COMPUTATIONAL STUDIES
ON THE STRUCTURAL ASPECTS OF
PROTEIN-PEPTIDE INTERACTIONS

NGUYEN THANH BINH

NATIONAL UNIVERSITY OF

SINGAPORE

2017

COMPUTATIONAL STUDIES ON NGUYEN THANH BINH 2017

THE STRUCTURAL ASPECTS OF PROTEIN-PEPTIDE INTERACTIONS

COMPUTATIONAL STUDIES
ON THE STRUCTURAL ASPECTS OF
PROTEIN-PEPTIDE INTERACTIONS

NGUYEN THANH BINH
(Master of Science, FSU, Russia)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF BIOLOGICAL SCIENCES
NATIONAL UNIVERSITY OF SINGAPORE

2017

Supervisors:

Associate Professor M. S. Madhusudhan, Main supervisor

Dr Chandra Shekhar Verma, Co-supervisor

Examiners:

Associate Professor Kunchithapadam Swaminathan

Assistant Professor Sebastian Maurer-Stroh

Professor Ramasubbu Sankararamakrishnan, Indian

Institute of Technology, Kanpur

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information that have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Nguyen Thanh Binh

12 Jan 2017

Acknowledgments

First I would like to thank my supervisor, Dr. M.S. Madhusudhan for all these years of mentoring and support. He has taught me not only bioinformatics, but also other scientific research skills, oral presentation and scientific writing. He has also given me valuable advices to overcome the stressful time of PhD program. I would like to thank my co-supervisor, Dr. Chandra Verma for his suggestion and guidance. I would like to thank my collaborator, Dr. Kim Chu Young for his insights into the research topics.

Secondly, I would like to thank our group members, especially, Minh Nguyen, Kuan Pern Tan, Neelesh Soni, Shipra Gupta, Parichit Sharma and many others for their helps and discussions on various projects. I would like to thank Sanjana Nair, Ankit Roy and Neeladri Sen for reading and revising this thesis.

I would like to thank Chandra's group, namely: Kannan Srinivasaraghavan, Lama Dilraj, Sim Adelene, Li Jianguo, Fox Stephen John, Irani Maryam Azimadeh, Zahra Ouaray and others, for deep insight discussion on the aspect of molecular dynamics simulations.

I would like to thank my TAC members, Dr. Maurer-Stroh Sebastian, Dr. Kunchithapadam Swaminathan, Dr. Greg Tucker-Kellogg, and Dr. Anders Poulsen for their valuable suggestions.

I would like to thank BII and DBS administration officers for their help in official matter during my PhD.

I would like to thank the A*STAR Singapore for funding me during my PhD program.

Lastly, I would like to thank my family, my husband and my daughter for their love, sacrifices and endless support.

This thesis is dedicated to my beloved daughter, Ha Minh An.

Table of Contents

Declaration i

Acknowledgments ii

Table of Contents iv

Summary ix

List of Tables xi

List of Figures xii

List of Abbreviation and Notation xiv

Chapter 1 Introduction 1

1.1 Introduction to Protein-Peptide Interactions 1

1.2 Thesis Organization 2

Chapter 2 Structural Basis of HLA-DQ2.5–CLIP Complexes 4

2.1 Background and Motivations 4

2.2 Methods and Experimental Procedures 6

2.2.1 Expression and Purification 6

2.2.2 Crystallization and Data Collection 7

2.2.3 Structure Determination and Analysis 8

2.2.4 Model Building of DQ2.5 (Wild Type)–CLIP1–DM and DQ2.5 (Eβ86G, Qα31I, Hα24F)–CLIP1–DM 8

2.2.5 Molecular Dynamics Simulations 9

2.2.6 Cavity Calculation 10

2.3 X-ray Crystal Structure Analysis 10

2.3.1 Crystal Structures of DQ2.5–CLIP1 and DQ2.5–CLIP2 10

2.3.2 Reason for the CLIP2 preference over the CLIP1 in DQ2.5 based on the crystal structures 15

2.3.3 The CLIP-rich Phenotype of DQ2.5 by Structural Basis 16

2.3.4 Hydrogen Bond Interaction between the P1 Backbone Nitrogen of

CLIP1 and the α52 Backbone Oxigen of DQ2.5 22

2.4 Discussion 23

Chapter 3 The Molecular Mechanism behind the Peptide-editing Process of MHCII by DM Catalyst: an MD study 27

3.1 Introduction 27

3.2 Introduction to MD simulations 31

3.2.1 Definition and history 31

3.2.2 Newton's equation of motion 32

3.2.3 Derivative of potential energy 33

3.3 Methodology 35

3.3.1 Preparation of Starting Structures for MD Simulations. 35

3.3.2 MD Simulations Protocol 36

3.3.3 Analyzing the MD Simulations 39

3.4 Results 39

3.4.1 Model Structures 39

3.4.2 Temperature and Potential Energy 40

3.4.3 Mobility of the Whole DR–DM Complex during 1.5 μs MD Simulations 42

3.4.4 Average Atomic Mobility Reveals Important Region of DR/DM in MD Simulations 45

3.4.5 Principal Component Analysis Shows the Dynamical Correlation of the DR β2 Domain 53

3.4.6 Dynamical cross-correlation 57

3.4.7 Peptide Editing from DR by DM 58

3.4.8 Correlation of the Size of Peptide Binding Groove in Peptide-free DR in the Presence and Absence of DM 60

3.4.9 Conformational Changes of DM with Peptide-bound DR and Peptide-free DR 61

3.5 Discussion 65

3.5.1 Effect of pH on the DR-DM interaction 65

3.5.2 Mechanism of peptide editing from DR 65

3.5.3 Stabilization of peptide-free DR by DM 66

3.5.4 Conformational change upon DR-DM complex formation and important residues for the interaction. 66

Chapter 4 Prediction of Polyproline Type II Helix Receptors 69

4.1 Background and Motivation 69

4.1.1 Definition and Properties of Polyproline II Helix (PPII) 69

4.1.2 Abundance of PPII 70

4.1.3 Significance of Predicting the Polyproline type II Helix Receptors 71

4.2 Methodology 73

4.2.1 Dataset for Identifying Important Requirements to Bind PPII 73

4.2.2 Features to Characterize the PPII Binding Site 73

4.2.3 Structural Alignments 74

4.2.4 Constructing the RMSD-SVM Models. 75

4.2.5 Modeling the PPII Peptide into the Predicted Location of the Query Protein 77

4.2.6 Dataset for Searching New PPII-binding Proteins 80

4.3 Results 81

4.3.1 Features to Characterize the PPII-binding Site 81

4.3.1.4 Entropy Conservation 86

4.3.2 Prediction Accuracy 86

4.3.3 Benchmarking with Other Protein-peptide Interaction Prediction Methods 89

4.3.4 Searching possible PPII receptors in the PDB 94

4.4 Discussion 104

Chapter 5 pK_a Prediction of Ionizable Amino Acid Residues in Proteins 107

5.1 Background and Motivations 107

5.1.1 Significance of the pK_a Prediction 107

5.1.2 Review of Previous Works 109

5.2 Residue Depth.. 111

5.3 DEPTH-based pK_a Prediction 111

5.3.1 Features Constructing the DEPTH Model 114

5.3.2 Dataset of experimental values of pK_a used in DEPTH Prediction 115

5.3.3 Results and Benchmarking 143

5.4 Meta-algorithm DEMM for pK_a Prediction...................................... 146

5.4.1 Improvement in the Electrostatic Calculation. 147

5.4.2 Dataset of experimental values of pK_a used in DEMM Prediction 147

5.4.3 Results and Performance Benchmark 148

5.5 Case Study and Web-server 149

5.5.1 Case Study 149

5.5.2 Web-server 150

5.6 Chapter Summary and Discussions 151

Chapter 6 Discussions and Future Directions 159

6.1 3D structures of DQ2.5-CLIP complexes 159

6.1.1 Summary 159

6.1.2 Suggestions and Future Directions 160

6.2 The Molecular Mechanism behind the Peptide-editing Process of MHCII by DM Catalyst: an MD study 161

6.2.1 Summary 161

6.2.2 Suggestions and future directions 161

6.3 Prediction of PPII receptors 162

6.3.1 Summary 162

6.3.2 Suggestions and future directions 162

6.4 pK_a prediction of ionizable residues 164

6.4.1 Summary 164

6.4.2 Suggestion and future direction 165

6.5 Conclusion remarks 165

Publications..167

Bibliography 169

Summary

Proteins mediate important biological processes by interacting with other biomolecules, namely, other proteins, peptides, sugars, lipids or nucleic acids. This thesis presents my works on protein-peptide interactions using computational approaches. The thesis is focused on a specific peptide conformation, namely the polyproline type II helix (PPII). The protein-PPII interactions are crucial to several processes such as signaling pathways, localization, immune response, post-translational modifications *etc*.

Three different aspects of protein-PPII interactions have been studied to decipher the specificity of these interactions. First, two X-ray crystal structures of major histocompatibility complex class II (MHCII) in the complexes with two peptides having PPII conformation have been refined. Based on the crystal structures and molecular dynamics (MD) simulations, the reasons why certain peptides are retained in that MHCII has been revealed. Next, MD simulations on MHCII-peptide complexes have been performed to understand the peptide editing mechanism by a catalyst, DM. The study showed that DM can stabilize the peptide-free MHCII by the interaction not only at α1 and β1 domains, but also at β2 domain. Third, we have also analyzed the protein-PPII interaction by studying PPII receptor proteins. The analysis suggests specific features for PPII-binding. These features have been used to predict the PPII-receptor propensity of a query protein.

As electrostatics play an important role in mediating these interactions, we have studied the protonation states of ionizable residues, pK_a, in proteins. This study, while applied to protein-PPII interactions is general and applicable to all protein structures. Our pK_a prediction protocol is a simple and relatively accurate method. The accuracy is within a fraction of a pH unit.

This thesis presents the intensive studies on different aspects of protein-PPII interactions and could contribute to the knowledge of these interactions as well as protein-peptide interactions.

List of Tables

Table 2.1: Systems for MD simulations..9

Table 2.2: Data collection and refinement statistics 11

Table 2.3: Binding inhibitory capacity of CLIP1 and CLIP2 peptides to different MHCII proteins 14

Table 3.1: Systems for MD simulations..36

Table 3.2: Protonation states of the ionizable residues that differ from their canonical states 38

Table 3.3: Residence times (in %) of hydrogen bond interactions between DR protein and HA peptide..59

Table 4.1: Statistics of RMSD between the transfer PPII and the native PPII in different SVM models..89

Table 4.2: RMSD benchmarking for CABS-dock, GalaxyPepDock and our methods. 91

Table 4.3: List of the predicted PPII-receptors 95

Table 5.1: RMSD of predicted pK_a (in pH units) from experimentally determined values 113

Table 5.2: Listing of experimentally determined pK_a values of ionizable residues and their sources. 116

Table 5.3: Benchmarking of pK_a prediction using DEPTH and other methods on a testing set of 60 ionizable residues...136

Table 5.4: Physical features tested (individually) for correlation with pK_a..145

Table 5.5: Optimized coefficients of linear recombination for the different ionzable amino acid. 145

Table 5.6: Performance benchmark of DEPTH, MMSCP and DEMM over 47 ionizable groups on the testing set. 150

Table 5.7: pK_a dataset for DEMM method. 153

List of Figures

Figure 2.1: Sequence and homo-trimer forms of human invariant chain.. 5

Figure 2.2: Crystal structures of CLIP1/CLIP2 peptides bound to DQ2.5.. ... 12

Figure 2.3: 2Fo-Fc electron density maps of CLIP peptides with a contour of 1.0σ. 13

Figure 2.4: Close up of the P4 pocket in four MHCII-CLIP1 complexes.. 16

Figure 2.5: Adaptive Poisson Boltzmann Solver (APBS)-generated electrostatic surface of MHC class II proteins at pH 7.0. 18

Figure 2.6: APBS-generated electrostatics surface of MHC class II proteins at pH 5.5.. 18

Figure 2.7: Conformation of the peptide binding groove in MHCII-peptide complexes.. 19

Figure 2.8: Hydrogen bond network at the bottom of the peptide binding groove in MHCII proteins. 21

Figure 2.9: Hydrogen bond interactions of the DQ2.5–CLIP1 complex during 50 ns MD simulations.. 21

Figure 2.10: The occupancy of P1 pocket in different complexes.. 22

Figure 2.11: The C^{center}- C^{α} distance between α51F and β82N during MD simulations. 23

Figure 3.1: The apo and the holo conformations of DR/DM proteins. 30

Figure 3.2: Potential energy (kJ/mol) during 500 ns MD simulations in the six systems.. 40

Figure 3.3: Temperature (in K) during 500 ns MD simulations of the six systems.. 41

Figure 3.4: RMSD of DR protein.. 43

Figure 3.5: RMSD of DM protein. 44

Figure 3.6: The conformation of DR in the DR–CLIP1 crystal structure and in the 500 ns snapshot.. 44

Figure 3.7: RMSF of DR in the six systems during 1.5 μs simulations.. 50

Figure 3.8: RMSF of DM in the four systems during 1.5 μs simulations. 53

Figure 3.9: PC1 Porcupine plots of DR in the six systems. 54

Figure 3.10: PC1 Porcupine plots of DM in the four systems.. 55

Figure 3.11: Dynamical cross validation matrix of DR in six systems. 56

Figure 3.12: Dynamical cross validation matrix of DM in four systems.. 57

Figure 3.13: Number of hydrogen bond between DR and HA peptide during the MD simulations.. 59

Figure 3.14: The conformational change of α-helices forming peptide binding groove.. 61

Figure 3.15: Conformation of the helical region in DM. 63

Figure 3.16: (ψ, ϕ) diheral angles residues α69-75 in DM. 64

Figure 4.1: Conformation of a PPII peptide.. 70

Figure 4.2: The number of hydrogen bond interactions between PPII and the receptor for the 44 structures.. 82

Figure 4.3: The atomic and residue depth of Trp and Donor residues for 44 structures.. 84

Figure 4.4: The $N^{\varepsilon1}$-NX distance in Å for 44 structures.. 85

Figure 4.5: Entropy values of the PPII-binding Trp residues in 44 structures.. 86

Figure 4.6: New cases of the PPII-binding proteins.. 103

Figure 5.1: Complementarity between DEPTH and MMSCP in pK_a prediction. 146

Figure 5.2: A ribbon representation of the ribonuclease H (PDB ID: 2RN2). 149

List of Abbreviation and Notation

RMSD	Root Mean Square Deviation
SVM	Support Vector Machine
MHC	Major Histocompatibility Complex
PPII	PolyProline II helix
RMSF	Root Mean Square Fluctuation
PCA	Principal Component Analysis
DCCM	Dynamics Cross-Correlation Matrix
MD	Molecular Dynamics
CLIP	Class II-associated invariant chain peptide
HA	Hemagglutinin peptide from influenza A virus
SH3	Src Homology-3 domain
GYF	Glycine-tyrosine-phenylalanine
EVH1	Enabled/VASP Homology-1
MYND	Myeloid, Nervy, and DEAF-1 domain
MC	Main Chain
SC	Side Chain
SO	Structure Overlap
SASA	Solvent Accessible Surface Area
LOFO	Leave-One-Family-Out
NMR	Nuclear Magnetic Resonance
X-ray	X-ray crystallography
BLAST	Basic Local Alignment Search Tool
PDB	Protein Data Bank

Chapter 1
Introduction

1.1 Introduction to Protein-Peptide Interactions

At the cellular level living organisms use protein to perform their essential biological functions, such as signaling network, DNA repair, metabolism, gene expression, replication, transporting and folding. These functions are performed when proteins interact with other molecules, such as other proteins, peptides, sugars, lipids or nucleic acids. Among these interactions, the most abundant are protein-protein interactions, 15 to 40% of which is mediated by a stretch of a small peptide[1]. The protein-peptide interactions involve in signaling, regulatory networks, cell localization, protein degradation, and immune response. Recently, it was shown that protein-peptide interactions could be a drug target, and the peptides, in addition, could be potential drug candidates[2].

Many experimental methods could be used for identifying protein-peptide interactions at the atomic resolution. The common techniques include but are not limited to X-ray crystallography[3], nuclear magnetic resonance (NMR) spectroscopy[4], alanine scanning mutagenesis[5] and mass spectrometry[6] approaches. These techniques are valuable and have contributed to the knowledge of protein-peptide interactions. However, these techniques have many drawbacks like difficulties in expression and purification of large proteins, obtaining high resolution X-ray structures or restriction of the protein size in the NMR method. Another limitation is that these techniques are time-consuming and labor intensive. Over the last 30 years, computational approaches for identifying the protein-peptide interactions have been developed. However, those approaches are still in the infant stage, despite the availability of more than 100, 000 protein structures in the protein data bank (PDB). The reason for this drawback could be due to insufficiency of protein-peptide interaction types, or the lack of knowledge in the features that contribute to the protein-peptide interactions. And hence, there are still open questions about protein-peptide interactions; such as: Are there common principles for peptide binding in different cellular functions? What factors help to stabilize protein-peptide interactions? Are such factors common in certain structural or functional families? Are there special conformations on the protein for recognizing the binding peptide? Is it possible to predict and/or design peptides that would have high affinity to the binding pocket of a particular protein? To address these questions, it is necessary to first categorize and gain insight from structural data on existing protein-peptide complexes. As more than one third of the protein-binding peptides have extended beta or polyproline II (PPII) helical conformation[7], this thesis focuses only on protein-PPII interactions.

1.2 Thesis Organization

In this thesis, the following issues have been tackled:

(1) In-depth learning on an example of the protein-peptide complex involved in the immune pathway, where the peptide has PPII conformation. This is a collaborative project, and we have worked with an experimental lab to solve

the complex structures with the follow up of analyzing the structures using molecular dynamics (MD) simulations;

(2) Conformational study on the PPII-peptide editing process of MHCII by a catalyst. This is an MD simulation study on different complex systems to understand this particular protein-PPII interaction.

(3) Predicting PPII-binding propensity of a protein. The target protein was aligned with a template of two residues from known PPII-binding proteins. All the possible alignments were then classified as the binding or nonbinding positions by support vector machine (SVM). The PPII peptide from the template structure was then used to build into the target structure using Monte Carlo simulations. We also applied this model to find the new PPII receptors in a non-redundant dataset of 30% sequence identity from PDB;

(4) As charges of ionizable residues play a critical role in protein functions and protein-peptide interactions, we have exploited the prediction protocol for identifying the charge or protonation state of these ionizable residues;

Details of all the four chapters listed above are in the following sections.

Chapter 2
Structural Basis of
HLA-DQ2.5–CLIP Complexes

In this chapter, an example of protein-peptide interaction, particularly DQ2.5-CLIP complex, was studied. This complex structure correlates with a particular disease. Understanding the interaction in this complex could give insight into the disease mechanism.

