron density map also revealed that residues W56 and W102, which sandwich the guanosine moiety of the m7GTP cap analogue in the cap-bound states, are significantly displaced. Hence they do not show a well-defined density profile in the map. W102 rotates out of the cap-binding interface and the loop bearing W56 swings out of the binding pocket. However, residues D90, R157 and N155 which form the hydrophilic triad and residues H200, F94 and W166 which pack against each other along with R112 are observed to have a distinct density which suggests a more structured nature of this set of residues. Thus, the crystallization data further reconfirms the presence of a water-mediated recognition module at the cap-binding region of eIF4E formed by protein residues (D90, R157, N155, H200, F94 and W166) and two specific water molecules and this further substantiates the current findings from comparative structural and MD analysis.

5.4 Conclusion

The role of water in mediating the interactions between the 5' cap of mRNA molecules and eIF4E has long been discussed and several crystal structures show a clear presence of the water molecules at the interface ([Brown et al, 2007](#); [Marcotrigiano et al, 1997](#); [Niedzwiecka et al, 2002](#); [Tomoo et al, 2003](#)). Some of these waters are spatially conserved across the different structures and are engaged in the formation of an elaborate hydrogen-bond network between the cap and the protein. Interestingly, a subset of these waters is structurally preserved at the interface even in the absence of the ligand. This is especially true for two bound waters which are present in a crevice located in the central portion of the cap-binding interface ([Figure 5.7D](#)). These waters are less mobile and more ordered as compared to other structurally conserved waters in the interface. They form a water-mediated interaction connecting the side-chains of residues N155 and W166. Both side chains seem crucial in maintaining local hydration which is lost if these residues are mutated. Besides, the spatial arrangement of these two residues also appears to be an important factor that would conserve the occupancy of the specific waters in the cleft. This conformational restraint on N155 and W166 is imposed by sub-networks formed around each of the two residues involving hydrophobic (F94-W166-H200) and hydrophilic interactions (R157-N155-D90). It is also interesting to see that this group of residues exhibits low degree of fluctuation and is relatively restrained in the cap-free state of eIF4E as compared to other residues, which are involved in the recognition of the cap. These residues along with the two water molecules form a local network in the cap-binding interface of eIF4E that provides

a structural scaffold for recognition of the cap ([Figure 5.7D](#)). Further, these sets of residues are conserved across different species suggesting that the water-bridge recognition module has been evolutionarily preserved.

Electrostatic interactions are one of the major driving forces for the attraction of the cap analogue towards the eIF4E binding interface ([Blachut-Okrasinska et al, 2007](#); [Niedzwiecka et al, 2002](#)). Biophysical and kinetic studies on the binding of the mRNA cap with eIF4E suggest a two-step binding mechanism ([Blachut-Okrasinska et al, 2007](#); [Niedzwiecka et al, 2002](#)). The phosphate groups in the cap act as an initial anchor followed by the formation of the fundamentally important cation- π stacking and hydrogen-bond interactions between W56, W102, E103 residues of eIF4E and the m7G guanosine moiety of the cap ([Niedzwiecka et al, 2002](#)). It is highly probable that the electrostatically steered mRNA cap encounters the water-bridge network identified in this work as an initial docking structural motif for the ligand. It is reported previously ([Niedzwiecka et al, 2002](#)) and also seen from the current comparative structural and simulation data as reported here, that the cap-binding event results in considerable uptake of water molecules into the binding interface. It can be hypothesized that the two structurally conserved waters in the bridge recruit other water molecules, thus, mobilizing the formation of an efficient interaction between the cap and the protein. They act as a strong framework on which to form the higher order interaction network. As such, it can be postulated that these waters are part of the other conserved residues in the interface that will be critical in the selectivity of the mRNA cap analogues. These specific waters and the network are

also seen in a new crystal structure of the apo state of the protein, which lends further credibility to the current proposition and the possible role of these waters in the process of recognition. Interestingly, a pair of structured water molecules present in the active site of the protein tyrosine kinase has been recently demonstrated to be critical for defining the selectivity of kinase inhibitor bosutinib towards its target. Further, conservation of two specific hydration sites in the protein Pin1 and its homologues has been shown to be critical components of the hydrogen-bond network in the active site and also for the stability of the helix-turn-coil structural motif ([Barman et al, 2015](#)).

The cap binding interface on eIF4E has been a target for the design and development of efficient inhibitors against the protein ([Chen et al, 2012](#); [Ghosh et al, 2009](#); [Ghosh et al, 2005](#)). A recent study reported on a structure-guided approach to design a modified cap analogue that binds eIF4E with more potency ([Chen et al, 2012](#)). However the nature of the hydration profile had not been considered explicitly in the design and development of these inhibitors, despite the general understanding that they have a significant role in the recognition process. The effect of displacing waters from specific hydration sites on ligand binding has been demonstrated for Scytalone dehydratase and EGFR as seen in the previous chapter ([Chapter 4](#)) and in the Adenosine A2A receptor. Substitutions on the Adenosine A2A receptor inhibitor triazolympurine resulted in varying activities of the ligand based on whether it displaced stable or unstable waters in the binding interface ([Higgs et al, 2010](#)). Thus, the current findings on the spatial conservation of specific water molecules in the cap-binding interface and their involvement in

the formation of a local structural scaffold could serve as an important starting point in the rational design of more efficient inhibitors against eIF4E. Additionally, the cap-analogue scaffold could be modified accordingly to see if displacement of certain specific water positions in the binding interface could be exploited for favorable binding with the protein. In conclusion, it is interesting to observe that the cap binding interface of eIF4E, which is primarily formed by flexible loop regions of the protein, preserves a preformed scaffold involving structurally conserved waters and a set of residues, which could be essential for the initial recognition of the ligand.

Chapter 6: Conclusions: summary, critiques and future directions

This chapter summarizes the study of the role of water in protein structure and function that I have carried out and presented in this thesis. The study is divided into four main sections (each corresponding to [Chapters 2-5 respectively](#)): (1) Hydration metric to predict protein aggregation (2) Mechanism for the loss of function in oncogenic mutants of the tumor suppressor protein p53 in cancer (3) Efficient characterization of hydration properties in structural ensembles of proteins (4) Data mining reveals a water-bridge that mediates recognition of mRNA cap in the translation initiation factor eIF4E. Here, I discuss the conclusions, critiques and probable future directions for each of these topics.

6.1 Hydration metric to predict protein aggregation

6.1.1 Summary

In this chapter ([Chapter 2](#)), I designed a metric based on hydration to identify aggregation associated regions in proteins. Ariel Fernandez has identified ‘Dehydrons’ as backbone hydrogen bonds in proteins that tend to be hydrated. Dehydrons were observed at protein-protein interaction interfaces, hence it was suggested that these dehydrons occur predominantly at residues which are evolutionarily conserved. However, a large concentration of dehydrons has also been noted in aggregating proteins. Given that protein aggregation is usually

deleterious for cells, I wondered why the dehydron-residues have been conserved during evolution. This led me to a systematic enquiry of dehydron-residue conservation in a non-redundant dataset of proteins in the PDB. I observed that dehydrons occur at both, conserved and non-conserved residues. Moreover, while the conserved dehydron-residues occurred at protein-protein interaction interfaces, the non-conserved dehydron-residues appeared to be present at aggregation interfaces in proteins. A study of aggregating proteins led me to conclude that the non-conserved dehydron-residues have a role in protein aggregation. Thus, the knowledge of the non-conserved dehydron-residues can be useful in attempts made to inhibit protein aggregation.

6.1.2 Critiques and future directions

One of the limitations of this work is its dependence on the 3D structure of a protein. So, the method is not easily applicable for a protein whose 3D structure is not known. In such cases, the 3D structures of proteins need to be modeled accurately and metrics can be developed for assessing whether these models are good enough to enable the computation of dehydrons accurately. As a future direction, I have been working towards designing a free web server to compute dehydrons and evolutionary conservation of residues in an input protein structure. I believe that this will facilitate the rapid prediction of aggregation-associated regions in proteins by other scientists, test (computationally and experimentally) the reliability and validity of dehydron distributions as predictors of properties such as aggregation and other properties that have not yet been related to dehydrons,

and, predict new ways of understanding the links between protein structure, dynamics and function.

6.2 Mechanism for the loss of function in oncogenic mutants of the tumor suppressor p53 in cancer

6.2.1 Summary

In this chapter ([Chapter 3](#)), I investigated the mechanism governing the loss of function (through unfolding and aggregation) of a class of mutants of the p53 protein that are implicated in cancers. The motivation behind this was to understand how p53 unfolds, as this will guide the design of inhibitors to restore the native state in the mutant forms of the protein, a significant unmet clinical need; although there are some compounds in trials ([Khoo et al, 2014](#); [Zandi et al, 2011](#)), their mechanism of action isn't exactly clear. To try and develop a mechanism for unfolding, I performed extensive all atom MD simulations of the WT and the mutant forms of p53 DBD. The analysis revealed a possible sequence of conformational changes that leads to the exposure of a hydrophobic region (APR) in the DBD. Additionally, I identified a novel 'druggable' pocket in the mutant forms of the protein that can be probed to evaluate if it is possible to restore the native structure in mutant p53 by targeting this pocket with small molecules.