2.1 Background and Motivations

Antigenic peptides are presented to T cell receptors of CD4 T cells[8] by MHCII proteins. These proteins have a peptide binding groove formed by one α chain and one β chain. Only three MHCII isotypes, namely DR, DP and DQ, are found in human. All these isotypes are encoded on chromosome 6. MHCII proteins are synthesized in the endoplasmic reticulum in a nanomeric complex with a chaperone protein called the invariant chain (Ii) or $\alpha_3\beta_3Ii_3$[9]. By the formation of this complex nascent MHCII is prevented from interacting with indiscriminate peptides and the MHCII-Ii complex is targeted to the endosome where MHC performs its function[10]. Once in the endosome, the invariant chain from MHCII-Ii complex is progressively proteolyzed until only a short fragment called class-II-associated invariant chain peptide (CLIP) remains in the peptide binding groove of the MHCII[11] (αβ). Subsequently, with the help of a catalyst, DM, CLIP is released and replaced by exogenous peptides. The MHCII-exogenous peptide complex is then transported to the cell surface and presented to $CD4^+$ T cell[12]. DM acts as a catalyst to edit either CLIP or low binding affinity peptides[12-14]. Currently, three regions in Ii that could bind

MHCII were detected. They are the canonical CLIP1 (residues 83-101), non-canonical CLIP2 (residues 92-107), and non-canonical CLIP3 (residues 98-111) (Figure 2.1). Among the three regions, CLIP1 is exclusively observed in most mouse and human MHCII proteins[12]. So far the only proteins, which are shown to bind both CLIP1 and CLIP2 peptides, are DQ2.2, DQ2.5, DQ7.5, and DQ8[15-17]. DQ7.5 also binds CLIP3[17]. Interestingly, those human MHCII alleles binding to CLIP2 and CLIP3 are associated with one or more autoimmune diseases; particularly celiac disease (DQ2.2, DQ2.5, DQ7.5 and DQ8)[18-20] and type 1 diabetes (DQ2.5 and DQ8)[19,20].

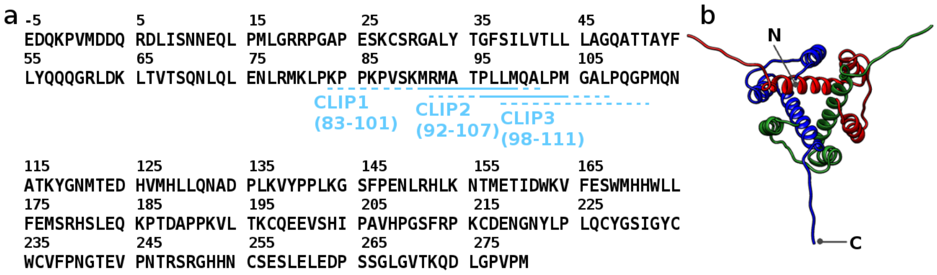

a
-5 EDQKPVMDDQ 5 RDLISNNEQL 15 PMLGRRPGAP 25 ESKCSRGALY 35 TGFSILVTLL 45 LAGQATTAYF
55 LYQQQGRLDK 65 LTVTSQNLQL 75 ENLRMKLPKP 85 PKPVSKMRMA 95 TPLLMQALPM 105 GALPQGPMQN
CLIP1 (83-101)
CLIP2 (92-107)
CLIP3 (98-111)
115 ATKYGNMTED 125 HVMHLLQNAD 135 PLKVYPPLKG 145 SFPENLRHLK 155 NTMETIDWKV 165 FESWMHHWLL
175 FEMSRHSLEQ 185 KPTDAPPKVL 195 TKCQEEVSHI 205 PAVHPGSFRP 215 KCDENGNYLP 225 LQCYGSIGYC
235 WCVFPNGTEV 245 PNTRSRGHHN 255 CSESLELEDP 265 SSGLGVTKQD 275 LGPVPM
b
N
C

Figure 2.1: Sequence and homo-trimer forms of human invariant chain. (a) Amino acid sequence. MHC binding core sequence of CLIP1 is MRMATPLLM and that of CLIP2 is PLLMQALPM. MHC binding core sequence of CLIP3 is unknown. (b) Solution NMR structure of the truncated human invariant chain protein (residues 118-192, PDB ID: 1IIE). The invariant chain exists as a homo-trimer and associates with three major histocompatibility complex proteins simultaneously in the endoplasmic reticulum.

DQ2.5 allele is associated with an autoimmune-like disorder, celiac disease, caused by a harmful immune response when wheat gluten and similar proteins from rye and barley[21] are ingested. About 95% of celiac disease patients express DQ2.5. This allele is encoded by the DQA1*05:01 and DQB1*02:01 genes of the DR3–DQ2 haplotype[18]. The gluten-specific CD4$^+$ T cells of celiac disease patients recognize a various set of gluten epitopes when they are presented in the complex with DQ2.5 but not in the complex with other

MHCII molecules[22-25]. DQ2.5 has unusually high CLIP phenotype. Up to 53% of exogenous displayed peptides[15-17,26] in DQ2.5 expressing B lymphoblastoid cells are CLIP peptides, either CLIP1 or CLIP2. While in general, only 10% of displayed peptides in other MHCII are CLIP[27]. Moreover, in DQ2.5 the amount of the non-canonical CLIP2 peptide is higher than that amount of the canonical CLIP1[15,16]. The CLIP-rich phenotype in DQ2.5 was explained by the poor interaction between DQ2.5 and DM[16,28]. The structural explanation for the unusual CLIP amount is not clear at the atomic level. Here, we have determined the crystal structures of DQ2.5–CLIP1 and DQ2.5–CLIP2 to have insight into the DQ2.5–CLIP interaction.

2.2 Methods and Experimental Procedures

2.2.1 Expression and Purification

The preparation of DQ2.5 containing covalently linked CLIP1 and CLIP2 is similar to DQ2.5–αI gliadin[22,29-31]. The Fos and Jun leucine zippers were attached to the C-termini of the α- and β-chains, respectively, through an intervening Factor Xa site to promote heterodimer stability[31,32]. A 15-residue linker was used to attach the CLIP1 (PVSKMRMATPLLMQA) and CLIP2 (MATPLLMQALPMGAL) peptides to the N terminus of the β-chain. A baculovirus expression system was used to coexpress the α- and β-chains in ExpresSF+ insect cells. The DQ2.5 heterodimer was purified using mAb 2.12. E11[29], concentrated, and washed using a size exclusion filter. The MD simulations were also applied in several MHCII–CLIP and MHCII–CLIP–DM systems to understand the mechanism of how CLIP is retained in DQ2.5 and how DM is less susceptible to DQ2.5.

2.2.2 Crystallization and Data Collection

Factor Xa was used to remove the leucine zippers from the DQ2.5–CLIP1 and DQ2.5–CLIP2 complexes for 16 hours at 24ºC. Purification of both complexes was conducted using anion exchange (buffer A: 25 mM Tris, pH 8.0, buffer B: 25 mM Tris, pH 8.0, 0.5 M NaCl) and size exclusion chromatography (buffer: 25 mM Tris, pH 8.0). The solution was then concentrated to 2 mg/ml. Both complexes were crystallized by combining 1 µl of the protein solution and 1 µl of respective precipitant buffer in a single hanging drop at 18 ºC. The buffer used for DQ2.5–CLIP1 was 0.1 M ammonium sulphate, 0.1 M sodium cacodylate, pH 6.5, 25% PEG 8000 and 6% glycerol, while for DQ2.5–CLIP2, the buffer was 0.1 M BIS-TRIS, pH 5.5, 22% PEG 3350. Small crystals of both DQ2.5–CLIP1 and DQ2.5–CLIP2 complexes appeared within one week and then grew to its full size in two weeks. Crystals were soaked in 5% glycerol + mother liquor and later flash frozen in liquid nitrogen. X-ray diffraction data were collected at the Stanford Synchrotron Radiation Laboratory at the beam line 9-3. HKL2000 program was used to index and integrate the diffraction data[33]. DQ2.5–CLIP1 crystal has the C121 space group. Its cell dimensions are a=128.86 Å, b=69.21 Å, c=146.69 Å and β = 110.3º. DQ2.5–CLIP2 crystallized in the I23 space group with cell dimensions a=b=c=137.01 Å.

2.2.3 Structure Determination and Analysis

Molecular replacement using Phaser[34,35] was used to determine both structures. The search model was DQ2.5–gliadin structure (PDB ID: 1S9V). Model refinement was carried out using Refmac[36], Phenix[37], and Coot[35]. Both CLIP1 and CLIP2 peptides were built at the end of the refinement process, using the F_o-F_c electron density map at 3.0 σ. Throughout the refinement isotropic B correction and bulk solvent correction were applied. Water molecules were identified in the $2F_o$–F_c map from electron density map greater than 1.0 σ. All the water molecules were checked for environment, valid geometry, and density shape before conducting additional model building and refinement cycles. The last two refinement rounds included TLS (translation, libration, and screw–rotation displacements) parameterization. PROCHECK[38] was used for checking the stereochemical quality of the final structures.

2.2.4 Model Building of DQ2.5 (Wild Type)–CLIP1–DM and DQ2.5 (Eβ86G, Qα31I, Hα24F)–CLIP1–DM

Both the wild type and mutant DQ2.5–CLIP1–DM complexes were modeled using the MODELLER program version 9.10[39,40]. The templates were the crystal structure of DR1–HA–DM (PDB code: 4FQX) and DQ2.5–CLIP1 (PDB code: 5KSU). In this model, CLIP1 was truncated to the same length (from P2 to P10) as the HA peptide in the DR1–HA–DM crystal structure. Total of five models evaluated by the DOPE statistical energy function[41] were created. Energy minimized models were achieved by slow refine option. All structural figures were generated using Chimera[42] and Pymol[43].

2.2.5 Molecular Dynamics Simulations

Introduction to MD simulations is on chapter 3.2. All-atom MD simulations were carried out on three systems, including, DQ2.5(wild type)–CLIP1, DQ2.5(wild type)–CLIP1–DM and DQ2.5(Eβ86G, Qα31I, Hα24F)–CLIP1–DM. All crystal water molecules were included in the starting MD structures because they are important for mediating the protein–peptide interactions[44-46]. The protonation states of all ionizable residues, including ASP, GLU, HIS, LYS and ARG were assigned according to the model pK_a at pH equal to 7. TIP3P water box with the minimum distance of 12 Å to any protein atom was used to solvate each system (Table 2.1). Sodium counter ions were used to neutralize the systems. Periodic boundary conditions were applied on the system.

Table 2.1: Systems for MD simulations

System	Peptide start (peptide length)	Net charge	Number of solvent waters
DQ2.5–CLIP1	P-4(14)	-6	21437
DQ–CLIP–DM	P2 (9)	-24	30139
DQ (mutant)–CLIP–DM	P2 (9)	-24	29243

First energy minimization using the steepest descent and then the conjugate gradient methods was applied to the complex. The system was heated to 300K within 800 ps under the NVT conditions. The system was then equilibrated for 1 ns under the NPT conditions. Later, under the NVE conditions the triplicate MD simulations were carried out for 50 ns.

SHAKE was applied for all bonds involving hydrogen. All the simulations were carried out using the ff99SB[47] force fields in the AMBER12 program[48]. The long-range interactions were calculated by the Particle Mesh Ewald (PME)[49] algorithm while the cutoff of 10.0 Å was applied for the short-range interactions. The integration time step was set to 1 fs. The analysis on MD trajectories was carried out using a combination of indigenously developed Python scripts and the Ptraj/Cpptraj module of Amber12.

2.2.6 Cavity Calculation

The Voronoi algorithm[50] was applied to calculate the cavity of the P4 pocket in MHCII.

2.3 X-ray Crystal Structure Analysis

2.3.1 Crystal Structures of DQ2.5–CLIP1 and DQ2.5–CLIP2

The crystal structures of DQ2.5–CLIP1 and DQ2.5–CLIP2 complexes were solved to 2.73 Å and 2.20 Å resolutions, respectively (Figure 2.2). Both structures do not have density for the β105-112 loop. Side chain atoms of α75K, α158E, α172K, β22E and β135D residues in the DQ2.5–CLIP1 structure could not be placed. Data collection and refinement statistics are presented in Table 2.2. The DQ2.5 conformation in the DQ2.5–CLIP1 structure is similar to that conformation in the DQ2.5–CLIP2, the DQ2.5–gliadin-α1a (PDB code 1S9V)[22] and the DQ2.5–gliadin-α2 (PDB code 4OZF, 4OZG, and 4OZH)[23] structures (C^{α} RMSD of 360 atoms ranging from 0.57 to 1.27 Å). The CLIP1 and CLIP2 peptides in the current structures have highly similar main chain (C^{α} RMSD of 0.47 Å) and side chain conformations (C^{β}

RMSD of 0.85 Å).

Table 2.2: Data collection and refinement statistics

Complex name	DQ2.5–CLIP 1	DQ2.5–CLIP 2
PDB code	5KSU	5KSV
Data collection		
Space group	C121	I23
Cell dimension		
a, b, c (Å)	128.86, 69.21, 146.69	137.01, 137.01, 137.01
α, β, γ (°)	90, 110.3, 90	90, 90, 90
Resolution (Å)	2.73 (2.80-2.73)	2.20 (2.30-2.20)
R_{merge} (%)	10.0	12.9
I/σI	11.7	12.7
Completeness (%)	93.7 (89.2)	99.7 (99.9)
Redundancy	3.5	6.5
Refinement		
Resolution (Å)	39.26-2.73	36.62-2.20
	(2.80-2.73)	(2.30-2.20)
Number of reflections	29676	21938
R_{work} / R_{free}	0.187/0.247	0.171/0.208
	(0.29-0.37)	(0.231-0.296)
Number of atoms	6144	3176
Protein	6027	3003
Water	117	173
B-factors (Å^2)	45.0	28.1
Protein	45.1	28.1
Water	35.9	29.0
r.m.s deviations		

Bond length (Å)	0.01	0.01
Bond angle (°)	1.24	1.12
Ramachandran favored	96.3	98.1

Values for highest resolution shell are in parentheses.

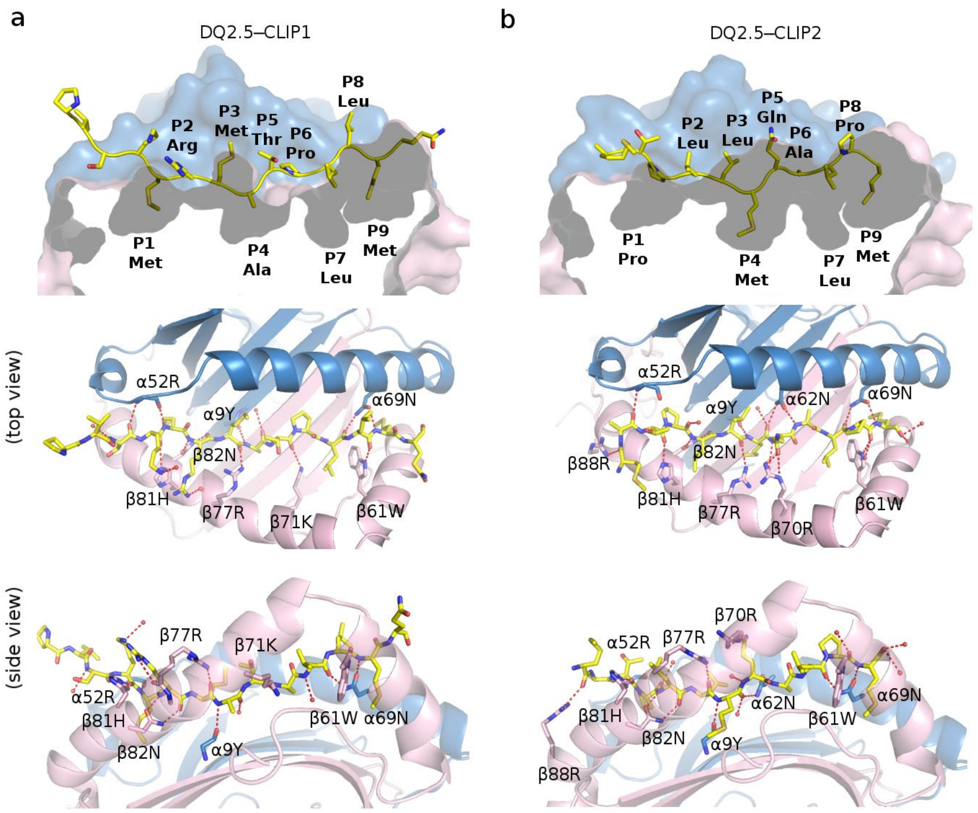


Figure 2.2: Crystal structures of CLIP1/CLIP2 peptides bound to DQ2.5. (a) Crystal structure of DQ2.5–CLIP1 (PDB ID: 5KSU). (b) Crystal structure of DQ2.5–CLIP2 (PDB ID: 5KSV). DQ2.5 α- and β-chains are in blue and pink, respectively. CLIP1 and CLIP2 peptides are shown in stick representation (light yellow, carbon; dark yellow, sulfur; blue, nitrogen; red, oxygen). Hydrogen bond interactions are represented in red dotted lines.

In the DQ2.5–CLIP1 structure, 14 residues of CLIP1 are clearly visible in the electron density map (Figure 2.3a). Residues MRMATPLLM in CLIP1 peptide (Ii 91-99) occupy the P1–P9 pockets of DQ2.5. This binding register is seen in all MHCII–CLIP1 crystal structures solved to date: DR1–CLIP1 (PDB code 3PDO), DR3–CLIP1 (PDB code 1A6A), and I-A^b–CLIP1 (PDB code

1MUJ)[51-53]. In the DQ2.5–CLIP2 structure, 12 residues of CLIP2 are clearly visible in the electron density map (Figure 2.3b). Residues PLLMQALPM in CLIP2 peptide (Ii 96-104) occupy the P1–P9 pockets of DQ2.5. This occupancy is in agreement with the binding register of CLIP2 by biochemically determination[15,16].

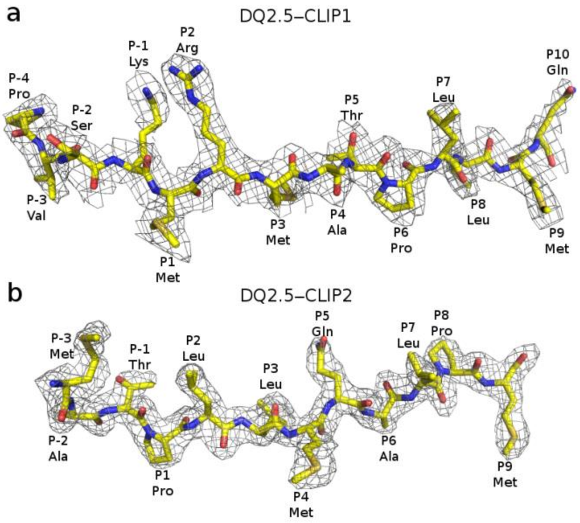


Figure 2.3: 2Fo-Fc electron density maps of CLIP peptides with a contour of 1.0σ. (a) CLIP1 and (b) CLIP2 peptides. CLIP1 and CLIP2 peptides are shown in stick representation (light yellow, carbon; dark yellow, sulfur; blue, nitrogen; red, oxygen).

CLIP1 has 12 direct and four water-mediated hydrogen bonds with DQ2.5, while CLIP2 has 14 direct and six water-mediated hydrogen bonds with DQ2.5 (Figure 2.2). There are two hydrogen bond interactions which are present in DQ2.5–CLIP1 but absent in DQ2.5–CLIP2; particularly ($N_{P1} - O_{\alpha 52N}$, and $N_{P2R}^{\eta 1} - O_{\beta 77R}$). The first interaction is not possible in DQ2.5–CLIP2 because the P1 residue in CLIP2 is a Pro, where the backbone nitrogen is in cyclic conformation and lacks the ability to make hydrogen bond interactions with backbone oxygen. There are three hydrogen bond interactions that are present in DQ2.5–CLIP2 but missing in DQ2.5–CLIP1 ($O_{P-3} - N_{\beta 88R}^{\eta 2}$, $O_{P5}^{\varepsilon 1} - N_{\beta 70R}^{\eta 2}$, $N_{P6} - O_{\beta 62N}^{\delta 1}$). Equivalent interactions are not possible in DQ2.5–CLIP1 because the main chain carbonyl C=O group of CLIP1 P-3 is rotated away from DQ2.5 β88, and because the P5 and P6 residues of CLIP1 are different from those of CLIP2. Overall, the DQ2.5 binding energy for CLIP1 and CLIP2 appear to be similar. This similarity was indicated by the experimentally measured dissociation time for DQ2.5–CLIP1 (140 hours) and DQ2.5–CLIP2 (140 hours) in the absence of DM[16].

Table 2.3: Binding inhibitory capacity of CLIP1 and CLIP2 peptides to different MHCII proteins

MHCII	***IC$_{50}$ (nM)**		**Reference**
	**CLIP1	CLIP2	
Mouse			Sette *et al.* Journal of Experimental Medicine. 181, 677-683 (1995)[12]
IA^b	32	Not tested	
IA^d	7.5		
IA^k	16666		
IA^g	49		
IE^d	682		
IE^k	364		
Human			
DR1	0.89		
DR2w2a	31		
DR3	118		
DR4w4	141		
DR4w14	12		
DR5	441		
DR7	40		
DR52a	16786		
DRw53	8.2		
DQ2.5	82500	6020	Vartdal *et al.* Eur. Journal of Immunology 26, 2764-2772 (1996)[26]

* Sette *et al.* and Vartdal *et al.* used different indicator peptides for IC$_{50}$ measurement.

** Human CLIP1 used by Sette *et al.* is Ii 80-103 (sequence: LPKPPKPVSKMRMATPLLMGALPM) and human CLIP1 used by Vartdal *et al.* is Ii 83-101. Mouse CLIP1 Is Ii 85-101 (sequence: KPVSQMRMATPLLMKPM). The core binding region (Ii 91-99) is underlined.

2.3.2 Reason for the CLIP2 preference over the CLIP1 in DQ2.5 based on the crystal structures

Four MHCII–CLIP1 crystal structures have been reported to date: DQ2.5–CLIP1 (PDB ID: 5KSU), DR1–CLIP1 (PDB ID: 3PDO)[52], DR3–CLIP1 (PDB ID: 1A6A)[51], and I-A^b–CLIP1 (PDB ID: 1MUJ)[53]. Among these four MHCII proteins, only DQ2.5 has been observed to bind CLIP2[15,16]. This specific binding to CLIP2 in DQ2.5 could be explained by its two structural features. Firstly, the P4 pocket in DQ2.5 is deeper and broader than that of DR1, DR3 and I-A^b because of the polymorphism at β13, β26 and β78 residues (Figure 2.4). Particularly, cavity size of P4 pocket in DQ2.5 is 566Å^3, that cavity from other three MHCII proteins ranges from 364 to 417 Å^3. In DQ2.5 the P4 pocket residues are β13G, β26L, and β78V. In DR1 they are β13F, β26L, and β78Y. In DR3 they are β13S, β26Y, and β78Y. In I-A^b they are β13G, β26Y, and β78V. Therefore, CLIP1, which has Ala at P4, binds to all four MHCII proteins whereas CLIP2, which has Met at P4, only binds to DQ2.5. Secondly, the DQ2.5 peptide binding groove has a positively charged (due to β70R, β71K and β77R), whereas in DR1, DR3, and I-A^b the grooves have a negative charge (due to β57D, β66D, α55E in DR1/DR3 and β57D, β66E, α55D in I-A^b) (Figure 2.5). CLIP1 peptide is positively charged (due to P-1 Lys and P2 Arg) while CLIP2 does not contain any charged amino acid residues. The long-range electrostatic interactions are crucial in the initial formation of the protein–protein complexes[54-58], and hence DQ2.5 is expected to interact more favorably with CLIP2 than with CLIP1. Indeed, it was shown by previous biochemical studies that CLIP2 binds to DQ2.5 with higher affinity than CLIP1 (IC_{50} of 6.0 μM vs. 82.5 μM)[26,59] (Table 2.3). Among all available three-dimensional structure of MHCII proteins, DQ8 is the only other MHCII which interacts with CLIP2 and, consistently, DQ8 has a large P4 pocket and a positively charged peptide binding groove similar to DQ2.5 (Figure 2.5).

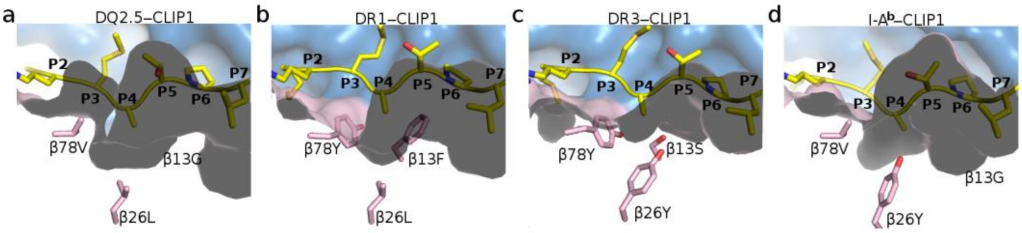


Figure 2.4: Close up of the P4 pocket in four MHCII-CLIP1 complexes. (a) DQ2.5–CLIP1 (PDB ID: 5KSU), (b) DR1–CLIP1 (PDB ID: 3PDO), (c) DR3–CLIP1 (PDB ID: 1A6A), and (d) I-A^b–CLIP1 (PDB ID: 1MUJ). The MHCII surface for α-chain and β-chain are shown in blue and pink, respectively. β-chain residues that line the P4 pocket are shown in stick representation. CLIP1 peptide is shown in yellow.

2.3.3 The CLIP-rich Phenotype of DQ2.5 by Structural Basis

While DQ2.5 expressing cells have an unusually high CLIP phenotype (up to 53%; CLIP1 and CLIP2 combined)[15-17,26], peptide content of other MHCIIs typically has only around 10% of CLIP peptide[20]. One possible explanation is that DQ2.5 binds CLIP with higher affinity compared to other MHCIIs. However, the binding affinity IC_{50} values in DR–CLIP1 and DQ2.5–CLIP1 are in nM[12] and µM[26] respectively (Table 2). These data do not support this notion. In addition, number of direct hydrogen bonds formed between CLIP1 (P-1 to P9 only) and DQ2.5, DR1, DR3, and I-A^b are 11, 13, 17 and 13 respectively. And hence, we propose that the CLIP-rich phenotype in DQ2.5 is due to an impaired interaction between DQ2.5 and the catalyst DM, whose function is to exchange CLIP peptide to higher binding affinity peptides. Much of the current structural and mechanistic understanding of MHCII–DM interaction is derived from the DR1–HA–DM crystal structure (PDB ID:

4FQX)[60]. Therefore, we investigated whether DQ2.5 has all the structural elements that facilitate the DM interaction as observing in the DR1–DM structure. To do this, a homology model of DQ2.5–CLIP1–DM was built (Figure 2.6). First, the electrostatic complementary of the contact surface areas between DQ2.5 and DM was examined. According to our model, DQ2.5 has two regions making direct contact with DM. The first region is located adjacent to the P1 pocket in the α_1 domain and the second region is located near the transmembrane segment in the β_2 domain. The surface charge distribution of these contact regions in DQ2.5 has more electrostatic complementarity to the corresponding surfaces of DM than DR. Therefore, the surface electrostatic charge distribution could be ruled out as the source of impaired DQ2.5–DM interaction.

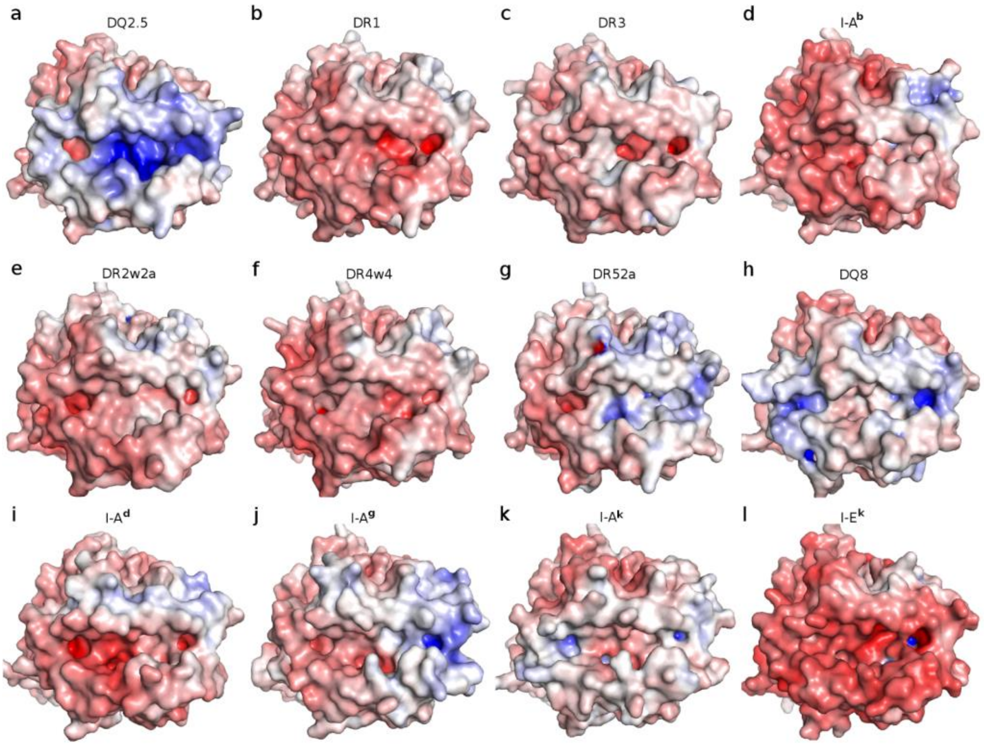


Figure 2.5: Adaptive Poisson Boltzmann Solver (APBS)-generated electrostatic surface of MHC class II proteins at pH 7.0. The negative, positive, and neutral electrostatics are in red, blue and white, respectively. The view is the top view of the peptide binding groove. (a) DQ2.5–CLIP1 (PDB ID: 5KSU), (b) DR1–CLIP1 (PDB ID: 3PDO), (c) DR3–CLIP1 (PDB ID: 1A6A), (d) I-A^b–CLIP1 (PDB ID: 1MUJ), (e) DR2w2a–Epstein Barr Virus DNA polymerase peptide (PDB ID: 1H15), (f) DR4w4–human collagen II peptide (PDB ID: 2SEB), (g) DR52a–integrin beta 3 peptide (PDB ID: 2Q6W), (h) DQ8–deamidated gluten peptide (PDB ID: 2NNA), (i) I-A^d–influenza hemagglutinin peptide (PDB ID: 2IAD), (j) I-A^g–hel 11-27 peptide (PDB ID: 3MBE), (k) I-A^k–conalbumin peptide (PDB ID: 1D9K), and (l) I-E^k–MCC peptide (PDB ID: 3QIU). The peptides bound to MHCII were omitted in the APBS electrostatics calculations.

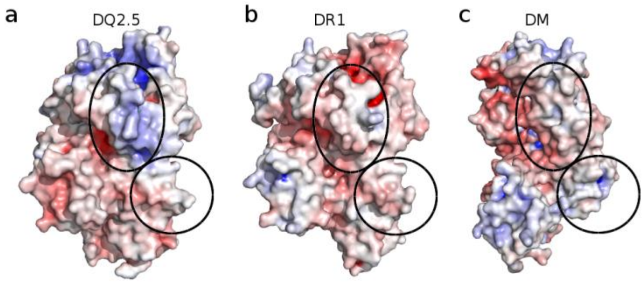


Figure 2.6: APBS-generated electrostatics surface of MHC class II proteins at pH 5.5. The negative, positive, and neutral electrostatics are shown in red, blue and white, respectively. (a) DQ2.5 in DQ2.5–CLIP1 (PDB ID: 5KSU), (b) DR1 in DR1–HA–DM (PDB ID: 4FQX) and (c) DM in DR1–HA–DM (PDB ID: 4FQX). Peptide has been removed to enhance clarity.

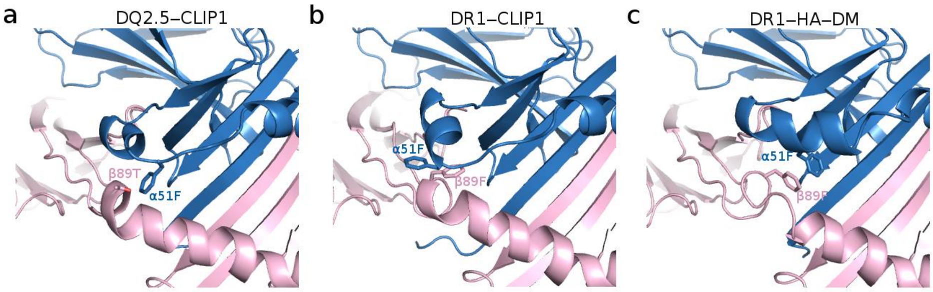


Figure 2.7: Conformation of the peptide binding groove in MHCII-peptide complexes. (a) DQ2.5–CLIP1 (PDB ID: 5KSU), (b) DR1–CLIP1 (PDB ID: 3PDO) and (c) DR1–HA–DM (PDB ID: 4FQX). The MHCII α- and β-chains are in blue and pink, respectively. Peptide bound to the MHCII is not shown.

Next, we examined whether DQ2.5 is able to undergo the same set of conformational changes that DR1 undergoes upon DM binding. The α51F in DR1 has been identified as a key DM binding residue[61,62]. When binding to DM, the α51-55 loop of DR1 transforms into an α-helix, this transformation results in the 13Å movement of α51F side chain from its initial solvent exposed position to the P1 pocket cavity. In the new location α51F forms a hydrophobic cluster with α24F, α31I, α32F, α48F, and β89F residues (Figure 2.7)[60]. This hydrophobic interaction is thought to stabilize the P1 pocket when P1 residue of the peptide is removed. In comparison with DR, DQ2.5 has a deletion mutation at α53. The insertion of Gly at this position results in partial restoring DM sensitivity in DQ2.5[63]. The deletion at α53 in DQ2.5 results in the α51F inaccessibility by DM. The relative conformational change in α51F may compromise the DQ2.5–DM interaction. Further, we suggest that the DM insensitivity of DQ2.5 is due to the presence of an extensive hydrogen bond network (involving α9Y, α22Y, α24H, α31Q, β86E, β90T, and a buried water molecule). This hydrogen bond network spans from the P1 to the P4 pockets of DQ2.5 (Figure 2.8). To assess the stability of this hydrogen bond network, a 50 ns MD simulations of the DQ2.5–CLIP1 complex was carried out. MD trajectories show that all hydrogen bonds in this network, with the exception of the peripheral β86E $O^{\varepsilon 1}$–β90T $O^{\gamma 1}$, are stable (Figure 2.9). Furthermore, during the course of the DQ2.5–CLIP1–DM MD simulations we observed that α51F does not enter the P1 pocket likely due to the presence of the α9–α22–α24–α31–β86–β90 hydrogen bond network. In particular, the P1 pocket is directly blocked by a water molecule mediated hydrogen bond interaction between α24H and α31Q (Figure 2.10). To examine the effect of the extended hydrogen bond network, we mutated the α24, α31, and β86 residues in our DQ2.5–CLIP1–DM model to their counterparts in DR1; particularly Eβ86G, Qα31I and Hα24F. Due to the ability to disrupt the α9–α22–α24–α31–β86–β90 hydrogen bond network, these mutations are expected to restore DM sensitivity in DQ2.5. Our 50 ns MD simulations of the triple mutant DQ2.5 shows that α51F does indeed occupy the P1 pocket, as seen in the DR1–HA–DM crystal structure (Figure 2.11). In DR1, DM binding also causes a change from an α-helix to a loop at the β85–90 region. This conformational change causes a 4.7 Å movement of β89F from the protein surface to the hydrophobic cluster of P1 pocket floor including α51F[60] (Figure 2.7). The similar rearrangement of the β85–90 region in DQ2.5 does not appear feasible because β86E and β90T residues are held in place by the extended α9–α22–α24–α31–β86–β90 hydrogen bond network (Figure 2.8a). In summary, the α9–α22–α24–α31–β86–β90 hydrogen bond network in DQ2.5 prevents repositioning of α51F and β89F, which is important for DR1–DM interactions.

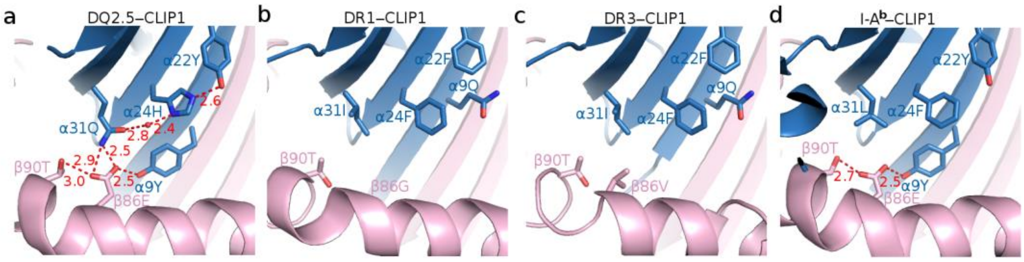


Figure 2.8: Hydrogen bond network at the bottom of the peptide binding groove in MHCII proteins. (a) DQ2.5–CLIP1 (PDB ID: 5KSU), (b) DR1–CLIP1 (PDB ID: 3PDO), (c) DR3–CLIP1 (PDB ID: 1A6A), and (d) I-A^b–CLIP1 (PDB ID: 1MUJ). MHCII α- and β-chains are shown in blue and pink, respectively. MHC bound peptides are represented in a stick model (light yellow, carbon; dark yellow, sulfur; blue, nitrogen; red, oxygen). Hydrogen bonds are shown as red dotted lines and their distances are given in Å. Water molecule is shown as a red sphere.