6.2.2 Critiques and future directions

I am in the process of screening compounds against the identified pocket with the goal to identify binders that may guide the development of molecules to restabilize mutant p53. For this purpose, I am screening a Natural Product Library

(~2500 compounds), a library of FDA approved drugs (~2500 compounds) and a library of pain killers; the latter two libraries are being explored with the aim of repurposing clinically approved drugs which is increasingly gaining popularity ([Huang et al, 2011](#)). Additionally, I have also been designing peptides to target this pocket in a structure based design approach. Peptide design is driven by observations made while examining the conformations adopted during the MD simulations of mutant p53 DBD with an extended N-terminus (residues 91-293). In these simulations, I observed that the N-terminus of DBD displaces from its crystal-like native position into a non-native position; however when I extract and dock a dipeptide made up of the two N-terminal residues of the DBD (Trp91 and Pro92), into the identified pocket ([Figure 6.1](#)), they remain stably bound during the MD simulations. Longer simulations are needed to see if the conformations adopted by the mutants revert back to the wild type conformation in the presence of this dipeptide and its analogues including longer cyclic versions that I am currently designing. In addition, the coordinates of the partially unfolded mutants that are seen in these MD simulations will be made available to the scientific community to enable the development of different restabilizing molecules. A major limitation of the work carried out thus far is the lack of an experimental validation of the hypotheses; attempts are being made to establish experimental collaborations.

Figure 6.1: Final frames from a 500 ns MD simulations of (A) WT and (B) G145S (destabilizing) mutant of p53 DBD are shown. The N-terminus residues Trp91 and Pro92 are shown in sticks. While the N-terminus of the wild type DBD samples a crystal-like native conformation, the G245S mutant assumes a non-native conformation and fits in the S6-S7 pocket formed during the course of the simulations.

During the course of running the additional simulations, we also developed two testable hypotheses to examine the validity of the pathway of unfolding of the DBD mutants that is proposed from our studies.

- (a) We examined the unfolding pathways and identified two sites for mutations to Cys residues in order to generate a disulphide bridge with the hypothesis that it will prevent unfolding. We demonstrate this by making in silico mutants P98C and T211C (see Figure 2) and joining them with a disulphide bond and carrying of further MD simulations on the wild type and two mutants (V143A and R175H). It is clear that the proteins do not unfold and that the S6-S7 turn retains the (wild type like) '*closed*' state of the S6-S7 turn (Figure 6.2) in all the simulations. Additionally, the N-terminus is

more likely to sample a crystal like native conformation due to the constraints imposed by the newly introduced C98-C211 disulfide linkage.

Figure 6.2: The S6-S7 turn samples a ‘closed’ state in the wild type and destabilizing mutants of p53 DBD when a disulfide bond was modeled between the N-terminus and the S6-S7 turn by incorporating mutations- P98C and T211C

(b) The Nterminus of the DBD has a Tryptophan (W91) that makes a stabilizing interaction with R174 and has been shown to be critical for the stability of the DBD (Natan, E et al., 2011). A fluorescence study should show fluorescence quenching if this region were involved in the unfolding of the mutant DBD as suggested from our simulations. However, the DBD contains another Tryptophan residue (W146) which is present on strand S3 that is part of the β -sheet structure and its dynamics will overlap with those of W91. W146 is oriented towards the solvent and has not been reported to have a role in any protein-protein/DNA interactions. Therefore we decided to mutate it to a His, noting that W146 is non-conserved (Figure 6.3) and therefore mutating it to a Histidine (that is structurally and chemically similar to tryptophan) is unlikely to affect the structure and function of p53

DBD. This mutation should then enable a fluorescence based investigation of the pathway of unfolding by comparing the fluorescence spectra of WT (with H146) with those of the mutants (with H146).

Figure 6.3: Position of W146 in the structure of p53DBD. Multiple sequence alignment of W146 shows that this residue is less conserved across the homologs of p53.

6.3 Efficient characterization of hydration properties in structural ensembles of biomolecules

6.3.1 Summary

In this chapter (Chapter 4), I have developed a comprehensive software package- 'JAL' for efficient characterization of hydration properties in structural ensembles of biomolecules. The highlight of the package is a collection of three

novel algorithms that aim to: (1) Compute buried waters in an MD trajectory of proteins that can provide insights into protein stability, (2) Identify multi-water bridges in proteins using residence time calculations, and (3) Compute the displaceability of interfacial waters in the ligand binding sites of proteins using a structure based search. The likelihood of favorably displacing an interfacial water molecule is routinely computed during drug optimization. I demonstrated the utility of ‘JAL’ using case studies of 4 proteins for which a variety of relevant experimental data is available: the DNA Binding Domain of the transcription factor p53, the translation initiation factor eIF4E, Scytalone dehydratase which catalyzes the biosynthesis of melanin, and the kinase domain of the signaling protein EGFR. The ability to rationalize experimental data available for Scytalone dehydratase and EGFR kinase suggests that the algorithm has promise.

6.3.2 Critiques and future directions

Presently, the programs need to run faster and I am attempting to modify the programs to enable faster computations. This involves parallelizing the programs to run on multiple CPUs and optimizing the algorithms. Additionally, driven by the observation that there lacks a sufficiently big and clean dataset of conserved and displaceable waters (as pointed out in [section 4.1](#)), I am planning on building one. The program needs to be extended to a larger dataset of examples. Additionally, during the course of my investigation of the displaceability of interfacial waters, I observed that the more easily displaceable waters were the ones with low residence times while the ones which are hard to displace exhibit high residence times. Once the dataset of conserved and displaceable waters has

been developed, it would be interesting to investigate the nature and the extent of correlation between the residence time of water at a site on a protein and the likelihood of favorably displacing it. Finally, I plan to test a hypothesis that the residence times of waters around a solute such as a molecule or a peptide can be an indicator of its solubility; this can have significant implications for modulating the solubility of small molecules and peptides.

6.4 Data mining reveals a water-bridge that mediates recognition of mRNA cap in eIF4E

6.4.1 Summary

In this chapter ([Chapter 5](#)), I used a combination of comparative structural analysis, extensive MD simulations and X-ray crystallography to provide a detailed characterization of waters at the mRNA cap-binding site of the translation initiation factor eIF4E. Furthermore, I discovered the existence of a pre-formed water-bridge in eIF4E which appears to be critical for efficient recognition between the protein and the mRNA cap. The structural and functional importance of this unit in the cap-binding capacity of eIF4E is further substantiated by its evolutionarily conservation across different species. In addition to providing major new insights into the biological mechanisms of cap recognition, this work will also be critical in providing a fresh direction towards the design of potent specific compounds against this major therapeutic target.

6.4.2 Critiques and future directions

Future directions involve the computational design of a competitive inhibitor of eIF4E in a structure guided approach. Displacing the energetically unfavorable waters is known to contribute favorably to binding affinities of ligands/proteins. So, I have carried out a structure based search (as discussed in [Chapter 4](#)) of the 14 conserved waters in the mRNA cap binding site of eIF4E (reported in [Chapter 5](#)) to compute the displaceability of these waters. Additionally, the structure based search also provides information about which ligand groups can favorably displace these interfacial waters (also discussed in [Chapter 4](#)). This information will be complemented with fragment screening that is performed in the cap binding pocket ([Figure 6.2](#)), which shows that different fragments have mapped to different regions of the pocket. The inhibitor design efforts now entail linking these fragments together (in several combinations) to displace the unfavorable waters at the interface, and thereby enabling the inhibitor to compete out the natural ligand (m7GTP). I already have an experimental collaborator who will test these new molecules.

Figure 6.4: Fragment mapping (using FTMap (Bohnuud et al, 2012; Brenke et al, 2009; Kozakov et al, 2015; Kozakov et al, 2011)) on eIF4E shows the potential to grow inhibitors using fragments. The natural binding ligand (m7GMP) is shown in orange thicker sticks. Fragments (colored in blue, cyan and pink, shown in thinner sticks and marked with arrows) map to regions unexplored by the natural ligand and can guide the design of competitive inhibitors.

6.5 Concluding remarks

In summary, in this study I have developed algorithms to computationally probe the role of water in modulating the dynamics and function of biomolecules and used the findings to develop hypotheses regarding the role of hydration in biomolecular form and function. I demonstrated the utility of these approaches using the examples of the human transcription factor protein p53 and the translation initiation factor protein eIF4E where the findings provided new biological insights. Several testable hypotheses have emerged en route to my scientific enquiry which includes novel routes towards the development of novel inhibitors and drugs. I am confident that the work presented here will have a

positive and significant impact in driving future research in protein structural stability, interactions, dynamics and function.

APPENDIX

We have now run several additional simulations (26 simulations ranging from 100-500ns each) for the wild type and the destabilizing mutants of p53 studied in Chapter 3 of the thesis. It is clear that the conclusions arrived at in this chapter are reproduced again. [Figure A1](#) shows the data for these new simulations and it is clear that this distribution mirrors those in the thesis. This confirms the robustness of this study. In the [Figure A1](#), the distance used to characterize the conformational state of S6-S7 turn (shown below) again suggests that these mutants are likely to sample the ‘open’ state for S6-S7 turn as hypothesized in the thesis.

Figure A1: Quantifying the conformational state of S6-S7 turn of p53 DBD ((A) Wild type (B) E258V (C) R110L (D) R175H (E) R248Q (F) V143A) by measuring the distance between CZ atom of Arg209 (S6-S7 turn) and backbone carbonyl oxygen atom of Asp259 (S9-S10 turn). Different colors indicate the replicates.