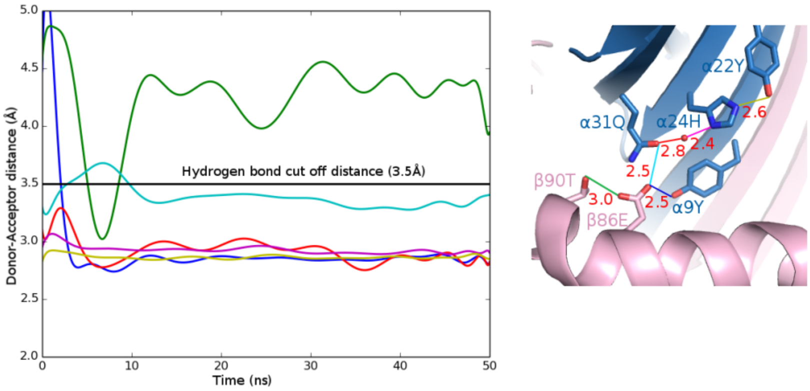


Figure 2.9: Hydrogen bond interactions of the DQ2.5–CLIP1 complex during 50 ns MD simulations. (Left) Distance trajectories of several atom pairs in the α22–α24–α31–β86–β90–α9 hydrogen bond network are shown; O$^{\varepsilon2}$ of β86E and O$^{\gamma1}$ of β90T (green), O$^{\varepsilon1}$ of β86E and O$^{\varepsilon1}$ of α31Q (cyan), O$^{\varepsilon1}$ of β86E and O$^{\eta}$ of α9Y (blue), O$^{\varepsilon1}$ of α31Q and water (red), N$^{\delta1}$ of α24H and water (violet), and N$^{\varepsilon2}$ of α24H and O$^{\eta}$ of α22Y (yellow). The threshold at 3.5Å for hydrogen bond distance is in black horizontal line. (Right) The color codes and distances (in Å) for hydrogen bond interactions from the crystal structure are indicated.

2.3.4 Hydrogen Bond Interaction between the P1 Backbone Nitrogen of CLIP1 and the α52 Backbone Oxygen of DQ2.5

All crystal structures of MHCII–peptide complexes have a hydrogen bond between the amide nitrogen of the P1 residue and the main chain carbonyl group of MHCII α53 if the P1 residue of the peptide is not Pro. Interestingly, DQ2.5 has a deletion at α53, and is able to bind gluten peptides that frequently have Pro at P1. That is why it was suggested that DQ2.5 is unable to form backbone α53-P1 hydrogen bond[63]. The significance of this peptide main chain hydrogen bond has been assessed by comparing binding of peptides being N-methylated at the P1 position with unmodified peptides. Such substitution gave decreased affinity for peptide binding to DR1 but no effect was seen for DQ2.5[64,65]. Interestingly, our DQ2.5–CLIP1 structure shows that there is indeed a hydrogen bond between the P1 main chain nitrogen of CLIP1 and the α52 carbonyl group of DQ2.5.

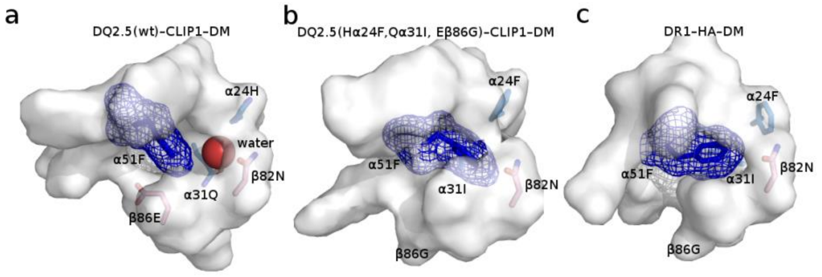


Figure 2.10: The occupancy of P1 pocket in different complexes. (a) DQ2.5–CLIP1–DM, (b) mutant DQ2.5(Eβ86G, Qα31I, Hα24F)–CLIP1–DM), and (c) DR1–HA–DM (PDB ID: 4FQX). The residues α24F, α31I, α32F, α48F, α49G, α50R, α52A, α53S, α54F, α55E, β82N, β85V, β86G and β89F that form a hydrophobic pocket in DR and equivalent residues in DQ2.5 are shown in surface representation, colored white. The residue α51F is shown in blue slid surface with side chain in stick representation. The water molecule is shown in red sphere in panel a. The P1 pockets in different complexes are shown in the same orientation.

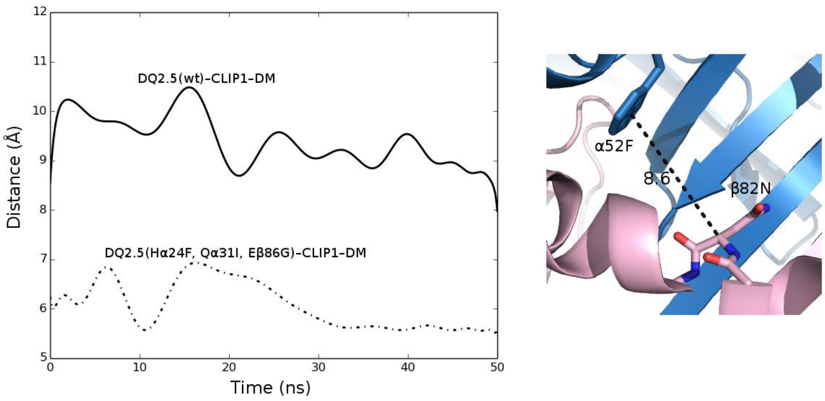


Figure 2.11: The C^{center}- C^{α} distance between α51F and β82N during MD simulations. (left) The distance trajectory of the center of the phenyl group in α51F to the C^{α} of β82N in DQ2.5 in the wild-type (solid line) and mutant (dashdot line) DQ2.5–CLIP1–DM complexes. (right) The corresponding distances in the DQ2.5–CLIP1 crystal structure.

2.4 Discussion

In this study, the crystal structures of DQ2.5–CLIP1 and DQ2.5–CLIP2 complexes have been determined at 2.73 Å and 2.20 Å resolutions, respectively. Although there are several available crystal structures of MHCII proteins with both canonical (the peptide orientation from N- to C-terminal in the peptide binding groove from P1 to P9) (PDB code 3PGD, 3PDO and 4AH2) and flipped (the peptide orientation is inverted from C to N-terminal in the peptide binding groove from P1 to P9) orientations (PDB code 3PGC and 4AEN)[52,66] of CLIP1 peptide, this is the first time a crystal structure of a MHCII–CLIP2 complex has been reported. DQ2.5 is unusual in the fact that it associates with the canonical CLIP1 (Ii 83-101) as well as the non-canonical CLIP2 (Ii 92-107)[26] peptides. Our study has revealed two unique structural features of DQ2.5 that may promote its association with CLIP2. Firstly, DQ2.5 has an unusually large P4 pocket that can accommodate the bulky P4 Met of CLIP2. Secondly, DQ2.5 has a positively charged peptide binding groove that is electrostatically more compatible with the neutral CLIP2 compared to the positively charged CLIP1 peptide.

It was suggested that DQ2.5 cannot have the back bone hydrogen bond interaction with the P1 residues of the bound peptide[63]. In addition, the gluten derived (gliadin-α1a, LQPFPQPELPY, where the P1 residue is underlined) binds to DQ2.5 with two-fold higher affinity than its analog peptide containing norvaline (Nva) at P1 (~25µM)[67]. Nva is a non-proteinogenic alpha amino acid that is isosteric to Pro. This amino acid residue, however, has a primary amine group that has the ability to participate in the formation of hydrogen bond interactions. Therefore, gliadin-α1a must have an overall energetic advantage over the Nva substituted analog peptide for binding to DQ2.5, despite of the deficiency of a hydrogen bond at the P1 position. We propose that gliadin-α1a, and other peptides containing Pro at P1, have an entropic advantage that compensates for the lost enthalpy associated with the P1 hydrogen bond.

Another unusual characteristic of DQ2.5 is its CLIP-rich phenotype, account for 53% of the eluted peptide pool[17]. It was proposed that the CLIP-rich phenotype of DQ2.5 is explained by DQ2.5–CLIP being poor substrates for DM[16,28]. During MHCII maturation, DM catalyzes the release of CLIP from the nascent MHCII[12]. Therefore, impaired DQ2.5–DM interaction will result in DQ2.5 molecules retaining their original CLIP cargo. In contrast, DR1 expressing cells have a low abundance of CLIP[27], which suggests that DR1 is a good substrate for DM. We found two structural elements in DQ2.5 that may lower its DM sensitivity. First, α51, which is a key DM contacting residue in DR1, is positioned internally in DQ2.5 due to the α53 deletion mutation.

Second, the peptide binding groove residues that form the α9–α22–α24–α31–β86–β90 hydrogen bond network are not as free to move as the corresponding residues in DR1 (Figure 2.8). Therefore, DQ2.5 is less predisposed to the drastic secondary structure changes that DR1 undergoes upon DM binding. Our MD study showed that the α9–α22–α24–α31–β86–β90 hydrogen bond network is stable and the α51F of DQ2.5 cannot move into the P1 pocket upon DM binding. This is due to the blockage of the P1 pocket entrance by a water molecule which is part of the α9–α22–α24–α31–β86–β90 hydrogen bond network. To further test this idea, we disrupted the hydrogen bond network by mutating α24, α31 and β86 residues to the hydrogen bond non-permissible residues and then repeated the MD exercise. This time, α51F did translocate to fill the P1 pocket, similar to what happens in DR1 when interacts with DM. Our hypothesis that the α9–α22–α24–α31–β86–β90 hydrogen bond network leads to diminished DM sensitivity and ultimately to the CLIP-rich phenotype is supported by a recent study[28]. In that study the authors showed that the mutation of βE86A in DQ8 resulted in the increase of its DM sensitivity. In comparison with DQ2.5, DQ8 allele has the similar α9–α22–α24–α31–β86–β90 hydrogen bond network, but lack of the α53 deletion mutation. Our hypothesis, however, is based on the assumption that the DR1–DM interaction mechanism is directly applicable to the DQ2.5–DM interaction. It remains to be seen if the DR1–DM interaction mechanism is truly universal. Even if this should prove not to be the case, the preferences of bulky hydrophobic anchor residues at the P1 pocket for both DR1[68,69] and DQ2.5[70,71] indicate that these two molecules likely share the mechanistic feature of α51F translocating to fill the P1 pocket in the interaction with DM.

Two other human MHCII alleles, which have a deletion mutation at α53 like DQ2.5 and also contain the same set of residues that make up the α9–α22–α24–α31–β86–β90 hydrogen bond network in DQ2.5, are DQ4.4 (DQA1*04:01-DQB1*04:02) and DQ7.5 (DQA1*05:05-DQB1*03:01)[72]. We predict that DQ4.4 and DQ7.5 could be poor substrates for DM and also have a CLIP-rich phenotype like DQ2.5. Interestingly, all three MHCIIs, namely DQ2.5, DQ4.4, and DQ7.5, are all associated with one or more human autoimmune disorders. DQ2.5 is associated with celiac disease and type 1 diabetes, DQ4.4 is associated with juvenile idiopathic arthritis[73], and DQ7.5 is associated with celiac disease[17,74]. Currently, there is no known mechanistic link between decreased DM sensitivity and human autoimmune disorders. Further experiments are needed to validate this hypothesis.

Chapter 3
The Molecular Mechanism behind the Peptide-editing Process of MHCII by DM Catalyst: an MD Study

DM is important for editing the peptide from MHCII. The mechanism of this process could help to have a better view on how peptide was presented to T-cell, as well as autoimmune disease. The process is studied by a pure computational technique, namely, MD simulations.

3.1 Introduction

MHCII protein forms a complex structure with an antigenic peptide that is later presented to T cell receptors. This antigen presentation triggers an immune response in the event of the pathogen entry (see chapter 2 for more details). The peptide repertoire of MHCII is accumulated with the help of DM, a non-classical MHCII protein. The 3D structure of DM is similar to that of MHCII, but lacks a peptide binding groove. Although MHCII proteins have great allelic variability, DM is non-polymorphic. This intracellular chaperone, DM, when in complex with other MHCII molecules, is responsible for (i) the removal of the CLIP peptide, (ii) the exchange of low to high binding affinity peptides[75] and (iii) stabilization of peptide-free MHCII against MHCII inactivation[13,76,77]. In the absence of DM, the CLIP peptide is retained in MHCII and only a few antigens are presented to the T-cell receptors, making the process insufficient. The absence of DM also results in the aggregation of the peptide-free MHCII proteins[14,78].

Currently, in the PDB there are only two crystal structures of the DR–Antigen–DM complex[60] at different pH conditions: at pH 6.5 (PDB ID: 4GBX), and at pH 5.5 (PDB ID: 4FQX). The MHCII-bound antigen in this case is the hemagglutinin 306–318 peptide from influenza A virus (HA). The peptide exchange activity of DM is promoted under slightly acidic pH conditions (pH 5.5). The crystal structures show that when DM interacts with DR, the α43W residue of DR flips its χ_1 side chain torsion angle by 120°. The α52-55 and β86-91 regions of DR also undergo secondary conformational changes (see chapter 2 for more details). The conformational change in α52-55 results in an intramolecular change in DR with α51F occupying the aromatic and hydrophobic P1 pocket. As a consequence, the pocket is stabilized when the P1 residue of the peptide is released. In addition, the β89F residue that is close to the P1 pocket is also conjectured to contribute to that stabilization.

Although the crystal structure clearly showed the interaction interface between DR and DM, the mechanism of conformational changes of residues in DR and DM from the non-interacting state (apo DR/DM) to the interacting state (holo DR/DM) remains unclear. The exact mechanism underlying DM catalysis and its pH dependence for the process of peptide exchange is also unknown. Even though both DR–HA–DM structures at pH 5.5 and pH 6.5 have the linked peptide starting at P2, the peptide is only visible in the electron density from P5 to P11 for the crystal structure at pH 5.5. The mechanism of peptide release from DR and the stabilization of empty DR by DM are poorly understood. This could result from the non-availability of the DR–DM complex structure in the absence of the peptide. It was previously shown that the peptide binding groove was closed during MD simulations of antigen peptide free DR protein[79]. It was proposed that when interacting with DM, DR could retain its receptive/open conformation.

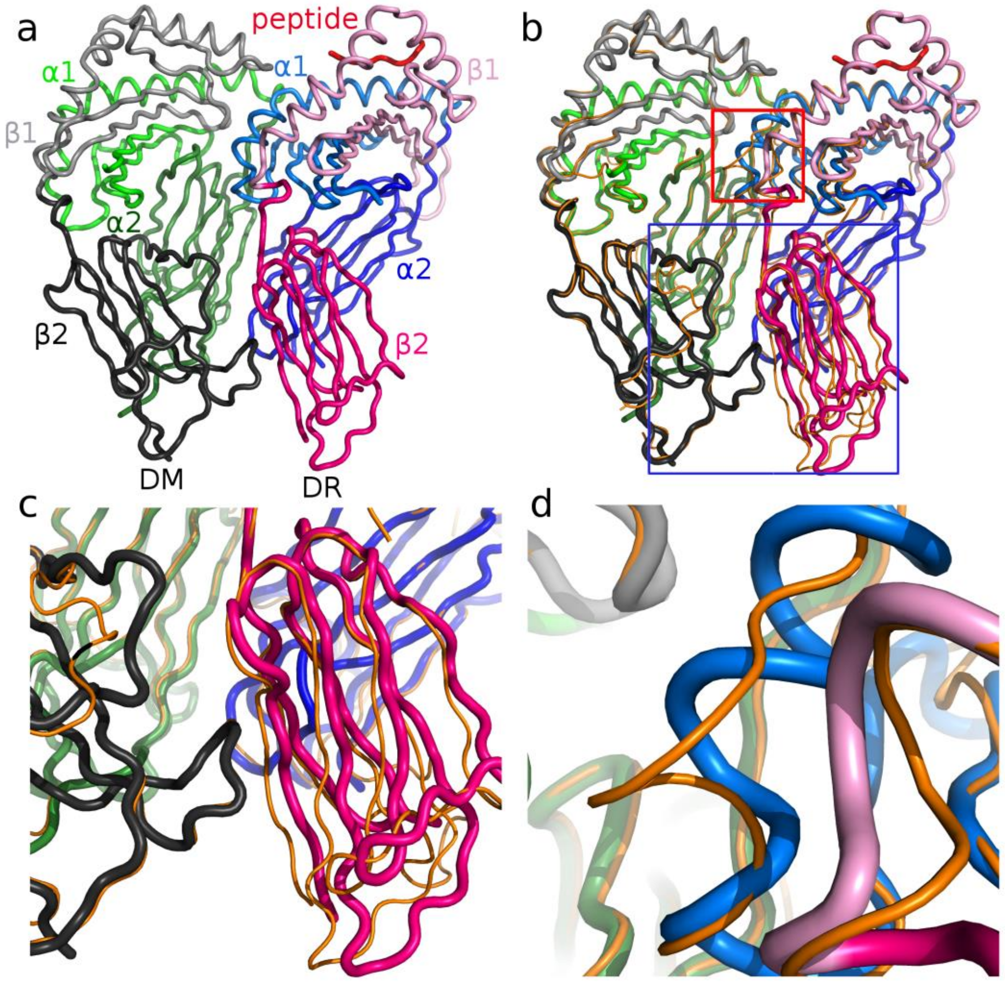


Figure 3.1: The apo and the holo conformations of DR/DM proteins. (a) Crystal structure of DR–HA–DM complex at pH 5.5 (holo) (PDB ID: 4FQX), (b) superimposition of DR–CLIP1 (apo) (PDB ID: 3PDO) and DM (apo) (PDB ID: 2BC4) onto DR–HA–DM complex. The α1 (α2-79), α2 (α80-181), β1 (β-5-88), and β2 (β89-190) domains of DR are in marine, blue, pink and hot pink, respectively. The α1 (α13-93), α2 (α94-200), β1 (β3-88), and β2 (β89-193) domains of DM are in green, forest, gray, and dark gray, respectively. The peptide of DR–HA–DM is in red. The peptide of DR–CLIP1 is omitted for clarity. The backbone of the apo DR and the apo DM is represented in orange. (c) and (d) are the zoomed in views of the region in (b) that have been boxed in blue (β2 domains of DR and DM) and red (α1 and β1 regions in DR, which are close to the P1 pocket), respectively. The two regions of (b) highlighted in boxes are the regions that have different conformations between the apo and the holo forms.

To answer the question regarding the mechanism describing the changes in conformations of residues in DR and DM as the system undergoes from the non-interacting state to the interacting state, we carried out MD simulations of the DR–HA–DM complex where DR and DM have been taken from the crystal structures describing their apo conformations (Figure 3.1). The protonation states of ionizable residues correspond to the pH where the DR–DM interaction occurs (pH 5.5). The apo and the holo terms for DM refer to the conformation of DM in its free state and in its DR-bound state, respectively. Similarly, the apo and the holo terms for DR refer to the conformations of the DR-peptide complex in its free state and DM-bound state, respectively. The effect of pH was studied by carrying out MD simulations of the same DR–HA–DM complex, but with the protonation state corresponding to pH 6.5. The peptide editing process by DM was studied by carrying out MD simulations of the DR–HA–DM crystal structure (PDB ID: 4FQX) using the DR–HA (PDB ID: 3PDO) system as a control. Finally, the stabilization of peptide-free DR was examined by carrying out MD simulations of the peptide-free DR–DM complex, with the peptide-free DR system as a control.

3.2 Introduction to MD simulations

3.2.1 Definition and history

Molecular dynamics simulation is a computational method that calculates the behavior of a molecular system as a function of time. This method provides detailed information on the fluctuations as well as conformational changes of biological molecules. Therefore, it is used to study the dynamic properties of biological molecules as well as their complexes. This method is also applied to determine and refine structures obtained by NMR and X-ray experiments.

MD simulations were first introduced in 1955 by Fermi, Pasta, and Ulam[80], in 1957 by Alder and Wainwright[81,82] and in 1964 by Rahman[83]. This was followed by simulations on a realistic system, *i.e.* liquid water[84]. The first simulations on a protein system[85], the bovine pancreatic trypsin inhibitor, were carried out in 1977. Nowadays, MD simulations are performed on various systems, including protein-DNA complexes, protein-protein complexes, lipid systems, ligand-bound systems, studying protein folding *etc.*

The basic algorithms of MD simulations include (i) dividing time into discrete time steps (ii) computing the forces on each atom at each time step depending on the molecular mechanic force field employed to model the interactions between atoms, (iii) determining the new position and velocity of each atom by numerically solving Newton's equations of motion (equation 3.1).

3.2.2 Newton's equation of motion

$$\frac{dv}{dt} = \frac{F(r)}{m} \quad (3.1)$$

where $\frac{dv}{dt}$ is the derivative of velocity v with respect to time t,

m is mass of the atom

$F(r)$ is the force on an atom i, and

r is the position of the atom i

An analytical solution of the equation 3.1 is impossible. However, the equation can be numerically solved as

$$v_{i+1} = v_i + \delta t \,.\frac{F(x_i)}{m} \text{ (3.2)}$$

$$\text{and } x_{i+1} = x_i + \delta t \,.v_i \text{ (3.3)}$$

where δt is the time step

3.2.3 Derivative of potential energy

An energy function (also referred to as the molecular mechanic force field) was used to determine the force on each atom. The potential energy $V(R)$ is calculated as a sum of internal (or bonded) and a sum of external (or non-bonded) terms

$$V(R) = E_{bonded} + E_{non-bonded} \text{ (3.4)}$$

3.2.3.1 Bonded energy (E_{bonded}) is calculated as

$$E_{bonded} = E_{bond-stretch} + E_{bond-bend} + E_{rotate\ along\ bond} \text{ (3.5)}$$

(1) The bond-stretching energy ($E_{bond-stretch}$) is the elastic interaction between a pair of atoms connected by a covalent bond and is calculated as follows

$$E_{bond-stretch} = \sum_m k_m^l \,(l_m - l_m^o)^2 \text{ (3.6)}$$

where l_m is the distance between two atoms of the *m*th bond

l_m^o is the bond length at equilibrium the *m*th bond

k_m^l is the force constant that determines the strength of the bond

(2) The angle bending energy ($E_{bond-bend}$) is the interaction among three covalently-bonded atoms that form a stable angle. This energy term is calculated as follows

$$E_{bond-bend} = \sum_m k_m^{\theta} \, (\theta_m - \theta_o^m)^2 \text{ (3.7)}$$

where θ_m is the mth angle of the two adjacent bonds sharing a common atom

θ_o^m is the bond angle at equilibrium

k_m^{θ} is the force constant that determines the geometry of the bond

(3) The torsional energy ($E_{rotate\ along\ bond}$) is the interaction among four covalently-bonded atoms that form a stable dihedral angle. This energy term is calculated as follows

$$E_{rotate\ along\ bond} = \sum_m k_m^{\emptyset} \, [1 + \cos(n_m \emptyset_m + \delta_m)] \text{ (3.8)}$$

where $\emptyset_m$ is the mth dihedral angle between two adjacent angles sharing a common bond

n_m is the periodicity factor which determines the number of equilibrium dihedral angles in a 360° rotation

δ_m is the phase shift

$k_m^{\emptyset}$ is the amplitute

3.2.3.2 Non-bonded energy ($E_{nonbonded}$) is calculated as follows

$$E_{nonbonded} = E_{vdw} + E_{elec} \text{ (3.9)}$$

(1) van der Waals interactions (E_{vdw}) are induced electrical interactions between two or more closely located, but not bonded, atoms or molecules.

$$E_{vdw} = \sum_i \sum_{j,i \neq j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] \text{ (3.10)}$$

where r_{ij} is the distance between the atom i and the atom j

ε_{ij} is the van der Waals dissociation energy

σ_{ij} is the collision diameter

(2) Electrostatic interaction E_{elec} is calculated as follows

$$E_{elec} = \sum \frac{q_i q_j}{\varepsilon r_{ij}} \quad (3.11)$$

where $q_i q_j$ are partial charges on the atom i and on the atom j

r_{ij} is the distance between the atom i and the atom j

ε is the dielectric constant

The neighbour list (*i.e.* list of each atom and its immediate neighbours) is recomputed every few steps.

The electrostatic potential is stronger and more long-range than the van der Waals potential.

In MD simulations, the non-bonded interactions are more important than the bonded interactions, because these non-bonded interactions are the intermolecular interactions which affect the secondary structures and the assemblies.

3.3 Methodology

3.3.1 Preparation of Starting Structures for MD Simulations.

500 ns MD simulations were performed on six systems. Each simulation carried out in triplicate with different initial conditions to ensure larger sampling (Table 3.1). The structural model of DR–HA–DM was built using MODELLER version 9.10[39,40].

Table 3.1: Systems for MD simulations

System	Peptide start (peptide length)	Starting structure	Net-charge	Number of solvent waters
Model_5.5	P1 (11)	2BC4 (DM) 3PDO (DR–CLIP1)*	-10	51855
Model_6.5	P1 (11)	2BC4 (DM) 3PDO (DR–CLIP1)*	-21	51612
DR–HA–DM	P5 (7)	4FQX&	-24	52331
DR–DM	-	4FQX&	-25	52358
DR–HA	P5(7)	4FQX (HA)& 3PDO (DR)	-12	39771
DR	-	3PDO	-13	39355

*The wild-type forms of α165D (DM), β46S (DM), and β92D (DM) were used because these residues were mutated in the crystal structures of DR–CLIP1 complex and DM protein.

& The wild-type forms of α65V (DR), and β30C (DR) were used because these residues were mutated in the crystal structures of DR–HA-DM.

3.3.2 MD Simulations Protocol

All-atom MD simulations were performed using the Gromacs-5.1.4[86] package with the AMBER99SB force field[47]. SPC (simple point charge) water[87] box with a minimum distance of 12 Å (Table 3.1) between any protein atom and the boundary of the box, was used to solvate each system. Periodic boundary conditions were applied. The protonation states of all ionizable residues were assigned according to the pK$_a$ predicted by the DEMM program[88,89] (also see chapter 5 for more details). The structures were also manually inspected to correct for protonation states according to plausible/potential interactions, especially in the case of assigning protonation states at N$^{\varepsilon1}$ and N$^{\delta2}$ in HIS. All systems were neutralized with sodium counter-ions (Table 3.1). All missing side chain atoms were built using MODELLER and hydrogen atoms were added using GROMACS. Four ASP, six GLU and 14 HIS residues had differing protonation states in the different MD systems or had pK_a values different from the standard state values (Table 3.2). The residues that had the HIP form of Histidine (protonation at both $N^{\varepsilon 1}$ and $N^{\delta 2}$) were α5, α177 (DR), β16, β81, β111, β112 (DR), α16, α137, and α138 (DM). All the N-terminal and C-terminal residues were capped by an acetyl group (ACE) and an ethylamine group (NHE), respectively.

The grid method was used to determine the neighbor list. This neighbour list and long-range forces were updated every 20 steps. The cut-off threshold for short-range forces, electrostatics and van der Waals was set to 10 Å. Particle Mesh Ewald method[90] was applied for the treatment of long-range electrostatics interactions. The system pressure was maintained by coupling to a Parrinello-Rahman barostat at 1 bar with a coupling constant $\tau P=2$ ps. The isothermal compressibility was set to 4.5×10^{-5} bar^{-1} along all box dimensions. The bond lengths were restrained using the LINCS algorithm[91]. The temperature of the system was coupled using velocity rescaling with a stochastic term.

The system was first minimized with a maximum force of 900, 950 or 1000.0 kJ/(mol.nm) in each of the triplicate runs respectively. The whole system was then submitted to MD simulations for 20 ns in the NVT ensemble, and 20 ns in the NPT ensemble, with the position restraints on heavy atoms of the proteins. This was followed by unrestrained NPT ensemble MD simulations which were performed for 500 ns, where T and P were set at 300 K and 1.0 bar, respectively. The MD time step for integration was 2 fs, and the trajectory was saved every 1ns for further analysis.

Table 3.2: Protonation states of the ionizable residues that differ from their canonical states

Residue number	Chain	model 5.5	model 6.5	DR–HA–DM*
21	A	GLH	GLU	GLH
29	A	ASH	ASP	ASH
30	A	GLH	GLU	GLH
33	A	HIE	HID	HIE
46	A	GLH	GLU	GLU
66	A	ASH	ASH	ASH
143	A	HID	HIE	HIE
149	A	HIE	HIE	HIP
167	A	HIE	HIE	HIE
2	B	-	-	ASP
176	B	GLH	GLU	GLU
177	B	HIE	HIE	HIE
20	C	HIE	HIE	HIP
35	C	GLH	GLH	GLU
180	C	HIE	HIE	HIE
6	D	HID	HID	HIP
31	D	ASH	ASP	ASP
47	D	GLH	GLU	GLU
61	D	HIP	HIE	HIP
82	D	HID	HID	HIP
141	D	HIP	HIE	HIP
145	D	HIP	HIE	HIP
161	D	HIP	HID	HIP
178	D	HIE	HIE	HIE

* the protonation state of ionizable residues in DR–DM , DR–HA and DR were the same as those in DR–HA–DM.

where: HIE is HIS with a proton at $N^{\varepsilon 1}$;

HID is HIS with a proton at $N^{\delta 2}$;

HIP is HIS with protons at both $N^{\varepsilon 1}$ and $N^{\delta 2}$;

ASH is ASP with a proton at $O^{\delta 2}$;

GLH is GLU with a proton at $O^{\varepsilon 2}$

3.3.3 Analysing the MD Simulations

The temperature and potential energy of all the systems during the MD simulations were generated using the g_traj module of the GROMACS package. In all analyses, residues at the N- and C-termini were omitted from consideration as these regions undergo large fluctuations and may confound dynamics/correlation analysis. Only the following regions were considered: α6-178 (DR), β6-186 (DR), α18-194 (DM), and β6-190 (DM). The triplicate simulations were combined and different analyses of C^{α} RMSD, and RMSF of MD trajectories were performed with the help of the R program (http://www.R-project.org.) using the Bio3d package[92]. The PCA and distance analyses were done using in-house scripts. Movies were made using the VMD[93] program. All the plots were obtained using the Python and GIMP programs. The representative complex structure was made using the Chimera[42] and Pymol[43] programs. The secondary structure assignment was calculated using the DSSP algorithm[94].

3.4 Results

3.4.1 Model Structures

The modelled structure of the DR–HA–DM complex has a C^{α} RMSD of 1.55, 1.13, 1.08, and 0.24 Å in comparison to crystal structures of the holo DR–HA–DM at pH 5.5 (PDB ID: 4FQX), the holo DR–HA–DM at pH 6.5 (PDB ID: 4GBX), the apo DR–CLIP (PDB ID: 3PDO), and the apo DM (PDB ID: 2BC4), respectively (Figure 3.1).

3.4.2 Temperature and Potential Energy

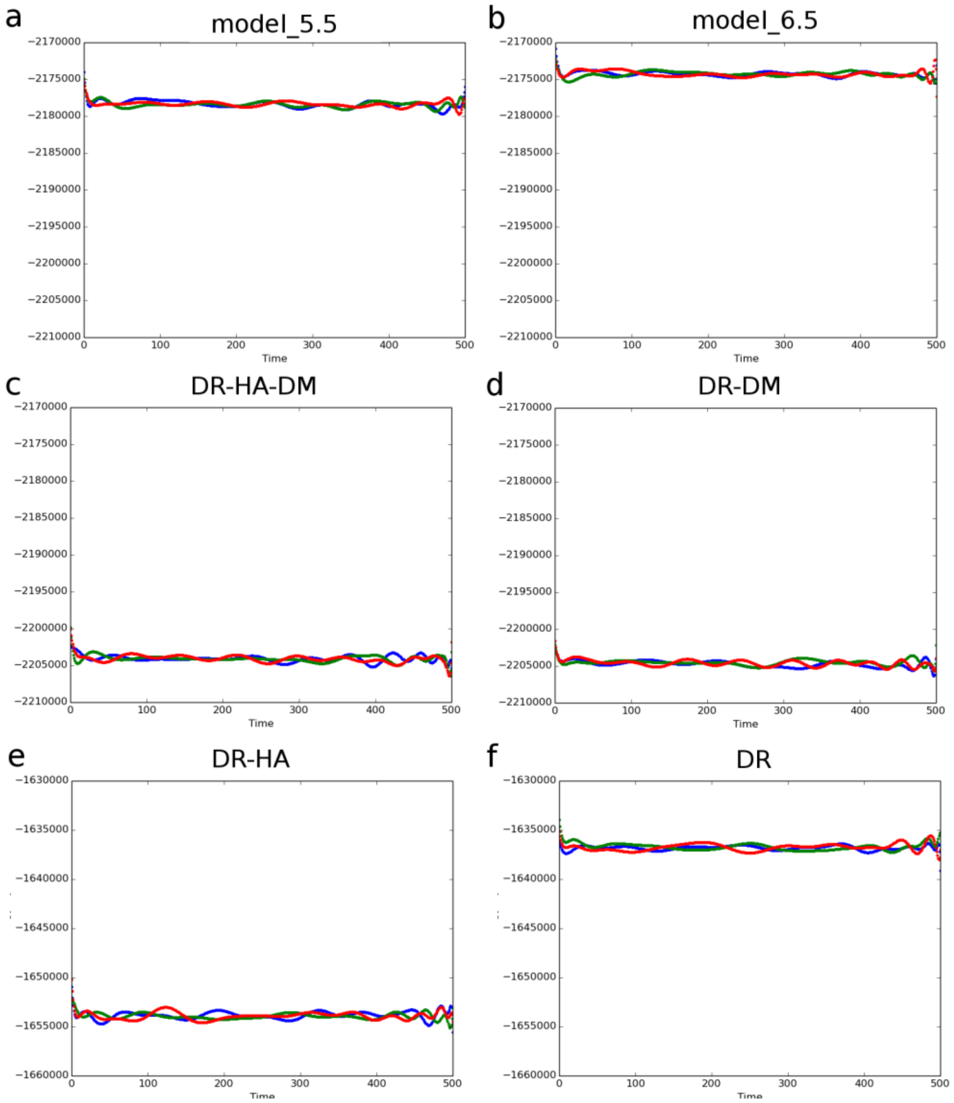


Figure 3.2: Potential energy (kJ/mol) during 500 ns MD simulations in the six systems. (a) Model_5.5, (b) model_6.5, (c) DR–HA–DM, (d) DR–DM, (e) DR–HA and (f) DR. Triplicate runs are shown in different colors (blue, green and red).

The potential energies of all MD systems (Figure 3.2) are stable. The energy fluctuations are around -2,175, 000 kJ/mol for model_5.5 and model_6.5. Those energies are around -2, 295, 000 kJ/mol for DR–HA–DM or DR–DM systems. The systems without DM have energies around -1,655, 000 (DR–HA) and -1,635, 000 (DR) kJ/mol. All the simulations have energy fluctuation less than 0.1%. The temperature of all systems (Figure 3.3) during MD simulations is kept constant at 300K.

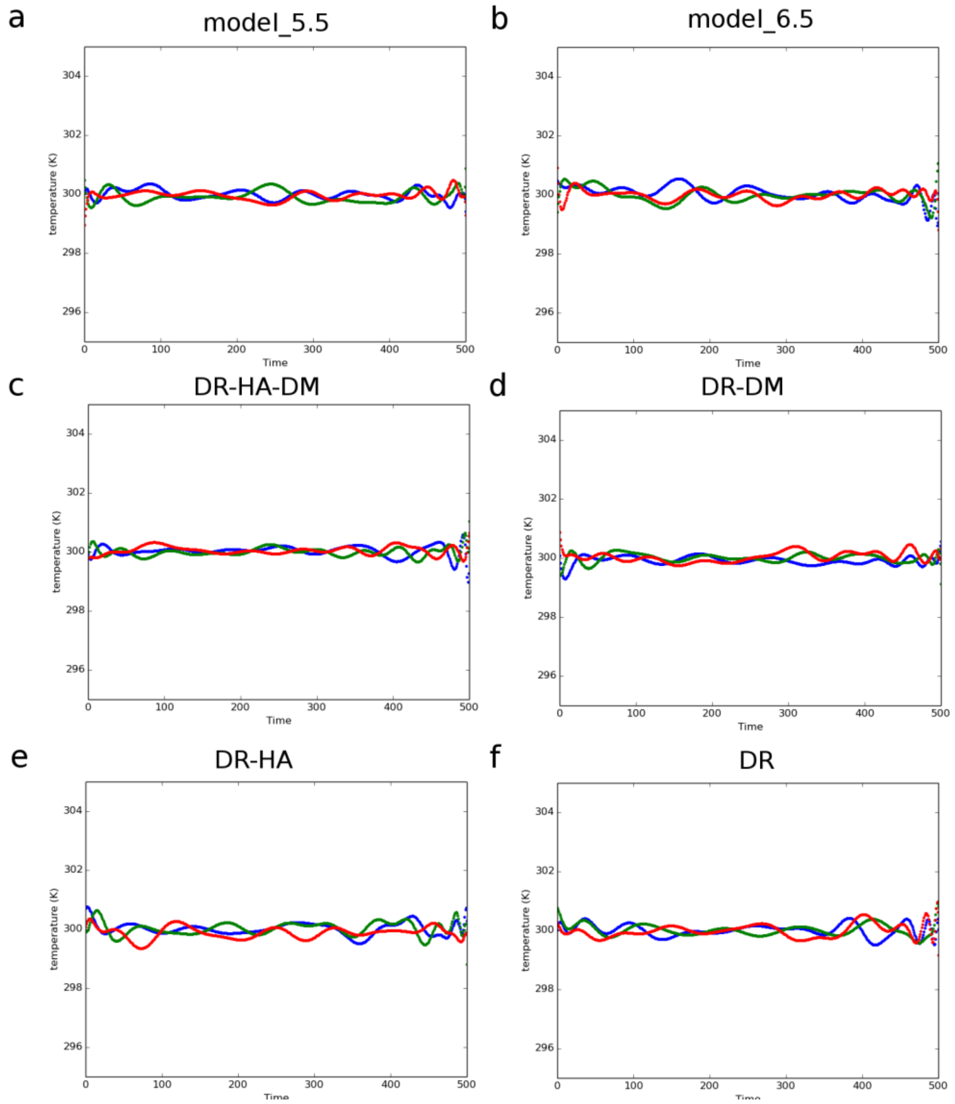


Figure 3.3: Temperature (in K) during 500 ns MD simulations of the six systems. (a) Model_5.5, (b) model_6.5, (c) DR–HA–DM, (d) DR–DM, (e) DR–HA and (f) DR. Triplicate runs are shown in different color (blue, green and red).