AUTHOR'S PUBLICATIONS

1. Pradhan MR, Pal A, Hu Z, Kannan S, Kwoh CK, Lane DP, Verma CS.
Wetting of non-conserved residue-backbones: A feature indicative of aggregation associated regions of proteins. *Proteins* (2015);84(2):254-66
(*Impact factor: 2.5*)
2. Pradhan MR, Minh N, Ouray Z, Kannan S, Kwoh CK, Lane DP, Verma CS.
Dynamics of aggregating mutants of p53 DNA Binding Domain reveal novel “druggable” pocket. Manuscript under review in *Plos Comp Bio*
(2016) (*Impact factor: 4.6*)
3. Pradhan MR*, Nguyen MN, Kwoh CK, Verma CS. Efficient computation of hydration properties in structural ensemble of biomolecules, Manuscript under review in *J. Chem. Theory Comput.* (2016) (*Impact factor: 5.3*)
4. Lama D*, Pradhan MR*, Brown CJ, Kwoh CK, Lane DP, Verma CS.
Water-Bridge Mediates Recognition of mRNA Cap in eIF4E. Manuscript under review. (* shared authorship) Under review in *Structure* (2016)
(*Impact factor: 5.6*)
5. Pradhan MR, Minh N, Ouray Z, Kannan S, Kwoh CK, Lane DP, Verma CS.
Dynamics of aggregating mutants of p53 DNA Binding Domain reveal novel “druggable” pocket. *Biophys. J* (2016): (Conference abstract) (*Impact factor: 3.9*)

REFERENCES:

Abel R, Young T, Farid R, Berne BJ, Friesner RA (2008) Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society* 130: 2817-2831

Accordini SR, Morini MA, Sierra MB, Fris JA, Appignanesi GA, Fernandez A (2012) Wrapping mimicking in drug-like small molecules disruptive of protein-protein interfaces. *Proteins* 80: 1755-1765

Accordini SR, Rodriguez Fris JA, Appignanesi GA (2013) Wrapping effects within a proposed function-rescue strategy for the Y220C oncogenic mutation of protein p53. *PloS one* 8: e55123

Aguzzi A, O'Connor T (2010) Protein aggregation diseases: pathogenicity and therapeutic perspectives. *Nature reviews Drug discovery* 9: 237-248

Ahmed MH, Spyraakis F, Cozzini P, Tripathi PK, Mozzarelli A, Scarsdale JN, Safo MA, Kellogg GE (2011) Bound Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. *Plos One* 6: e24712

Allen FH (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B* 58: 380-388

Ang HC, Joerger AC, Mayer S, Fersht AR (2006) Effects of common cancer mutations on stability and DNA binding of full-length p53 compared with isolated core domains. *The Journal of biological chemistry* 281: 21934-21941

Ano Bom AP, Rangel LP, Costa DC, de Oliveira GA, Sanches D, Braga CA, Gava LM, Ramos CH, Cepeda AO, Stumbo AC, De Moura Gallo CV, Cordeiro Y, Silva JL (2012) Mutant p53 aggregates into prion-like amyloid oligomers and fibrils: implications for cancer. *The Journal of biological chemistry* 287: 28152-28162

Ashby JA, Stevenson CE, Jarvis GE, Lawson DM, Maule AJ (2011) Structure-based mutational analysis of eIF4E in relation to sbm1 resistance to pea seed-borne mosaic virus in pea. *PLoS One* 6: e15873

Ashur-Fabian O, Avivi A, Trakhtenbrot L, Adamsky K, Cohen M, Kajakaro G, Joel A, Amariglio N, Nevo E, Rechavi G (2004) Evolution of p53 in hypoxia-stressed Spalax

mimics human tumor mutation. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12236-12241

B. Balamurugan MNAMR, B. Shaahul Hameed, K. Sumathi, R. Senthilkumar, A. Udayakumar, K. H. Venkatesh Babu, M. Kalaivani, G. Sowmiya, P. Sivasankari, S. Saravanan, C. Vasuki Ranjani, K. Gopalakrishnan, K. N. Selvakumar, M. Jaikumar, T. Brindha, D. Michael and K. Sekar (2007) PSAP: protein structure analysis package. *Journal of Applied Crystallography*: 773-777

Bagchi B (2013) Water in Biological and Chemical Processes: From Structure and Dynamics to Function. *Cambridge University Press: Cambridge*

Bagchi K, Roy S (2014) Sensitivity of water dynamics to biologically significant surfaces of monomeric insulin: role of topology and electrostatic interactions. *The journal of physical chemistry B* 118: 3805-3813

Ball P (2003) Portrait of a molecule. *Nature* 421: 421-422

Ball P (2008a) Water as a biomolecule. *Chemphyschem : a European journal of chemical physics and physical chemistry* 9: 2677-2685

Ball P (2008b) Water as an active constituent in cell biology. *Chemical reviews* 108: 74-108

Barakat K, Issack BB, Stepanova M, Tuszyński J (2011) Effects of temperature on the p53-DNA binding interactions and their dynamical behavior: comparing the wild type to the R248Q mutant. *PloS one* 6: e27651

Barillari C, Taylor J, Viner R, Essex JW (2007) Classification of water molecules in protein binding sites. *Journal of the American Chemical Society* 129: 2577-2587

Barman A, Smitherman C, Souffrant M, Gadda G, Hamelberg D (2015) Conserved Hydration Sites in Pin1 Reveal a Distinctive Water Recognition Motif in Proteins. *J Chem Inf Model*

Baroni TE, Wang T, Qian H, Dearth LR, Truong LN, Zeng J, Denes AE, Chen SW, Brachmann RK (2004) A global suppressor motif for p53 cancer mutants. *Proceedings of the National Academy of Sciences of the United States of America* 101: 4930-4935

Bayden AS, Moustakas DT, Joseph-McCarthy D, Lamb ML (2015) Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. *Journal of chemical information and modeling* 55: 1552-1565

Beglov DR, B (1997) An Integral Equation to Describe the Solvation of Polar Molecules in Liquid Water. *J Phys Chem B* 101: 7821–7826

Berendsen HJC, Postma JPM, Van Gunsteren WF, DiNola A, Haak JR (1984) Molecular Dynamics with coupling to an external bath. *J Chem Phys* 81: 3684-3690

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic acids research* 28: 235-242

Bhattacharyya D, Nguyen K, Basu S (2014) Rationally induced RNA:DNA G-quadruplex structures elicit an anticancer effect by inhibiting endogenous eIF-4E expression. *Biochemistry* 53: 5461-5470

Blachut-Okrasinska E, Bojarska E, Stepinski J, Antosiewicz JM (2007) Kinetics of binding the mRNA cap analogues to the translation initiation factor eIF4E under second-order reaction conditions. *Biophys Chem* 129: 289-297

Blanden AR, Yu X, Wolfe AJ, Gilleran JA, Augeri DJ, O'Dell RS, Olson EC, Kimball SD, Emge TJ, Movileanu L, Carpizo DR, Loh SN (2015) Synthetic metallochaperone ZMC1 rescues mutant p53 conformation by transporting zinc into cells as an ionophore. *Mol Pharmacol* 87: 825-831

Boeckler FM, Joerger AC, Jaggi G, Rutherford TJ, Veprintsev DB, Fersht AR (2008) Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. *Proceedings of the National Academy of Sciences of the United States of America* 105: 10360-10365

Bohnuud T, Beglov D, Ngan CH, Zerbe B, Hall DR, Brenke R, Vajda S, Frank-Kamenetskii MD, Kozakov D (2012) Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde. *Nucleic acids research* 40: 7644-7652

Borden KL (2011) Targeting the oncogene eIF4E in cancer: From the bench to clinical trials. *Clin Invest Med* 34: E315

Bottoms CA, Smith PE, Tanner JJ (2002) A structurally conserved water molecule in Rossmann dinucleotide-binding domains. *Protein Sci* 11: 2125-2137

Brachmann RK, Yu K, Eby Y, Pavletich NP, Boeke JD (1998) Genetic selection of intragenic suppressor mutations that reverse the effect of common p53 cancer mutations. *The EMBO journal* 17: 1847-1859

Brandt T, Kaar JL, Fersht AR, Veprintsev DB (2012) Stability of p53 homologs. *PloS one* 7: e47889

Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* 25: 621-627

Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoseck M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. *Journal of computational chemistry* 30: 1545-1614

Brosh R, Rotter V (2009) When mutants gain new powers: news from the mutant p53 field. *Nature reviews Cancer* 9: 701-713

Brown CJ, Cheok CF, Verma CS, Lane DP (2011a) Reactivation of p53: from peptides to small molecules. *Trends in pharmacological sciences* 32: 53-62

Brown CJ, Lain S, Verma CS, Fersht AR, Lane DP (2009a) Awakening guardian angels: drugging the p53 pathway. *Nature reviews Cancer* 9: 862-873

Brown CJ, Lim JJ, Leonard T, Lim HC, Chia CS, Verma CS, Lane DP (2011b) Stabilizing the eIF4G1 alpha-helix increases its binding affinity with eIF4E: implications for peptidomimetic design strategies. *J Mol Biol* 405: 736-753

Brown CJ, McNae I, Fischer PM, Walkinshaw MD (2007) Crystallographic and mass spectrometric characterisation of eIF4E with N7-alkylated cap derivatives. *Journal of molecular biology* 372: 7-15

Brown CJ, Verma CS, Walkinshaw MD, Lane DP (2009b) Crystallization of eIF4E complexed with eIF4GI peptide and glycerol reveals distinct structural differences around the cap-binding site. *Cell Cycle* 8: 1905-1911

Brownlow S, Morais Cabral JH, Cooper R, Flower DR, Yewdall SJ, Polikarpov I, North AC, Sawyer L (1997) Bovine beta-lactoglobulin at 1.8 Å resolution--still an enigmatic lipocalin. *Structure* 5: 481-495