3.4.3 Mobility of the Whole DR–DM Complex during 1.5 μs MD Simulations

(a) DR mobility

In general, the RMSD value with respect to the average structure (green line) is lower than the value with respect to the starting structure (blue line) (Figure 3.4). The RMSD plot of the MD simulations for DR–HA was only shown up to 1μs because in one of the triplicate simulations the DR β2 domain rotates by about 90° compared to the crystal structure (Figure 3.6). We did look at the torsion angle but did not find that the rotation is due to a transition of a single torsion angle. It could instead be because of several angles. The rotation could be an artifact of the MD simulations, so we left that simulation out. Among the six systems, DR in peptide-free DR simulation has the most fluctuations (2.65 Å and 1.93 Å with respect to the starting and average structures), while the DR protein in model_6.5 has the least fluctuation (average RMSD of 1.88 Å and 1.52 Å with respect to the starting and average structures, respectively).

(b) DM mobility

For the DM fluctuation, among the four systems the DM protein in model_6.5 has the most fluctuation (average RMSD of 2.21 Å and 1.38 Å with respect to the starting and average structures, respectively), while this protein in DR–DM systems has the least fluctuation (average RMSD of 1.82 Å and 1.07 Å with respect to the starting and average structures, respectively).

This analysis gives the overall view of how each system stabilizes during 1.5 μs MD simulations. Next we analyzed how each residue in each system has changed in the simulations.

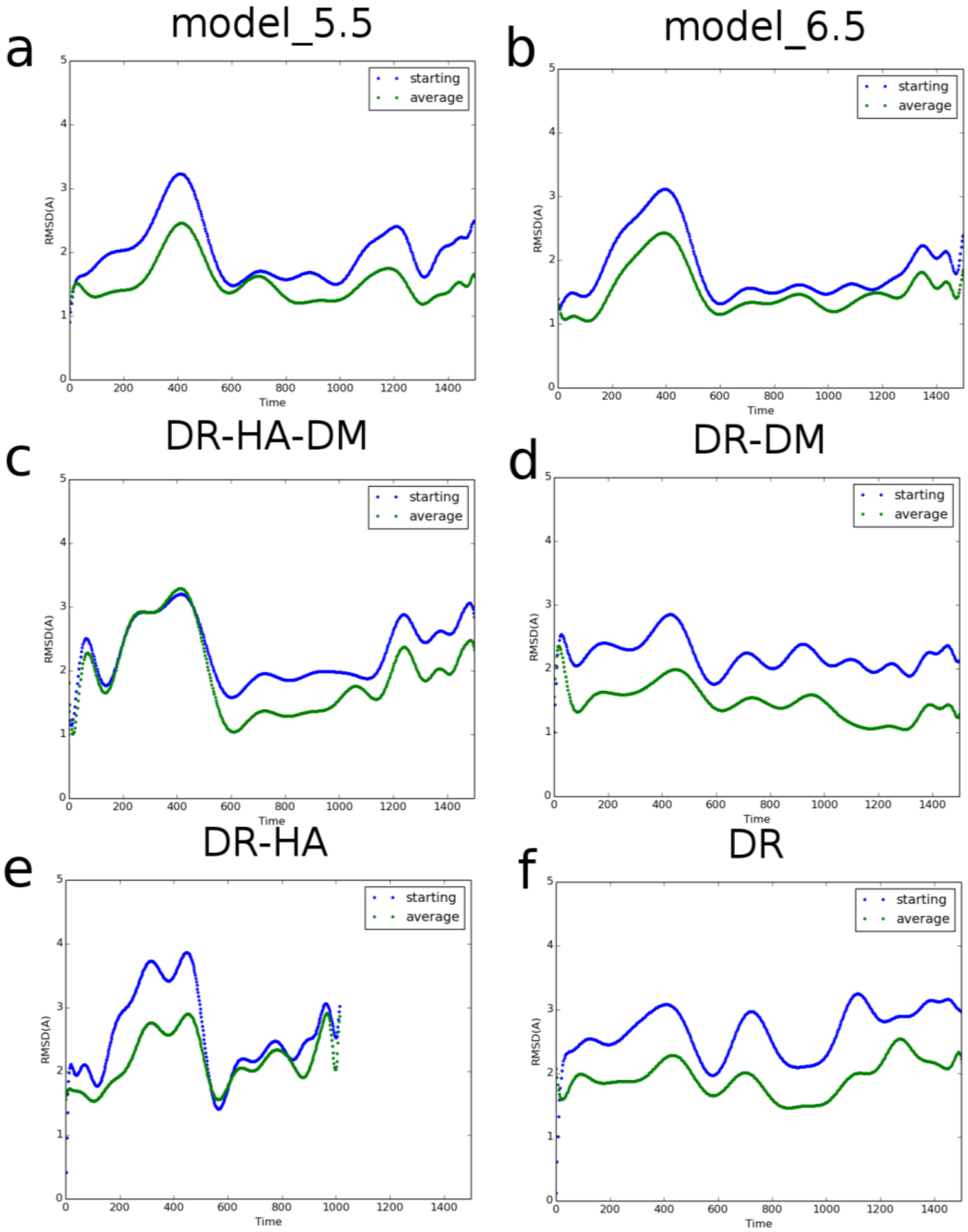


Figure 3.4: RMSD of DR protein. (a) Model_5.5, (b) model_6.5, (c) DR–HA–DM, (d) DR–DM, (e) DR–HA and (f) DR. The RMSD values with respect to the starting and the average structures are in blue, and green, respectively. The x axis is the running time (in ns). The y axis is the RMSD values (in Å).

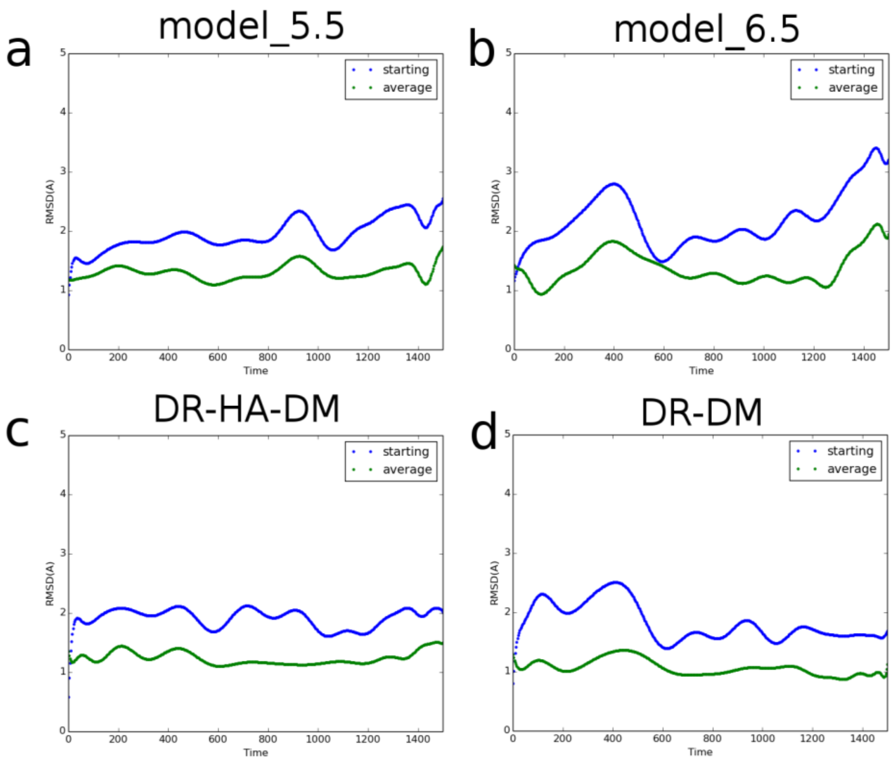


Figure 3.5: RMSD of DM protein. (a) Model_5.5, (b) model_6.5, (c) DR–HA–DM and (d) DR–DM. The RMSD values with respect to the starting and the average structures are in blue, and green, respectively. The x axis is the running time (in ns). The y axis is the RMSD values (in Å).

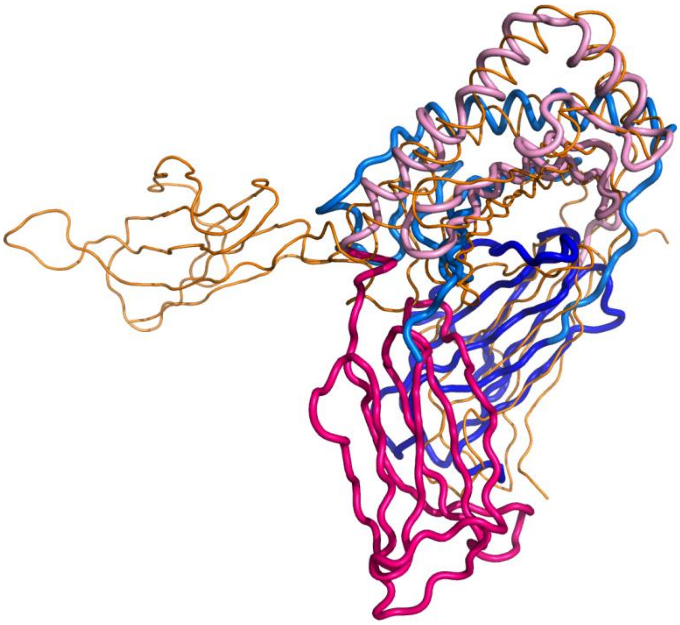


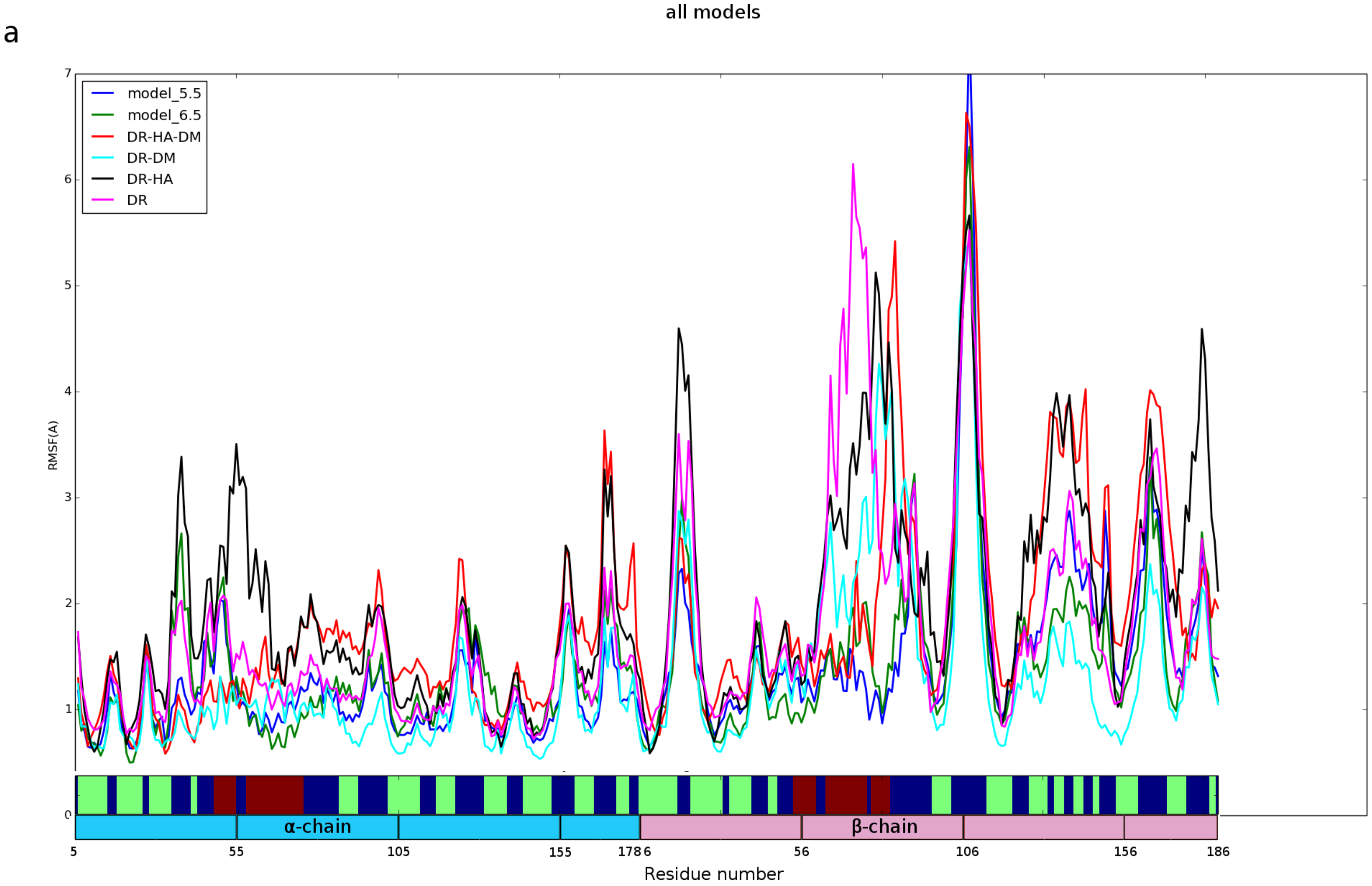
Figure 3.6: The conformation of DR in the DR–CLIP1 crystal structure and in the 500 ns snapshot. The crystal structure is represented in tube with the color code is the same as Figure 3.1. The snapshot is represented in orange.

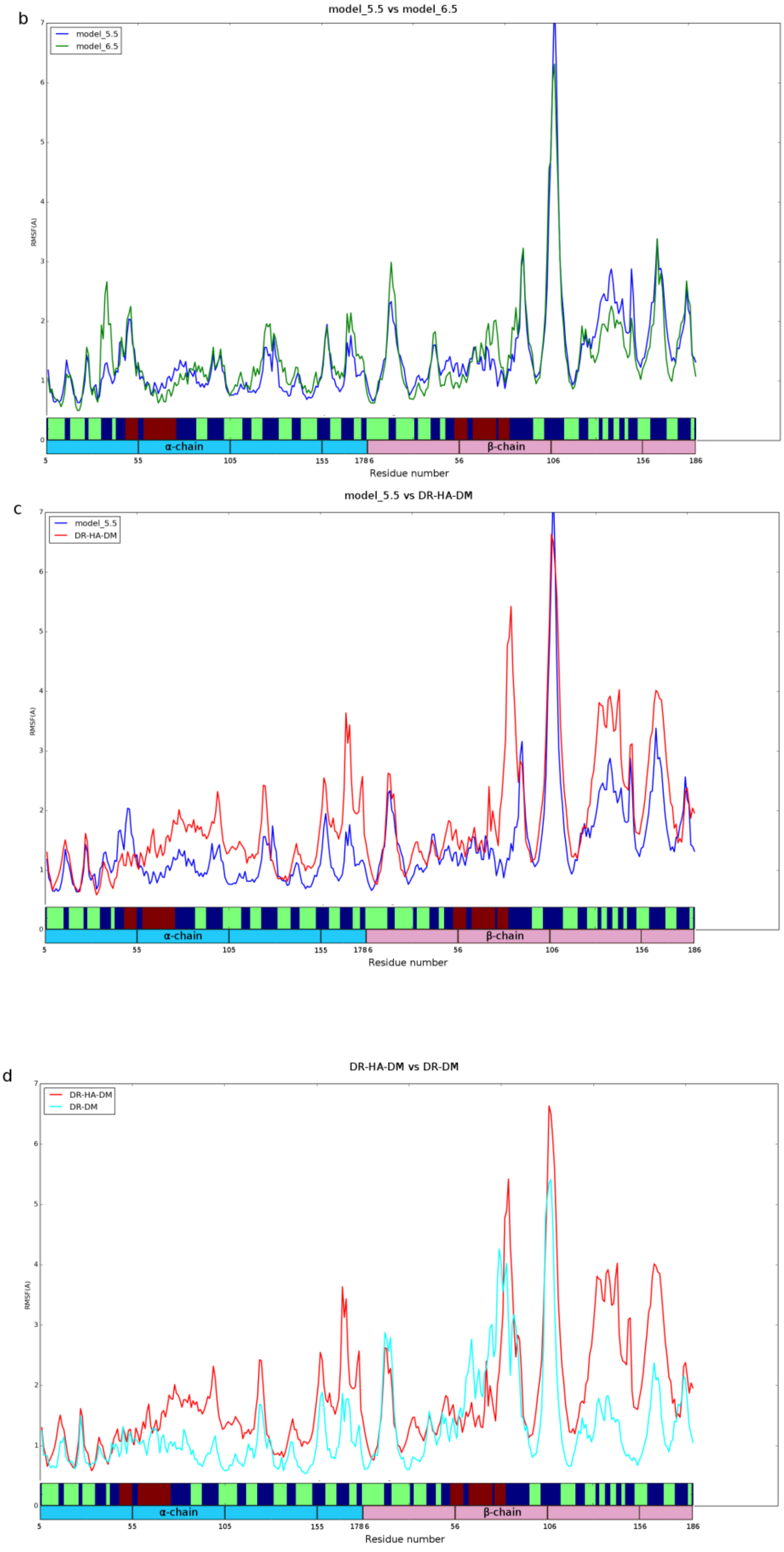
3.4.4 Average Atomic Mobility Reveals Important Region of DR/DM in MD Simulations

(a) DR mobility

The fluctuations of each residue in the six systems were analyzed using RMSF (root-mean-square fluctuation). The RMSF plot of DR in the six MD systems (Figure 3.7a) shows the different conformational behaviors in different parts of the protein. There are several regions where the fluctuations are higher than 2 Å, namely, α35-40, α46-47, α49-65, α78, α99-100, α124-125, α156-159, α168-173, α177-178, β17-24, β43, β64-94, β104-116, β126-152 and β160-186. The missing region in the crystal structures, β104-116, has the fluctuation higher than 3 Å, with the peak at β108-109 in all the six systems. The β2 domains in DR–HA–DM and DR-HA complexes have a fluctuation of more than 2 Å in almost all their component residues.

Different pair-wise comparisons have been made. Only the regions that have RMSF differences higher than 1 Å are discussed here. Model_5.5 has residues where the RMSF values differ by more than 1 Å from model_6.5 at residues α37-38 (Figure 3.7b). This could be due to the fact that they are located in the loop region of the α1 domain. The RMSF values in model_5.5 are different from those in DR–HA–DM (Figure 3.7c) at the following regions: α168-173 (α2 domain), β84-88 (peptide binding domains), and β133-145 (β2 domain). The RMSF value differences between DR–HA–DM and DR–DM (Figure 3.7d) are at the α2 (α169-171) and the β2 (β133-145) domains. The DR proteins in the DR–HA–DM and DR–HA complexes (Figure 3.7e) show differences in the peptide binding region (α37-38, α53-58, β18-23, and β72-82) and β2 domain (β178-183). The DR–HA complex has a fluctuation in the peptide binding groove of the α chain (α53-58). This is close to the DM interaction site and P1 pocket. The RMSF values of DR–DM and DR are different only at the β72-82 region (Figure 3.7f). The DR–HA and DR systems have fluctuation differences at α17-24, α53-58, β67-71, β141-143 and β178-183 (Figure 3.7g).


a
all models
model_5.5
model_6.5
DR-HA-DM
DR-DM
DR-HA
DR
RMSF(A)
7
6
5
4
3
2
1
0
α-chain
β-chain
5
55
105
155
178
6
56
106
156
186
Residue number


b
model_5.5 vs model_6.5
model_5.5
model_6.5
RMSF(A)
0
1
2
3
4
5
6
7
α-chain
β-chain
5
55
105
155
178
6
56
106
156
186
Residue number
c
model_5.5 vs DR-HA-DM
model_5.5
DR-HA-DM
d
DR-HA-DM vs DR-DM
DR-HA-DM
DR-DM

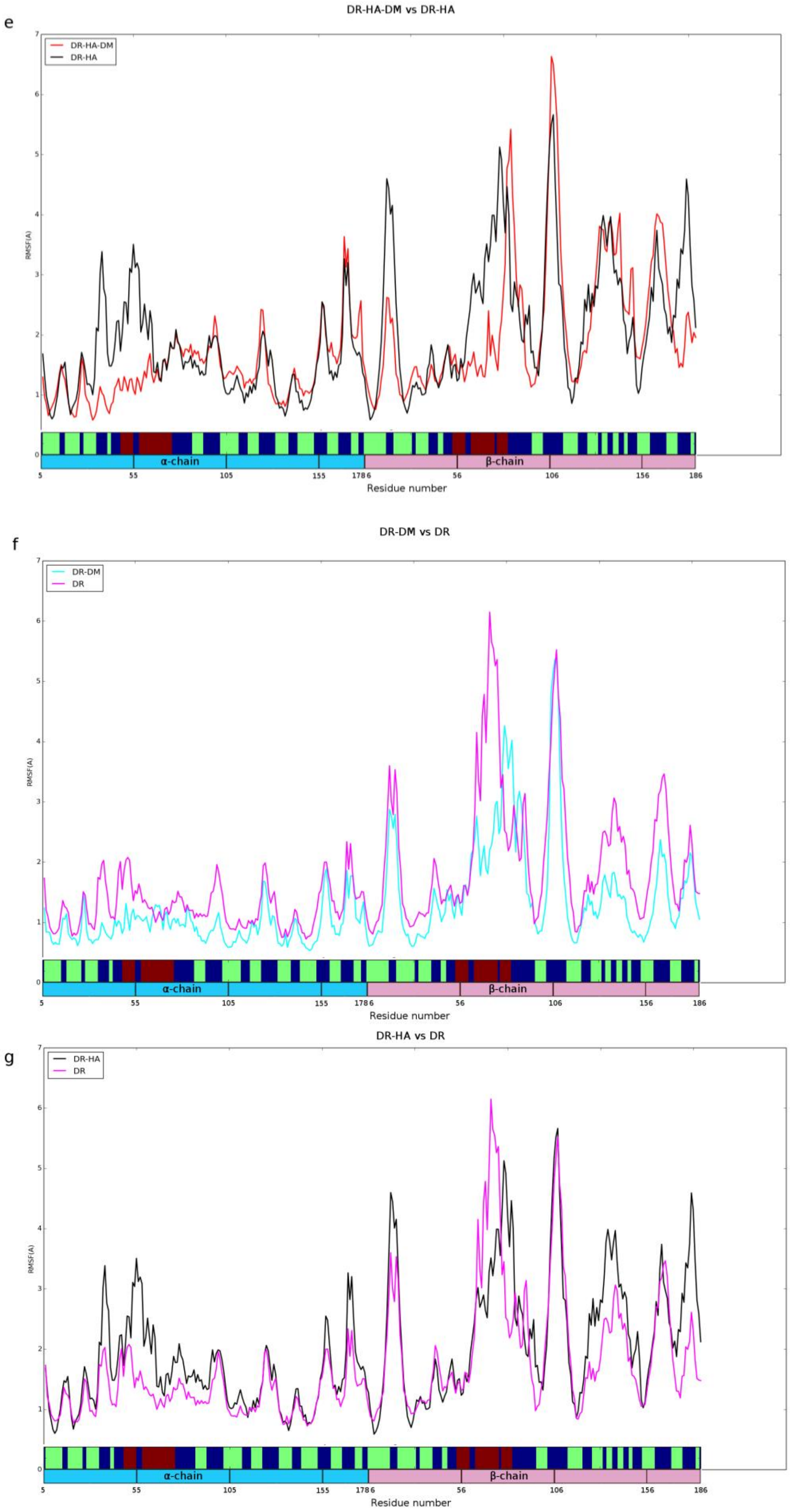
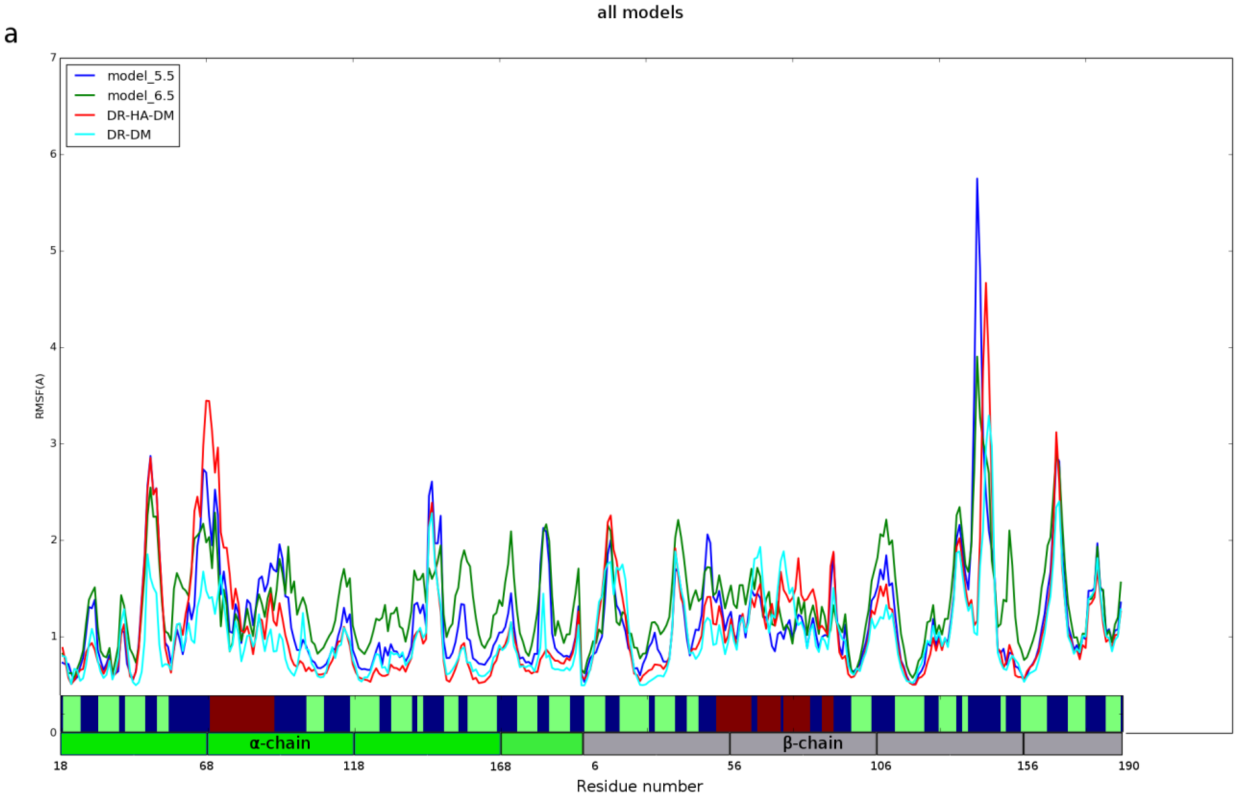

e
DR-HA-DM vs DR-HA
DR-HA-DM
DR-HA
RMSF(A)
0
1
2
3
4
5
6
7
α-chain
β-chain
5
55
105
155
178 6
56
106
156
186
Residue number
f
DR-DM vs DR
DR-DM
DR
RMSF(A)
0
1
2
3
4
5
6
7
α-chain
β-chain
5
55
105
155
178 6
56
106
156
186
Residue number
g
DR-HA vs DR
DR-HA
DR
RMSF(A)
0
1
2
3
4
5
6
7
α-chain
β-chain
5
55
105
155
178 6
56
106
156
186
Residue number

Figure 3.7: RMSF of DR in the six systems during 1.5 µs simulations. The RMSF values of model_5.5, model_6.5, DR–HA–DM, DR–DM, DR–HA and DR are represented in blue, green, red, cyan, black, and magenta, respectively. (a) All systems combination, (b) model_5.5 vs. model_6.5, (c) model_5.5 vs. DR–HA–DM, (d) DR–HA–DM vs. DR–DM, (e) DR–HA–DM vs. DR–HA, (f) DR–DM vs. DR, and (g) DR–HA vs. DR. The secondary structure assigned by DSSP are along the x-axis (above), where α-helix, β-strand and coil are in red, green and blue, respectively. The residue numbers of α and β chains (below) are shown in cyan and pink as in the Figure 3.1.

(b) DM mobility

The DM fluctuations of these four systems (Figure 3.8) are more consistent than the DR fluctuations. Fluctuations higher than 2 Å are observed at the following regions: α47-50, α63-72, α143-144, α147, α171, α182-183, β15-16, β38-40, β49, β108-110, β134-136, β152, and β166-170. The regions of α67-69 and β140-146 have RMSF fluctuation higher than 3 Å in three out of the four systems (model_5.5, model_6.5 and DR–HA–DM). Among the four systems, only the β152 residue in model_6.5 has a fluctuation higher than 2 Å.

all models
a
model_5.5
model_6.5
DR-HA-DM
DR-DM
RMSF(A)
0
1
2
3
4
5
6
7
α-chain
β-chain
18
68
118
168
6
56
106
156
190
Residue number

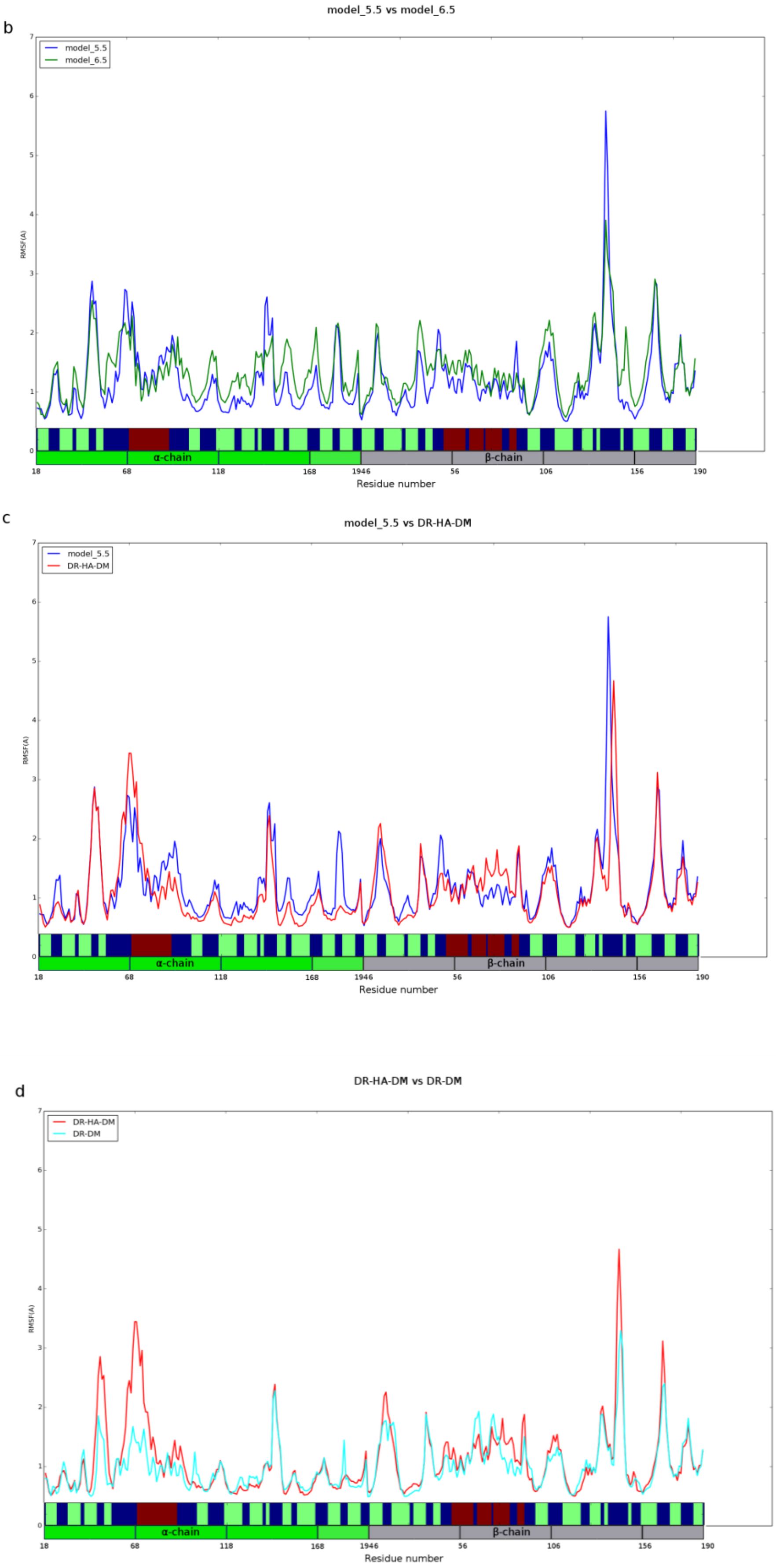

b
model_5.5 vs model_6.5
model_5.5
model_6.5
RMSF(A)
0
1
2
3
4
5
6
7
α-chain
β-chain
18
68
118
168
1946
56
106
156
190
Residue number
c
model_5.5 vs DR-HA-DM
model_5.5
DR-HA-DM
RMSF(A)
α-chain
β-chain
Residue number
d
DR-HA-DM vs DR-DM
DR-HA-DM
DR-DM
RMSF(A)
α-chain
β-chain
Residue number

Figure 3.8: RMSF of DM in the four systems during 1.5 μs simulations. The RMSF values of model_5.5, model_6.5, DR–HA–DM, DR–DM, DR–HA and DR are represented in blue, green, red and cyan, respectively. (a) All systems combination, (b) model_5.5 vs. model_6.5, (c) model_5.5 vs. DR–HA–DM, and (d) DR–HA–DM vs. DR–DM. The secondary structure assigned by DSSP are along the x-axis (above), where α-helix, β-strand and coil are in red, green and blue, respectively. The residue numbers of α and β chains (below) are shown in green and gray as in the Figure 3.1.

3.4.5 Principal Component Analysis Shows the Dynamical Correlation of the DR β2 Domain

The essential dynamics of the six systems during the MD simulations were monitored by PCA (principal component analysis).

(a) PCA of DR

The porcupine plot of the first principal component shows the fluctuation of α-helical region of peptide binding domain (α1 and β1), β2 domain and α2 domain. The fluctuation of peptide binding region and β2 domain in DR increases in the absence of DM *i.e.* when comparing DR–HA–DM with DR–HA and DR–DM with DR (Figure 3.9c, d, e and f). The β2 domain of DR in DR–HA–DM complex moves towards DM (Figure 3.9c), while in other five cases this domain moves away from DM (Figure 3.9d). Both model_5.5 and model_6.5 systems show less fluctuations, in comparison with DR–HA–DM.

(b) PCA of DM

Similar to RMSF, the fluctuation amplitude of PC1 in DM is smaller than in DR. The porcupine plots for DM show that only the β2 domains of DM in model_5.5 and DR–HA–DM have the tendency to come closer to the β2 domains of DR (Figure 10). In model_5.5, model_6.5 and DR–HA–DM complexes the motion is localized to the region surrounding α64 residue.

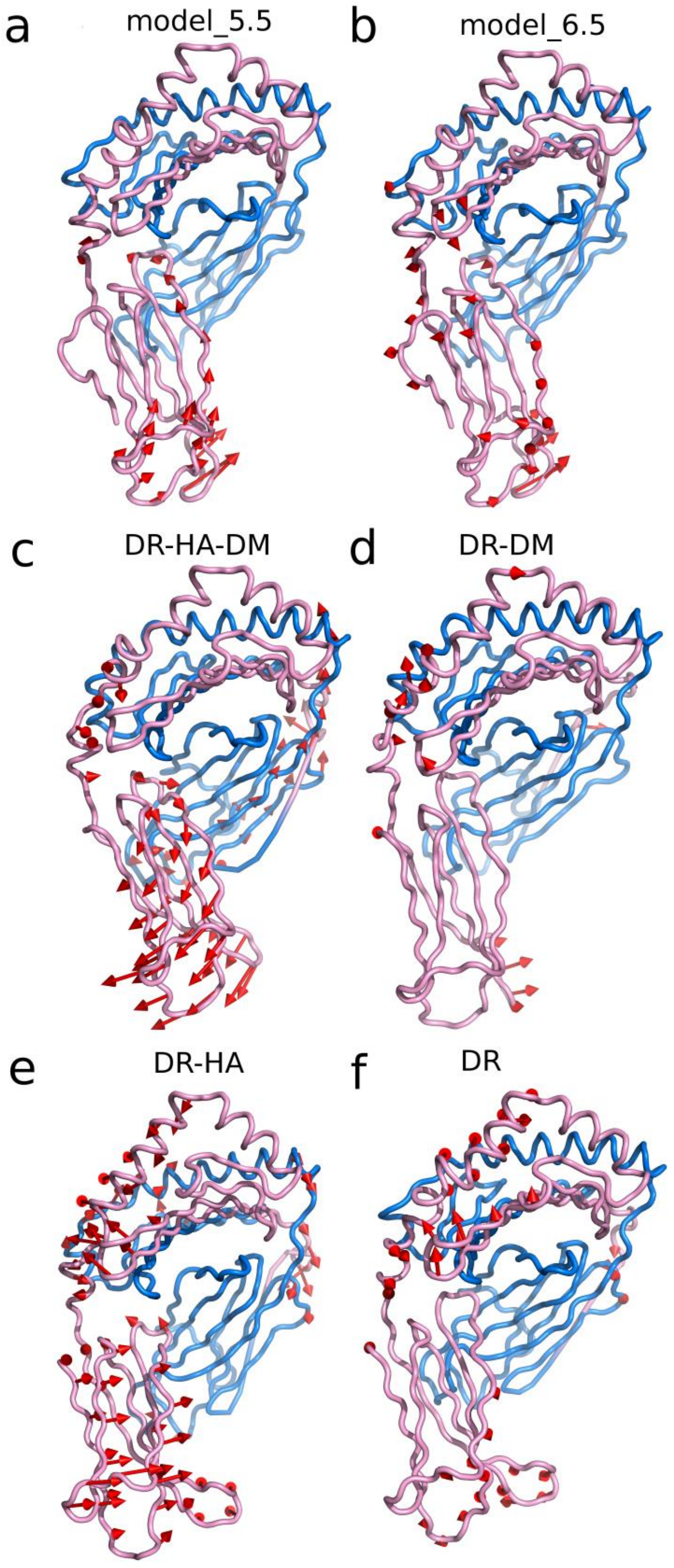


Figure 3.9: PC1 Porcupine plots of DR in the six systems. (a) Model_5.5, (b) model_6.5, (c) DR–HA–DM, (d) DR–DM, (e) DR–HA and (f) DR. The alpha chain is in blue and the beta chain is in pink. The red arrows depict the direction and amplitude of the motion.