Buck PM, Kumar S, Wang X, Agrawal NJ, Trout BL, Singh SK (2012) Computational methods to predict therapeutic protein aggregation. *Methods in molecular biology* 899: 425-451

Bullock AN, Fersht AR (2001) Rescuing the function of mutant p53. *Nature reviews Cancer* 1: 68-76

Bullock AN, Henckel J, DeDecker BS, Johnson CM, Nikolova PV, Proctor MR, Lane DP, Fersht AR (1997) Thermodynamic stability of wild-type and mutant p53 core domain. *Proceedings of the National Academy of Sciences of the United States of America* 94: 14338-14342

Bullock AN, Henckel J, Fersht AR (2000) Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene* 19: 1245-1256

Bykov VJ, Issaeva N, Selivanova G, Wiman KG (2002a) Mutant p53-dependent growth suppression distinguishes PRIMA-1 from known anticancer drugs: a statistical analysis of information in the National Cancer Institute database. *Carcinogenesis* 23: 2011-2018

Bykov VJ, Issaeva N, Shilov A, Hultcrantz M, Pugacheva E, Chumakov P, Bergman J, Wiman KG, Selivanova G (2002b) Restoration of the tumor suppressor function to mutant p53 by a low-molecular-weight compound. *Nat Med* 8: 282-288

Cai Q, Xiao B, Si H, Cervini A, Gao J, Lu J, Upadhyay SK, Verma SC, Robertson ES (2012) Kaposi's sarcoma herpesvirus upregulates Aurora A expression to promote p53 phosphorylation and ubiquitylation. *PLoS Pathog* 8: e1002566

Carroll M, Borden KL (2013) The oncogene eIF4E: using biochemical insights to target cancer. *J Interferon Cytokine Res* 33: 227-238

Carugo O (2015) Structure and function of water molecules buried in the protein core. *Curr Protein Pept Sci* 16: 259-265

Case DA, Darden TA, Cheatham TE, 3rd, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Hayik S, Roitberg A, Seabra G, Swails J, Gotz AW, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wolf RM, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh M-J, Cui G, Roe DR, Mathews DH, Seetin MG, Salomon-Ferrer R, Sagui C, Babin V, Luchko S, Gusarov S, Kovalenko A, Kollman PA (2012) AMBER 12. *University of California, San Francisco*

Cellmer T, Bratko D, Prausnitz JM, Blanch HW (2007) Protein aggregation in silico. *Trends in biotechnology* 25: 254-261

Chaplin M (2006) Do we underestimate the importance of water in cell biology? *Nature reviews Molecular cell biology* 7: 861-866

Chen C, Gorlatova N, Kelman Z, Herzberg O (2011) Structures of p63 DNA binding domain in complexes with half-site and with spacer-containing full response elements. *Proceedings of the National Academy of Sciences of the United States of America* 108: 6456-6461

Chen JM, Xu SL, Wawrzak Z, Basarab GS, Jordan DB (1998) Structure-based design of potent inhibitors of scytalone dehydratase: displacement of a water molecule from the active site. *Biochemistry* 37: 17735-17744

Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica Section D, Biological crystallography* 66: 12-21

Chen X, Kopecky DJ, Mihalic J, Jeffries S, Min X, Heath J, Deignan J, Lai S, Fu Z, Guimaraes C, Shen S, Li S, Johnstone S, Thibault S, Xu H, Cardozo M, Shen W, Walker N, Kayser F, Wang Z (2012) Structure-guided design, synthesis, and evaluation of guanine-derived inhibitors of the eIF4E mRNA-cap interaction. *J Med Chem* 55: 3837-3851

Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL (2009) Design of therapeutic proteins with enhanced stability. *Proceedings of the National Academy of Sciences of the United States of America* 106: 11937-11942

Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annual review of biochemistry* 75: 333-366

Chiti F, Dobson CM (2009) Amyloid formation by globular proteins under native conditions. *Nature chemical biology* 5: 15-22

Chong SH, Ham S (2014) Interaction with the surrounding water plays a key role in determining the aggregation propensity of proteins. *Angewandte Chemie* 53: 3961-3964

Cohen SI, Vendruscolo M, Dobson CM, Knowles TP (2012) From macroscopic measurements to microscopic mechanisms of protein aggregation. *Journal of molecular biology* 421: 160-171

Cooper DR, Boczek T, Grelewski K, Pinkowska M, Sikorska M, Zawadzki M, Derewenda Z (2007) Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallogr D Biol Crystallogr* 63: 636-645

Csermely P (2001) Water and cellular folding processes. *Cellular and molecular biology* 47: 791-800

D.A. Case JTB, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman. (2015) AMBER 2015, University of California, San Francisco.

Darden T, York D, Pedersen L (1993a) Particle mesh Ewald-an $N\log(N)$ method for Ewald sums in large systems. *J Chem Phys* 98: 10089-10092

Darden T, York D, Pedersen L (1993b) Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of chemical physics* 98: 10089-10092

Das P, Kapoor D, Halloran KT, Zhou R, Matthews CR (2013) Interplay between drying and stability of a TIM barrel protein: a combined simulation-experimental study. *Journal of the American Chemical Society* 135: 1882-1890

Das P, Matysiak S (2012) Direct characterization of hydrophobic hydration during cold and pressure denaturation. *The journal of physical chemistry B* 116: 5342-5348

De Simone A, Dodson GG, Verma CS, Zagari A, Fraternali F (2005) Prion and water: tight and dynamical hydration sites have a key role in structural stability. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7535-7540

DeLano WL (2009) PyMOL (TM) Molecular Graphics System, Version 1.2r1. *DeLano Scientific LLC*

Derbyshire DJ, Basu BP, Serpell LC, Joo WS, Date T, Iwabuchi K, Doherty AJ (2002) Crystal structure of human 53BP1 BRCT domains bound to p53 tumour suppressor. *The EMBO journal* 21: 3863-3872

Dima RI, Thirumalai D (2002) Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein science : a publication of the Protein Society* 11: 1036-1049

Dodson GG, Lane DP, Verma CS (2008) Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO reports* 9: 144-150

Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic acids research* 35: W522-525

Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic acids research* 32: W665-667

Donald Hamelberg JAM. (2008) Dealing with bound waters in a site: Do they leave or stay? In Robert M Stroud JF-M (ed.), *Computational and structural approaches to drug discovery*, pp. 95-106.

Dua K, Williams TM, Beretta L (2001) Translational control of the proteome: relevance to cancer. *Proteomics* 1: 1191-1199

Dupradeau FY, Cezard C, Lelong R, Stanislawiak E, Pecher J, Delepine JC, Cieplak P (2008) R.E.DD.B.: a database for RESP and ESP atomic charges, and force field libraries. *Nucleic Acids Res* 36: D360-367

Eric Martz WD. (2014) Temperature value. *Proteopedia*, ISPC at the Weizmann Institute of Science in Israel.

Ethayathulla AS, Tse PW, Monti P, Nguyen S, Inga A, Fronza G, Viadiu H (2012) Structure of p73 DNA-binding domain tetramer modulates p73 transactivation. *Proceedings of the National Academy of Sciences of the United States of America* 109: 6066-6071

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology* 22: 1302-1306

Fernandez A (2004a) Functionality of wrapping defects in soluble proteins: what cannot be kept dry must be conserved. *Journal of molecular biology* 337: 477-483

Fernandez A (2004b) Keeping dry and crossing membranes. *Nature biotechnology* 22: 1081-1084

Fernandez A, Berry RS (2002) Extent of hydrogen-bond protection in folded proteins: a constraint on packing architectures. *Biophysical journal* 83: 2475-2481

Fernandez A, Crespo A (2008) Protein wrapping: a molecular marker for association, aggregation and drug design. *Chemical Society reviews* 37: 2373-2382

Fernandez A, Kardos J, Scott LR, Goto Y, Berry RS (2003) Structural defects and the diagnosis of amyloidogenic propensity. *Proceedings of the National Academy of Sciences of the United States of America* 100: 6446-6451

Fernandez A, Sanguino A, Peng Z, Ozturk E, Chen J, Crespo A, Wulf S, Shavrin A, Qin C, Ma J, Trent J, Lin Y, Han HD, Mangala LS, Bankson JA, Gelovani J, Samarel A, Bornmann W, Sood AK, Lopez-Berestein G (2007) An anticancer C-Kit kinase inhibitor is reengineered to make it more active and less cardiotoxic. *The Journal of clinical investigation* 117: 4044-4054

Fernandez A, Scheraga HA (2003) Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* 100: 113-118

Fernandez A, Scott R (2003) Dehydron: a structurally encoded signal for protein interaction. *Biophysical journal* 85: 1914-1928

Fischer S, Verma CS (1999) Binding of buried structural water increases the flexibility of proteins. *Proceedings of the National Academy of Sciences of the United States of America* 96: 9613-9615

Fiser A, Sali A (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics* 19: 2500-2501

Forest KT, Filutowicz MS (2003) Remodeling of replication initiator proteins. *Nature structural biology* 10: 496-498

Foster BA, Coffey HA, Morin MJ, Rastinejad F (1999) Pharmacological rescue of mutant p53 conformation and function. *Science* 286: 2507-2510

Foster TJ, MacKerell AD, Jr., Guvench O (2012) Balancing target flexibility and target denaturation in computational fragment-based inhibitor discovery. *Journal of computational chemistry* 33: 1880-1891

Fraser JA, Madhumalar A, Blackburn E, Bramham J, Walkinshaw MD, Verma C, Hupp TR (2010) A novel p53 phosphorylation site within the MDM2 ubiquitination signal: II. a model in which phosphorylation at SER269 induces a mutant conformation to p53. *The Journal of biological chemistry* 285: 37773-37786

Friedler A, Hansson LO, Veprintsev DB, Freund SM, Rippin TM, Nikolova PV, Proctor MR, Rudiger S, Fersht AR (2002) A peptide that binds and stabilizes p53 core domain: chaperone strategy for rescue of oncogenic mutants. *Proceedings of the National Academy of Sciences of the United States of America* 99: 937-942

Friedler A, Veprintsev DB, Hansson LO, Fersht AR (2003) Kinetic instability of p53 core domain mutants: implications for rescue by small molecules. *The Journal of biological chemistry* 278: 24108-24112

Gaiddon C, Lokshin M, Ahn J, Zhang T, Prives C (2001) A subset of tumor-derived mutant forms of p53 down-regulate p63 and p73 through a direct interaction with the p53 core domain. *Mol Cell Biol* 21: 1874-1887

Garcia-Sosa AT, Mancera RL, Dean PM (2003) WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J Mol Model* 9: 172-182

Ghosh B, Benyumov AO, Ghosh P, Jia Y, Avdulov S, Dahlberg PS, Peterson M, Smith K, Polunovsky VA, Bitterman PB, Wagner CR (2009) Nontoxic chemical interdiction of the epithelial-to-mesenchymal transition by targeting cap-dependent translation. *ACS Chem Biol* 4: 367-377

Ghosh P, Park C, Peterson MS, Bitterman PB, Polunovsky VA, Wagner CR (2005) Synthesis and evaluation of potential inhibitors of eIF4E cap binding to 7-methyl GTP. *Bioorg Med Chem Lett* 15: 2177-2180

Ghosh S, Ghosh D, Ranganathan S, Anoop A, P SK, Jha NN, Padinhateeri R, Maji SK (2014) Investigating the intrinsic aggregation potential of evolutionarily conserved segments in p53. *Biochemistry* 53: 5995-6010

Goldschmidt L, Teng PK, Riek R, Eisenberg D (2010) Identifying the amyloids, proteins capable of forming amyloid-like fibrils. *Proceedings of the National Academy of Sciences of the United States of America* 107: 3487-3492

Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28: 849-857

Gotz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC (2012) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of chemical theory and computation* 8: 1542-1555

Grossfield A. (2.0.7) WHAM: the weighted histogram analysis method. Vol. 2.0.7.

Gsponer J, Haberthur U, Caflisch A (2003) The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proceedings of the National Academy of Sciences of the United States of America* 100: 5154-5159

Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* 102: 15447-15452

Guvench O, MacKerell AD, Jr. (2009) Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput Biol* 5: e1000435

Harrington DJ, Adachi K, Royer WE, Jr. (1997) The high resolution crystal structure of deoxyhemoglobin S. *Journal of molecular biology* 272: 398-407

Helms V, Wade RC (1995) Thermodynamics of water mediating protein-ligand interactions in cytochrome P450cam: a molecular dynamics study. *Biophysical journal* 69: 810-824

Helms V, Wade RC (1998) Hydration energy landscape of the active site cavity in cytochrome P450cam. *Proteins* 32: 381-396

Higgs C, Beuming T, Sherman W (2010) Hydration Site Thermodynamics Explain SARs for Triazolylpurines Analogues Binding to the A2A Receptor. *ACS Med Chem Lett* 1: 160-164

Ho WC, Fitzgerald MX, Marmorstein R (2006) Structure of the p53 core domain dimer bound to DNA. *The Journal of biological chemistry* 281: 20494-20502

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65: 712-725

Hsieh AC, Ruggero D (2010) Targeting eukaryotic translation initiation factor 4E (eIF4E) in cancer. *Clin Cancer Res* 16: 4914-4920

Hu B, Lill MA (2014) WATsite: hydration site prediction program with PyMOL interface. *Journal of computational chemistry* 35: 1255-1260

Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen DT, Austin CP (2011) The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci Transl Med* 3: 80ps16

Huggins DJ, Payne MC (2013) Assessing the accuracy of inhomogeneous fluid solvation theory in predicting hydration free energies of simple solutes. *The journal of physical chemistry B* 117: 8232-8244

Hummer G, Tokmakoff A (2014) Preface: special topic on biological water. *J Chem Phys* 141: 22D101

Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *Journal of molecular graphics* 14: 33-38, 27-38

Hunt NT, Kattner L, Shanks RP, Wynne K (2007) The dynamics of water-protein interaction studied by ultrafast optical Kerr-effect spectroscopy. *Journal of the American Chemical Society* 129: 3168-3172

Ingram VM (1957) Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* 180: 326-328

Issaeva N, Friedler A, Bozko P, Wiman KG, Fersht AR, Selivanova G (2003) Rescue of mutants of the tumor suppressor p53 in cancer cells by a designed peptide. *Proceedings of the National Academy of Sciences of the United States of America* 100: 13303-13307

Jakalian A, Jack DB, Bayly CI (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of computational chemistry* 23: 1623-1641

Janowski R, Kozak M, Jankowska E, Grzonka Z, Grubb A, Abrahamson M, Jaskolski M (2001) Human cystatin C, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. *Nature structural biology* 8: 316-320

Jeszenoi N, Horvath I, Balint M, van der Spoel D, Hetenyi C (2015) Mobility-based prediction of hydration structures of protein surfaces. *Bioinformatics* 31: 1959-1965

Jia Y, Polunovsky V, Bitterman PB, Wagner CR (2012) Cap-dependent translation initiation factor eIF4E: an emerging anticancer drug target. *Med Res Rev* 32: 786-814

Joerger AC, Allen MD, Fersht AR (2004) Crystal structure of a superstable mutant of human p53 core domain. Insights into the mechanism of rescuing oncogenic mutations. *The Journal of biological chemistry* 279: 1291-1296

Joerger AC, Ang HC, Fersht AR (2006) Structural basis for understanding oncogenic p53 mutations and designing rescue drugs. *Proceedings of the National Academy of Sciences of the United States of America* 103: 15056-15061

Joerger AC, Ang HC, Veprintsev DB, Blair CM, Fersht AR (2005) Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. *The Journal of biological chemistry* 280: 16030-16037

Joerger AC, Bauer MR, Wilcken R, Baud MG, Harbrecht H, Exner TE, Boeckler FM, Spencer J, Fersht AR (2015) Exploiting Transient Protein States for the Design of Small-Molecule Stabilizers of Mutant p53. *Structure* 23: 2246-2255

Joerger AC, Fersht AR (2008) Structural biology of the tumor suppressor p53. *Annual review of biochemistry* 77: 557-582

Joerger AC, Rajagopalan S, Natan E, Veprintsev DB, Robinson CV, Fersht AR (2009) Structural evolution of p53, p63, and p73: implication for heterotetramer formation. *Proceedings of the National Academy of Sciences of the United States of America* 106: 17705-17710

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983a) Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* 79: 926-935

Jorgensen WL, Chandrasekhar J, Madura JD, Impey WR, Klein ML (1983b) Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* 79: 926-935

Kakudo Y, Shibata H, Otsuka K, Kato S, Ishioka C (2005) Lack of correlation between p53-dependent transcriptional activity and the ability to induce apoptosis among 179 mutant p53s. *Cancer Res* 65: 2108-2114

Kanaujia SP, Sekar K (2009) Structural and functional role of water molecules in bovine pancreatic phospholipase A(2): a data-mining approach. *Acta Crystallogr D Biol Crystallogr* 65: 74-84

Kazmirski SL, Isaacson RL, An C, Buckle A, Johnson CM, Daggett V, Fersht AR (2002) Loss of a metal-binding site in gelsolin leads to familial amyloidosis-Finnish type. *Nature structural biology* 9: 112-116

Khoo KH, Verma CS, Lane DP (2014) Drugging the p53 pathway: understanding the route to clinical efficacy. *Nature reviews Drug discovery* 13: 217-236

Kitayner M, Rozenberg H, Kessler N, Rabinovich D, Shaulov L, Haran TE, Shakked Z (2006) Structural basis of DNA recognition by p53 tetramers. *Mol Cell* 22: 741-753

Kitayner M, Rozenberg H, Rohs R, Suad O, Rabinovich D, Honig B, Shakked Z (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat Struct Mol Biol* 17: 423-429

Klein C, Georges G, Kunkele KP, Huber R, Engh RA, Hansen S (2001) High thermostability and lack of cooperative DNA binding distinguish the p63 core domain from the homologous tumor suppressor p53. *The Journal of biological chemistry* 276: 37390-37401

Klimov DK, Thirumalai D (2003) Dissecting the assembly of Abeta16-22 amyloid peptides into antiparallel beta sheets. *Structure* 11: 295-307

Knight JD, Hamelberg D, McCammon JA, Kothary R (2009) The role of conserved water molecules in the catalytic domain of protein kinases. *Proteins* 76: 527-535

Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins* 64: 559-574

Kovalenko AH, F (1998) Three-Dimensional Density Profiles of Water in Contact with a Solute of Arbitrary Shape: a RISM Approach. *Chem Phys Lett* 290: 237-244

Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, Xia B, Beglov D, Vajda S (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc* 10: 733-755

Kozakov D, Hall DR, Chuang GY, Cencic R, Brenke R, Grove LE, Beglov D, Pelletier J, Whitty A, Vajda S (2011) Structural conservation of druggable hot spots in protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* 108: 13528-13533

Kuffel A, Zielkiewicz J (2012) Why the solvation water around proteins is more dense than bulk water. *The journal of physical chemistry B* 116: 12113-12124

Kumar S, Rosenberg, J. M., Bouzida, D., Swendsen, R. H. and Kollman, P. A. (1995) Multidimensional free-energy calculations using the weighted histogram analysis method. *J Comput Chem* 16: 1339-1350

Lama D, Brown CJ, Lane DP, Verma CS (2015) Gating by Tryptophan 73 Exposes a Cryptic Pocket at the Protein-Binding Interface of the Oncogenic eIF4E Protein. *Biochemistry* 54: 6535-6544

Lama D, Modi V, Sankararamakrishnan R (2013) Behavior of solvent-exposed hydrophobic groove in the anti-apoptotic Bcl-XL protein: clues for its ability to bind diverse BH3 ligands from MD simulations. *PLoS One* 8: e54397

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* 26: 283-291

Lazaridis T (1998) Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *J Phys Chem B* 102: 3531–3541

Lazaridis T, Karplus M (1997) "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* 278: 1928-1931

Levinson NM, Boxer SG (2014) A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity. *Nature chemical biology* 10: 127-132

Levitt M, Park BH (1993) Water: now you see it, now you don't. *Structure* 1: 223-226

Levy Y, Onuchic JN (2004) Water and proteins: a love-hate relationship. *Proceedings of the National Academy of Sciences of the United States of America* 101: 3325-3326

Likic VA, Juranic N, Macura S, Prendergast FG (2000) A "structural" water molecule in the family of fatty acid binding proteins. *Protein Sci* 9: 497-504

Liu Q, Kaneko S, Yang L, Feldman RI, Nicosia SV, Chen J, Cheng JQ (2004) Aurora-A abrogation of p53 DNA binding and transactivation activity by phosphorylation of serine 215. *The Journal of biological chemistry* 279: 52175-52182

Liu W, Zhao R, McFarland C, Kieft J, Niedzwiecka A, Jankowska-Anyszka M, Stepinski J, Darzynkiewicz E, Jones DN, Davis RE (2009) Structural insights into parasite eIF4E binding specificity for m7G and m2,2,7G mRNA caps. *J Biol Chem* 284: 31336-31349

Liu X, Wilcken R, Joerger AC, Chuckowree IS, Amin J, Spencer J, Fersht AR (2013) Small molecule induced reactivation of mutant p53 in cancer cells. *Nucleic acids research* 41: 6034-6044

Loncharich RJ, Brooks BR, Pastor RW (1992) Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* 32: 523-535

Lubin DJ, Butler JS, Loh SN (2010) Folding of tetrameric p53: oligomerization and tumorigenic mutations induce misfolding and loss of function. *Journal of molecular biology* 395: 705-716

Lukman S, Lane DP, Verma CS (2013) Mapping the structural and dynamical features of multiple p53 DNA binding domains: insights into loop 1 intrinsic dynamics. *PloS one* 8: e80221

Madhusudhan MS, Vishveshwara S (2001) Deducing hydration sites of a protein from molecular dynamics simulations. *J Biomol Struct Dyn* 19: 105-114

Makarov VA, Andrews BK, Smith PE, Pettitt BM (2000) Residence times of water molecules in the hydration sites of myoglobin. *Biophysical journal* 79: 2966-2974

Mamane Y, Petroulakis E, Rong L, Yoshida K, Ler LW, Sonenberg N (2004) eIF4E--from translation to transformation. *Oncogene* 23: 3172-3179

Marcotrigiano J, Gingras AC, Sonenberg N, Burley SK (1997) Cocystal structure of the messenger RNA 5' cap-binding protein (eIF4E) bound to 7-methyl-GDP. *Cell* 89: 951-961

Marcotrigiano J, Gingras AC, Sonenberg N, Burley SK (1999) Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G. *Mol Cell* 3: 707-716

Marquart M, Walter, J., Deisenhofer, J., Bode, W., Huber, R. (1983) The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and its Complexes with Inhibitors. *Acta Crystallogr*, 39

Marquart M, Walter, J., Deisenhofer, J., Bode, W., Huber, R. (1983) The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and its Complexes with Inhibitors. *Acta Crystallogr, Sect B* 39: 480-490

Martinez L, Nascimento AS, Nunes FM, Phillips K, Aparicio R, Dias SM, Figueira AC, Lin JH, Nguyen P, Apriletti JW, Neves FA, Baxter JD, Webb P, Skaf MS, Polikarpov I (2009) Gaining ligand selectivity in thyroid hormone receptors via entropy. *Proceedings of the National Academy of Sciences of the United States of America* 106: 20717-20722

Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature methods* 7: 237-242

Maurus R, Overall CM, Bogumil R, Luo Y, Mauk AG, Smith M, Brayer GD (1997) A myoglobin variant with a polar substitution in a conserved hydrophobic cluster in the heme binding pocket. *Biochimica et biophysica acta* 1341: 1-13

McClendon CL, Friedland G, Mobley DL, Amirkhani H, Jacobson MP (2009) Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *Journal of chemical theory and computation* 5: 2486-2502

McClendon CL, Hua L, Barreiro A, Jacobson MP (2012) Comparing Conformational Ensembles Using the Kullback-Leibler Divergence Expansion. *Journal of chemical theory and computation* 8: 2115-2126

McClendon CL, Kornev AP, Gilson MK, Taylor SS (2014) Dynamic architecture of a protein kinase. *Proceedings of the National Academy of Sciences of the United States of America* 111: E4623-4631

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* 40: 658-674

McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernandez CX, Schwantes CR, Wang LP, Lane TJ, Pande VS (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical journal* 109: 1528-1532

Meplan C, Richard MJ, Hainaut P (2000) Metalloregulation of the tumor suppressor protein p53: zinc mediates the renaturation of p53 after exposure to metal chelators in vitro and in intact cells. *Oncogene* 19: 5227-5236

Merabet A, Houleberghs H, Maclagan K, Akanho E, Bui TT, Pagano B, Drake AF, Fraternali F, Nikolova PV (2010) Mutants of the tumour suppressor p53 L1 loop as second-site suppressors for restoring DNA binding to oncogenic p53 mutations: structural and biochemical insights. *Biochem J* 427: 225-236

Merrick WC (2004) Cap-dependent and cap-independent translation in eukaryotic systems. *Gene* 332: 1-11

Michel J (2014) Current and emerging opportunities for molecular simulations in structure-based drug design. *Phys Chem Chem Phys* 16: 4465-4477

Michel J, Tirado-Rives J, Jorgensen WL (2009a) Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *Journal of the American Chemical Society* 131: 15403-15411

Michel J, Tirado-Rives J, Jorgensen WL (2009b) Prediction of the water content in protein binding sites. *The journal of physical chemistry B* 113: 13337-13346

Milano T, Di Salvo ML, Angelaccio S, Pascarella S (2015) Conserved water molecules in bacterial serine hydroxymethyltransferases. *Protein Eng Des Sel* 28: 415-426

Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of molecular biology* 291: 177-196

Monzingo AF, Dhaliwal S, Dutt-Chaudhuri A, Lyon A, Sadow JH, Hoffman DW, Robertus JD, Browning KS (2007) The structure of eukaryotic translation initiation factor-4E from wheat reveals a novel disulfide bond. *Plant Physiol* 143: 1504-1518

Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* 30: 2785-2791

Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53: 240-255

Natan E, Baloglu C, Pagel K, Freund SM, Morgner N, Robinson CV, Fersht AR, Joerger AC (2011) Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. *Journal of molecular biology* 409: 358-368

Ng JW, Lama D, Lukman S, Lane DP, Verma CS, Sim AY (2015) R248Q mutation-Beyond p53-DNA binding. *Proteins*

Nguyen CN, Young TK, Gilson MK (2012) Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *The Journal of chemical physics* 137: 044101

Nguyen MN, Madhusudhan MS (2011) Biological insights from topology independent comparison of protein 3D structures. *Nucleic acids research* 39: e94

Nguyen MN, Tan KP, Madhusudhan MS (2011) CLICK--topology-independent comparison of biomolecular 3D structures. *Nucleic acids research* 39: W24-28

Nguyen MN, Verma C (2015) Rclick: a web server for comparison of RNA 3D structures. *Bioinformatics* 31: 966-968

Niedzwiecka A, Marcotrigiano J, Stepinski J, Jankowska-Anyszka M, Wyslouch-Cieszyńska A, Dadlez M, Gingras AC, Mak P, Darzynkiewicz E, Sonenberg N, Burley SK, Stolarski R (2002) Biophysical studies of eIF4E cap-binding protein: recognition of mRNA 5' cap structure and synthetic fragments of eIF4G and 4E-BP1 proteins. *Journal of molecular biology* 319: 615-635

Nikolova PV, Wong KB, DeDecker B, Henckel J, Fersht AR (2000) Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations. *The EMBO journal* 19: 370-378

Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P (2002) The IARC TP53 database: new online mutation analysis and recommendations to users. *Human mutation* 19: 607-614

Otsuka K, Kato S, Kakudo Y, Mashiko S, Shibata H, Ishioka C (2007) The screening of the second-site suppressor mutations of the common p53 mutants. *Int J Cancer* 121: 559-566

Otting G, Liepinsh E, Wuthrich K (1991) Protein hydration in aqueous solution. *Science* 254: 974-980

Padlan EA, Love WE (1985) Refined crystal structure of deoxyhemoglobin S. I. Restrained least-squares refinement at 3.0-Å resolution. *The Journal of biological chemistry* 260: 8272-8279

Pagano B, Jama A, Martinez P, Akanho E, Bui TT, Drake AF, Fraternali F, Nikolova PV (2013) Structure and stability insights into tumour suppressor p53 evolutionary related proteins. *PloS one* 8: e76014

Panman MR, Bakker BH, den Uyl D, Kay ER, Leigh DA, Buma WJ, Brouwer AM, Geenevasen JA, Woutersen S (2013) Water lubricates hydrogen-bonded molecular machines. *Nature chemistry* 5: 929-934

Papadopoulos E, Jenni S, Kabha E, Takrouri KJ, Yi T, Salvi N, Luna RE, Gavathiotis E, Mahalingam P, Arthanari H, Rodriguez-Mias R, Yefidoff-Freedman R, Aktas BH, Chorev M, Halperin JA, Wagner G (2014) Structure of the eukaryotic translation initiation factor eIF4E in complex with 4EGI-1 reveals an allosteric mechanism for dissociating eIF4G. *Proc Natl Acad Sci U S A* 111: E3187-3195

Park JH, Liu Y, Lemmon MA, Radhakrishnan R (2012) Erlotinib binds both inactive and active conformations of the EGFR tyrosine kinase domain. *Biochem J* 448: 417-423

Park S, Saven JG (2005) Statistical and molecular dynamics studies of buried waters in globular proteins. *Proteins* 60: 450-463

Pastor RW, Brooks BR, Szabo A (1988) An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol Phys* 65: 1409-1419

Patel H, Gruning BA, Gunther S, Merfort I (2014) PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering. *Bioinformatics* 30: 2978-2980

Pawlicki S, Le Behec A, Delamarche C (2008) AMYPdb: a database dedicated to amyloid precursor proteins. *BMC bioinformatics* 9: 273

Peter D, Igreja C, Weber R, Wohlbold L, Weiler C, Ebertsch L, Weichenrieder O, Izaurralde E (2015a) Molecular architecture of 4E-BP translational inhibitors bound to eIF4E. *Mol Cell* 57: 1074-1087

Peter D, Weber R, Kone C, Chung MY, Ebertsch L, Truffault V, Weichenrieder O, Igreja C, Izaurralde E (2015b) Mextli proteins use both canonical bipartite and novel tripartite binding modes to form eIF4E complexes that display differential sensitivity to 4E-BP regulation. *Genes Dev* 29: 1835-1849

Peters MB, Yang Y, Wang B, Fusti-Molnar L, Weaver MN, Merz KM, Jr. (2010) Structural Survey of Zinc Containing Proteins and the Development of the Zinc AMBER Force Field (ZAFF). *Journal of chemical theory and computation* 6: 2935-2947

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* 25: 1605-1612

Prakash P, Sayyed-Ahmad A, Gorfe AA (2012) The role of conserved waters in conformational transitions of Q61H K-ras. *PLoS Comput Biol* 8: e1002394

Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, Hess B, Lindahl E (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29: 845-854

Quiocho FA, Wilson DK, Vyas NK (1989) Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature* 340: 404-407

Rajagopalan S, Huang F, Fersht AR (2011) Single-Molecule characterization of oligomerization kinetics and equilibria of the tumor suppressor p53. *Nucleic acids research* 39: 2294-2303

Raman EP, Yu W, Guvench O, Mackerell AD (2011) Reproducing crystal binding modes of ligand functional groups using Site-Identification by Ligand Competitive Saturation (SILCS) simulations. *Journal of chemical information and modeling* 51: 877-896

Rangel LP, Costa DC, Vieira TC, Silva JL (2014) The aggregation of mutant p53 produces prion-like properties in cancer. *Prion* 8: 75-84

Raschke TM (2006) Water structure and interactions with protein surfaces. *Curr Opin Struct Biol* 16: 152-159

Raymer ML, Sanschagrin PC, Punch WF, Venkataraman S, Goodman ED, Kuhn LA (1997) Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *Journal of molecular biology* 265: 445-464

Reddy G, Straub JE, Thirumalai D (2009) Dynamics of locking of peptides onto growing amyloid fibrils. *Proceedings of the National Academy of Sciences of the United States of America* 106: 11948-11953

Rhoads RE (2009) eIF4E: new family members, new binding partners, new roles. *J Biol Chem* 284: 16711-16715

Ross CA, Poirier MA (2005) Opinion: What is the role of protein aggregation in neurodegeneration? *Nature reviews Molecular cell biology* 6: 891-898

Ross GA, Bodnarchuk MS, Essex JW (2015) Water Sites, Networks, And Free Energies with Grand Canonical Monte Carlo. *Journal of the American Chemical Society* 137: 14930-14943

Ross GA, Morris GM, Biggin PC (2012) Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS one* 7: e32036

Royer WE, Jr., Pardanani A, Gibson QH, Peterson ES, Friedman JM (1996) Ordered water molecules as key allosteric mediators in a cooperative dimeric hemoglobin. *Proceedings of the National Academy of Sciences of the United States of America* 93: 14526-14531

Ryckaert J-P, Ciccotti G, Berendsen HJ (1977a) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 23: 327-341

Ryckaert J-P, Ciccotti G, Berendsen HJC (1977b) Numerical Integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J Comput Phys* 23: 327-341

S. Praveen JR, P. Sivasankari, G. Sowmiya and K. Sekar (2008) WAP (version 2.0): an updated computing and visualization server for water molecules. *Journal of Applied Crystallography* 41: 952–954

Sachs R, Max KE, Heinemann U, Balbach J (2012) RNA single strands bind to a conserved surface of the major cold shock protein in crystals and solution. *Rna* 18: 65-76

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* 234: 779-815

Sander C, Schneider R (1994) The HSSP database of protein structure-sequence alignments. *Nucleic acids research* 22: 3597-3599

Savage H, Wlodawer A (1986) Determination of water structure around biomolecules using X-ray and neutron diffraction methods. *Methods in enzymology* 127: 162-183

Schrodinger, LLC. (2010) The PyMOL Molecular Graphics System, Version 1.4r1.

Schuijjer M, Berns EM (2003) TP53 and ovarian cancer. *Human mutation* 21: 285-291

Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences of the United States of America* 102: 10147-10152

Sekiyama N, Arthanari H, Papadopoulos E, Rodriguez-Mias RA, Wagner G, Leger-Abraham M (2015) Molecular mechanism of the dual activity of 4EGI-1: Dissociating eIF4G from eIF4E but stabilizing the binding of unphosphorylated 4E-BP1. *Proc Natl Acad Sci U S A* 112: E4036-4045

Selivanova G, Wiman KG (2007) Reactivation of mutant p53: molecular mechanisms and therapeutic potential. *Oncogene* 26: 2243-2254

Selkoe DJ (2003) Folding proteins in fatal ways. *Nature* 426: 900-904

Setny P (2015) Prediction of Water Binding to Protein Hydration Sites with a Discrete, Semiexplicit Solvent Model. *Journal of chemical theory and computation* 11: 5961-5972

Setny P, Zacharias M (2010) Hydration in discrete water. A mean field, cellular automata based approach to calculating hydration free energies. *The journal of physical chemistry B* 114: 8667-8675

Shrake A, Rupley JA (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of molecular biology* 79: 351-371

Siddiqui N, Tempel W, Nedyalkova L, Volpon L, Wernimont AK, Osborne MJ, Park HW, Borden KL (2012) Structural insights into the allosteric effects of 4EBP1 on the eukaryotic translation initiation factor eIF4E. *J Mol Biol* 415: 781-792

Silva JL, Vieira TC, Gomes MP, Bom AP, Lima LM, Freitas MS, Ishimaru D, Cordeiro Y, Foguel D (2010) Ligand binding and hydration in protein misfolding: insights from studies of prion and p53 tumor suppressor proteins. *Accounts of chemical research* 43: 271-279

Smith KJ, Reid SW, Stuart DI, McMichael AJ, Jones EY, Bell JI (1996) An altered position of the alpha 2 helix of MHC class I is revealed by the crystal structure of HLA-B*3501. *Immunity* 4: 203-213

Soragni A, Janzen DM, Johnson LM, Lindgren AG, Thai-Quynh Nguyen A, Tiourin E, Soriaga AB, Lu J, Jiang L, Faull KF, Pellegrini M, Memarzadeh S, Eisenberg DS (2016) A Designed Inhibitor of p53 Aggregation Rescues p53 Tumor Suppression in Ovarian Carcinomas. *Cancer Cell* 29: 90-103

Sumathi K, Ananthalakshmi P, Roshan MN, Sekar K (2006) 3dSS: 3D structural superposition. *Nucleic acids research* 34: W128-132

Tame JR, Murshudov GN, Dodson EJ, Neil TK, Dodson GG, Higgins CF, Wilkinson AJ (1994) The structural basis of sequence-independent peptide binding by OppA protein. *Science* 264: 1578-1581

Tame JR, Sleight SH, Wilkinson AJ, Ladbury JE (1996) The role of water in sequence-independent ligand binding by an oligopeptide transporter protein. *Nature structural biology* 3: 998-1001

Tan YS, Sledz P, Lang S, Stubbs CJ, Spring DR, Abell C, Best RB (2012) Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1. *Angewandte Chemie* 51: 10078-10081

Tan YS, Spring DR, Abell C, Verma CS (2015) The Application of Ligand-Mapping Molecular Dynamics Simulations to the Rational Design of Peptidic Modulators of Protein-Protein Interactions. *Journal of chemical theory and computation* 11: 3199-3210

Tan YS, Spring, D. R., Abell, C. & Verma, C. S. (2015) The Application of Ligand-Mapping Molecular Dynamics Simulations to the Rational Design of Peptidic Modulators of Protein-Protein Interactions. *J Chem Theory Comput* 11: 3199-3210

Thanki N, Thornton JM, Goodfellow JM (1988) Distributions of water around amino acid residues in proteins. *Journal of molecular biology* 202: 637-657

Thirumalai D, Reddy G, Straub JE (2012) Role of water in protein aggregation and amyloid polymorphism. *Accounts of chemical research* 45: 83-92

Thumma SC, Jacobson BA, Patel MR, Konicek BW, Franklin MJ, Jay-Dixon J, Sadiq A, De A, Graff JR, Kratzke RA (2015) Antisense oligonucleotide targeting eukaryotic translation initiation factor 4E reduces growth and enhances chemosensitivity of non-small-cell lung cancer cells. *Cancer Gene Ther* 22: 396-401

Tomoo K, Matsushita Y, Fujisaki H, Abiko F, Shen X, Taniguchi T, Miyagawa H, Kitamura K, Miura K, Ishida T (2005) Structural basis for mRNA Cap-Binding regulation of eukaryotic initiation factor 4E by 4E-binding protein, studied by spectroscopic, X-ray crystal structural, and molecular dynamics simulation methods. *Biochim Biophys Acta* 1753: 191-208

Tomoo K, Shen X, Okabe K, Nozoe Y, Fukuhara S, Morino S, Sasaki M, Taniguchi T, Miyagawa H, Kitamura K, Miura K, Ishida T (2003) Structural features of human initiation factor 4E, studied by X-ray crystal analyses and molecular dynamics simulations. *J Mol Biol* 328: 365-383

Tsolis AC, Papandreou NC, Ionomidou VA, Hamodrakas SJ (2013) A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. *PloS one* 8: e54175

Umezawa K, Higo J, Shimotakahara S, Shindo H (2007) Collective solvent flows around a protein investigated by molecular dynamics simulation. *The Journal of chemical physics* 127: 045101

Vanquelef E, Simon S, Marquant G, Garcia E, Klimerak G, Delepine JC, Cieplak P, Dupradeau FY (2011) R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res* 39: W511-517

Verdonk ML, Cole JC, Taylor R (1999) SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *Journal of molecular biology* 289: 1093-1108

Vijayan R, Sahai MA, Czajkowski T, Biggin PC (2010) A comparative analysis of the role of water in the binding pockets of ionotropic glutamate receptors. *Phys Chem Chem Phys* 12: 14057-14066

Virtanen JJ, Makowski L, Sosnick TR, Freed KF (2010) Modeling the hydration layer around proteins: HyPred. *Biophysical journal* 99: 1611-1619

Vogelstein B, Lane D, Levine AJ (2000) Surfing the p53 network. *Nature* 408: 307-310

Vriend G (1990) WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics* 8: 52-56, 29

Wade RC, Mazor MH, McCammon JA, Quiocho FA (1991) A molecular dynamics study of thermodynamic and structural aspects of the hydration of cavities in proteins. *Biopolymers* 31: 919-931

Wagner JR, Lee CT, Durrant JD, Malmstrom RD, Feher VA, Amaro RE (2016) Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chemical reviews* 116: 6370-6390

Walsh I, Seno F, Tosatto SC, Trovato A (2014) PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic acids research* 42: W301-307

Wang G, Dunbrack RL, Jr. (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic acids research* 33: W94-98

Wang G, Fersht AR (2012) First-order rate-determining aggregation mechanism of p53 and its implications. *Proceedings of the National Academy of Sciences of the United States of America* 109: 13590-13595

Wang G, Fersht AR (2015) Mechanism of initiation of aggregation of p53 revealed by Phi-value analysis. *Proceedings of the National Academy of Sciences of the United States of America* 112: 2437-2442

Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 25: 247-260

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J Comput Chem* 25: 1157-1174

Wang PL, Sait F, Winter G (2001) The 'wildtype' conformation of p53: epitope mapping using hybrid proteins. *Oncogene* 20: 2318-2324

Wang Y, Rosengarth A, Luecke H (2007) Structure of the human p53 core domain in the absence of DNA. *Acta crystallographica Section D, Biological crystallography* 63: 276-281

Warnock LJ, Raines SA, Milner J (2011) Aurora A mediates cross-talk between N- and C-terminal post-translational modifications of p53. *Cancer Biol Ther* 12: 1059-1068

Wassman CD, Baronio R, Demir O, Wallentine BD, Chen CK, Hall LV, Salehi F, Lin DW, Chung BP, Hatfield GW, Richard Chamberlin A, Luecke H, Lathrop RH, Kaiser P, Amaro RE (2013) Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nat Commun* 4: 1407

Wawrzak Z, Sandalova T, Steffens JJ, Basarab GS, Lundqvist T, Lindqvist Y, Jordan DB (1999) High-resolution structures of scytalone dehydratase-inhibitor complexes crystallized at physiological pH. *Proteins* 35: 425-439

Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* 106: 765-784

Wendy D. Cornell PC, Christopher I. Bayly , Ian R. Gould , Kenneth M. Merz , Jr., David M. Ferguson , David C. Spellmeyer , Thomas Fox , James W. Caldwell , and Peter A.

Kollman * (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* 117: 5179

White AD, Keefe AJ, Ella-Menye JR, Nowinski AK, Shao Q, Pfaendtner J, Jiang S (2013) Free energy of solvated salt bridges: a simulation and experimental study. *The journal of physical chemistry B* 117: 7254-7259

Wilcken R, Wang G, Boeckler FM, Fersht AR (2012) Kinetic mechanism of p53 oncogenic mutant aggregation and its inhibition. *Proceedings of the National Academy of Sciences of the United States of America* 109: 13584-13589

Wiman KG (2010) Pharmacological reactivation of mutant p53: from protein structure to the cancer patient. *Oncogene* 29: 4245-4252

Wissner A, Berger DM, Boschelli DH, Floyd MB, Jr., Greenberger LM, Gruber BC, Johnson BD, Mamuya N, Nilakantan R, Reich MF, Shen R, Tsou HR, Upeslakis E, Wang YF, Wu B, Ye F, Zhang N (2000) 4-Anilino-6,7-dialkoxyquinoline-3-carbonitrile inhibitors of epidermal growth factor receptor kinase and their bioisosteric relationship to the 4-anilino-6,7-dialkoxyquinazoline inhibitors. *J Med Chem* 43: 3244-3256

Wojtczak A, Cody V, Luft JR, Pangborn W (2001) Structure of rat transthyretin (rTTR) complex with thyroxine at 2.5 Å resolution: first non-biased insight into thyroxine binding reveals different hormone orientation in two binding sites. *Acta crystallographica Section D, Biological crystallography* 57: 1061-1070

Wolozin B (2012) Regulated protein aggregation: stress granules and neurodegeneration. *Molecular neurodegeneration* 7: 56

Wong KB, DeDecker BS, Freund SM, Proctor MR, Bycroft M, Fersht AR (1999) Hot-spot mutants of p53 core domain evince characteristic local structural changes. *Proceedings of the National Academy of Sciences of the United States of America* 96: 8438-8442

Wong SE, Lightstone FC (2011) Accounting for water molecules in drug design. *Expert Opin Drug Discov* 6: 65-74

Xu J, Reumers J, Couceiro JR, De Smet F, Gallardo R, Rudyak S, Cornelis A, Rozenski J, Zwolinska A, Marine JC, Lambrechts D, Suh YA, Rousseau F, Schymkowitz J (2011) Gain of function of mutant p53 by coaggregation with multiple tumor suppressors. *Nature chemical biology* 7: 285-295

Xu X, Ma Z, Wang X, Xiao ZT, Li Y, Xue ZH, Wang YH (2012) Water's potential role: Insights from studies of the p53 core domain. *Journal of structural biology* 177: 358-366

Yang DS, Zewail AH (2009) Ordered water structure at hydrophobic graphite interfaces observed by 4D, ultrafast electron crystallography. *Proceedings of the National Academy of Sciences of the United States of America* 106: 4122-4126

Young T, Abel R, Kim B, Berne BJ, Friesner RA (2007) Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proceedings of the National Academy of Sciences of the United States of America* 104: 808-813

Yu X, Blanden AR, Narayanan S, Jayakumar L, Lubin D, Augeri D, Kimball SD, Loh SN, Carpizo DR (2014) Small molecule restoration of wildtype structure and function of mutant p53 using a novel zinc-metallochaperone based mechanism. *Oncotarget* 5: 8879-8892

Zandi R, Selivanova G, Christensen CL, Gerds TA, Willumsen BM, Poulsen HS (2011) PRIMA-1Met/APR-246 induces apoptosis and tumor growth delay in small cell lung cancer expressing mutant p53. *Clin Cancer Res* 17: 2830-2841

Zhong D, Pal SK, Zewail AH (2008) Biological Water: A Critique. *Chem Phys Lett* 503: 1-11

Zhong D, Pal SK, Zewail AH (2011) Biological water: A critique. *Chemical Physics Letters* 503: 1-11

Zhou W, Quah ST, Verma CS, Liu Y, Lane DP, Brown CJ (2012) Improved eIF4E binding peptides by phage display guided design: plasticity of interacting surfaces yield collective effects. *PLoS One* 7: e47235