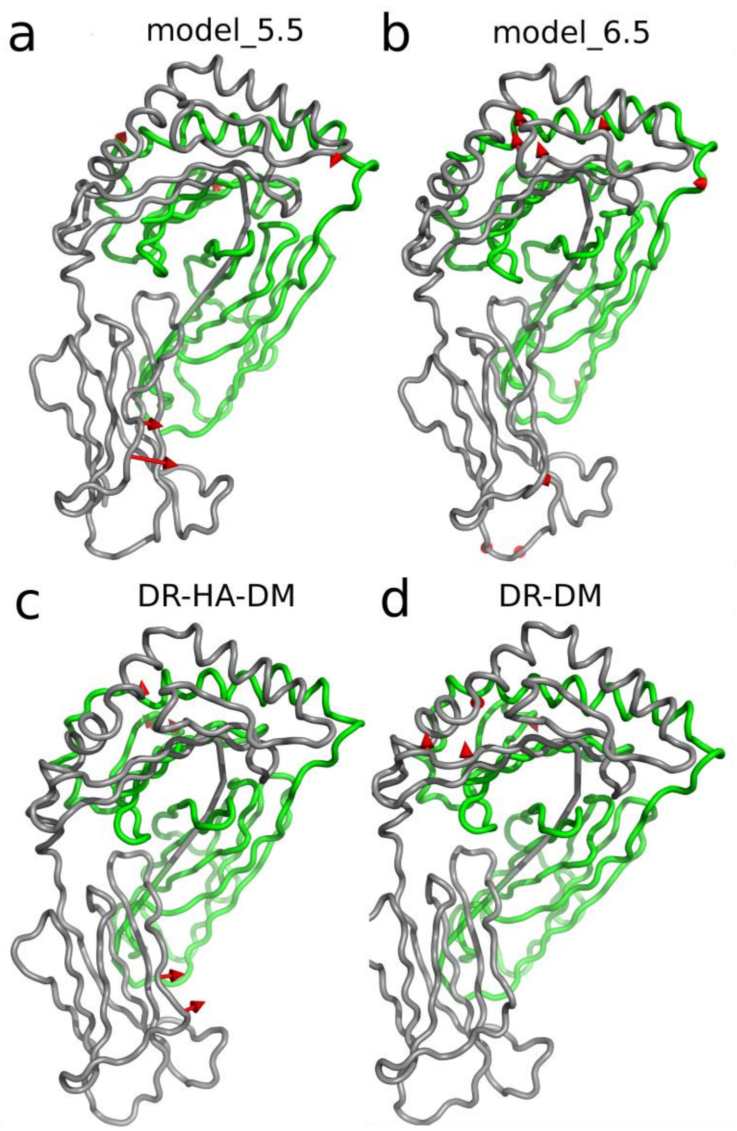


Figure 3.10: PC1 Porcupine plots of DM in the four systems. (a) Model_5.5, (b) model DR–HA–DM at pH 6.5, (c) DR–HA–DM, and (d) DR–DM. The alpha chain is in green and the beta chain is in gray. The red arrows depict the direction and amplitude of the motion.

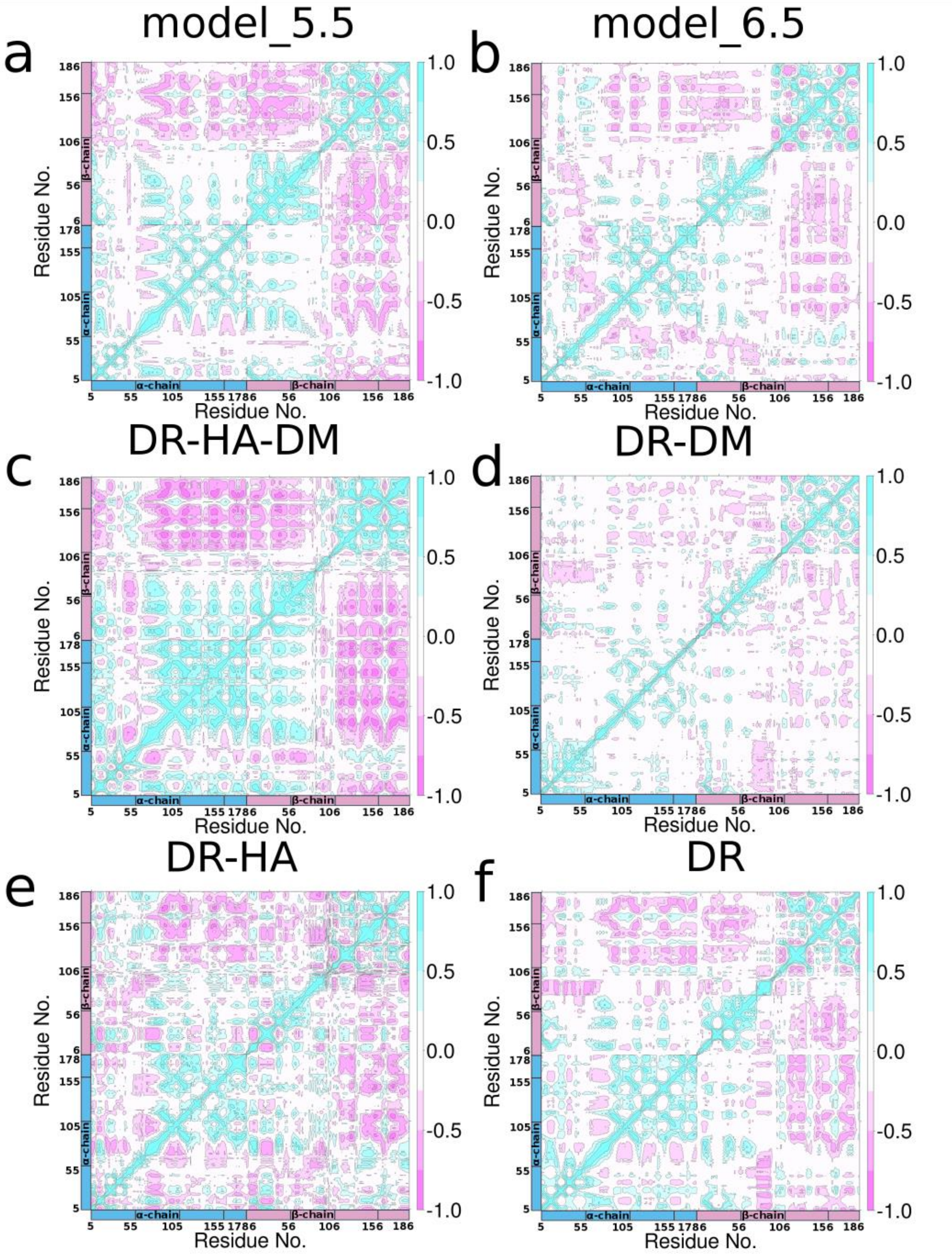


Figure 3.11: Dynamical cross validation matrix of DR in six systems. (a) Model_5.5, (b) model_6.5, (c) DR–HA–DM, (d) DR–DM, (e) DR–HA and (f) DR. The pink indicates negative correlation (anticorrelation), while the blue indicates positive correlation (correlation). The α and β chains along x and y axes are in blue and pink as in the Figure 3.1.

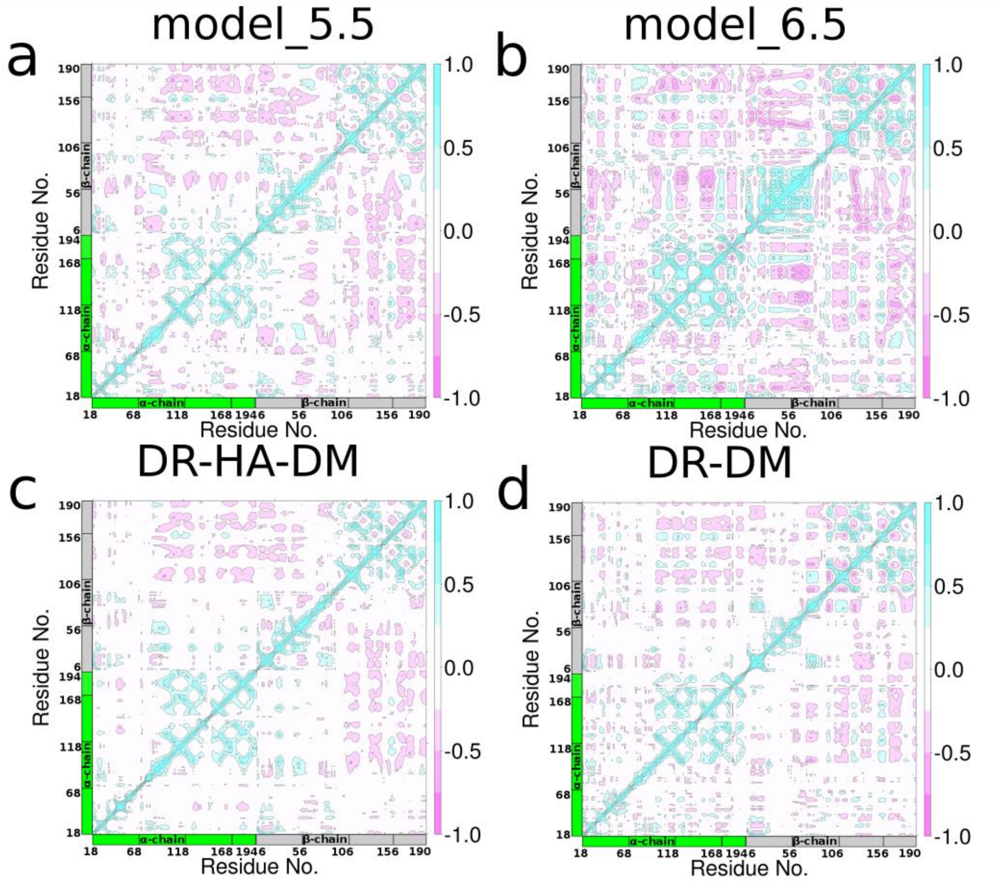


Figure 3.12: Dynamical cross validation matrix of DM in four systems. (a) Model_5.5, (b) model_6.5, (c) DR–HA–DM and (d) DR–DM. The pink indicates negative correlation (anticorrelation), while the blue indicates positive correlation (correlation). The α and β chains along x and y axes are in green and gray as in the Figure 3.1.

3.4.6 Dynamical cross-correlation

The dynamical cross-correlation matrix (DCCM) of the C^{α} atoms indicates complex correlations in DR (Figure 3.11), but not in DM (Figure 3.12). DR has both positive and negative correlations, while DM is characterized by negative correlations.

(a) DR correlation

In DR–HA–DM and model_5.5 systems, the positive and negative correlations are more extensive than in the other four systems (DR–DM, model_6.5, DR–HA and DR). In general, the β2 domain shows the anticorrelation (in pink) with other regions, namely, α1, α2 and β1, but correlation (in blue) with itself. The β1 correlates with α2 but anticorrelates with α1. The α1/α2 correlation is not clear.

(b) DM correlation

The model_6.5 has more negatively correlated motion than the other three systems (model_5.5, DR–HA–DM and DR–DM). Positive correlation is less frequently observed than the negative correlation.

3.4.7 Peptide Editing from DR by DM

Peptide editing from DR by DM was also investigated in the DR–HA–DM and DR–HA complexes where HA starts from P5. The DR–peptide interaction was examined by the number of hydrogen bonds between HA peptide and DR protein (Figure 3.13 and Table 3.3). The side chain-side chain hydrogen bond interactions do not favor the HA-DR interaction (less than 22% for each hydrogen bond, see the first two rows of Table 3.3). Most of the hydrogen bond interactions are from the main chain atoms of the HA peptide and the side chain atoms of the DR protein. During 1.5 μs (3 x 500 ns) MD simulations, the average number of hydrogen bonds has increased to 6.5 (DR–HA–DM) and 6.9 (DR–HA) in comparison to the crystal structure (five hydrogen bonds). Even though the peptide has not left the DR groove, in both the cases, the hydrogen bond number decreases for the last 50 ns of all triplicate simulations of the DR–HA system and one simulation of the DR–HA–DM system. The decrease suggests the possibility of the HA peptide leaving from the DR groove in both the DR–HA–DM and DR–HA complexes.

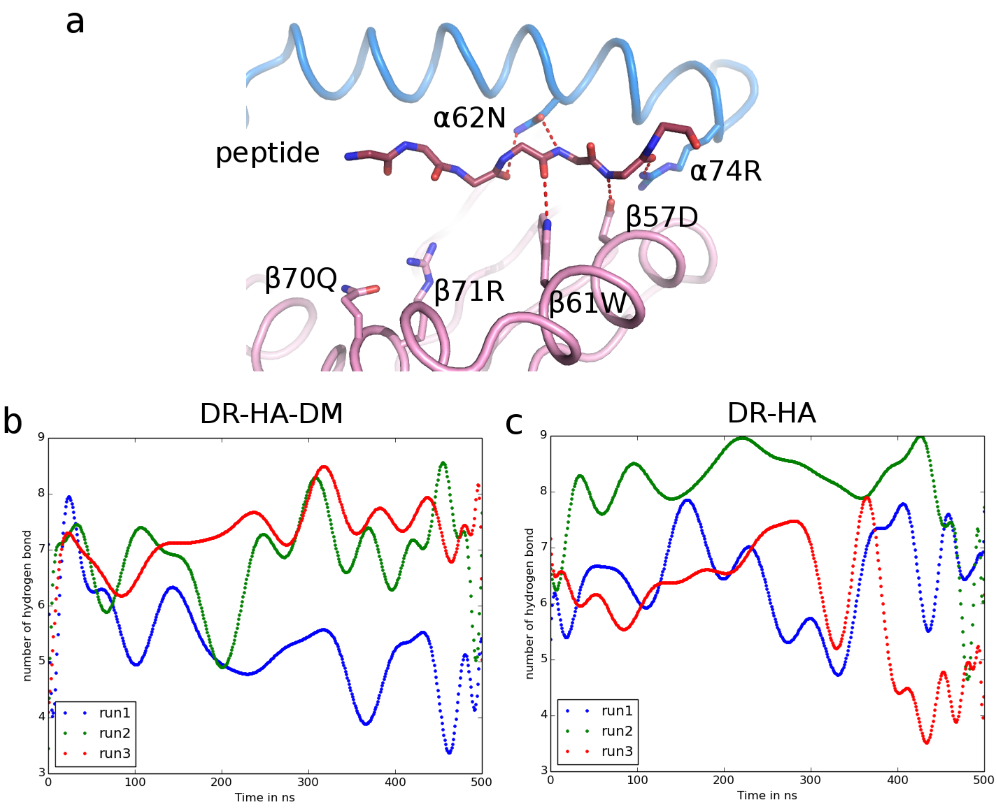


Figure 3.13: Number of hydrogen bond between DR and HA peptide during the MD simulations. (a) Location of residues that could make hydrogen bonds with HA, the hydrogen bond number in the (b) DR–HA–DM and (c) DR–HA complexes during triplicate 500 ns MD simulations.

Table 3.3: Residence times (in %) of hydrogen bond interactions between DR protein and HA peptide

Peptide residue	Atom name	Protein residue	Atom name	DR–HA (%)	DR–HA–DM (%)
P5	$O^{\delta1}$	β71	$N^{\eta1}$	21.9	17.3
P5	$O^{\delta1}$	β71	$N^{\eta2}$	13.7	7.5
P5	O	β71	$N^{\eta1}$	33.5	31.2
P5	O	β71	$N^{\eta2}$	48.1	41.5
P5	N	α62	$O^{\delta1}$	42.7	18.6
P7	O	α69	$N^{\delta2}$	90.6	89.4
P8	O	β61	$N^{\varepsilon1}$	81.1	95.3

P9	N	α69	$O^{\delta1}$	81.4	89.3
P10	O	α76	$N^{\eta1}$	0.0	77.7
P10	O	α76	$N^{\eta2}$	86.0	19.3
P10	N	β57	$O^{\delta1}$	93.9	39.7
P10	N	β57	$O^{\delta2}$	0.0	27.1

[*]only hydrogen bond interactions that occur more than 10% in at least one of the two systems were shown in this table. The residence time is the number of MD simulation snapshots in which the hydrogen bond is formed over the total number of MD simulation snapshots (in percentage).

3.4.8 Correlation of the Size of Peptide Binding Groove in Peptide-free DR in the Presence and Absence of DM

Previous studies[79] have shown that during 20 ns MD simulations the peptide-free MHCII (in the absence of DM), some snapshots depict the closing of the peptide binding groove. Experimental studies[95] showed that DM can stabilize the peptide-free DR. In this study, we used peptide-free DR as a control simulation in order to compare with the peptide-free DR–DM complex. The size of the peptide-binding groove in both systems was studied by analyzing the C^{α}-C^{α} distance of the periphery binding groove pairs; namely, αE55-βN82, αN62-βR71, and αN69-βW61[79] (Figure 3.14). In the absence of DM, the C^{α}-C^{α} distance between the two ends of the peptide binding groove either insignificantly increases (~3 Å) or decreases (~1 Å). The center of the groove, measured by the C^{α}-C^{α} distance of αN62-βR71 residues in one of the trajectories, decreases from ~16 Å to ~6 Å (Figure 3.14 and movie at http://cospi.iiserpune.ac.in/cospi/data/). The corresponding C^{α}-C^{α} distances in the DR–DM simulation was also investigated. The smallest distance between αN62 and βR71 resides in DR–DM simulation is 11.6 Å (Figure 3.14b), which is higher than in the case of the peptide-free DR simulations (6 Å).

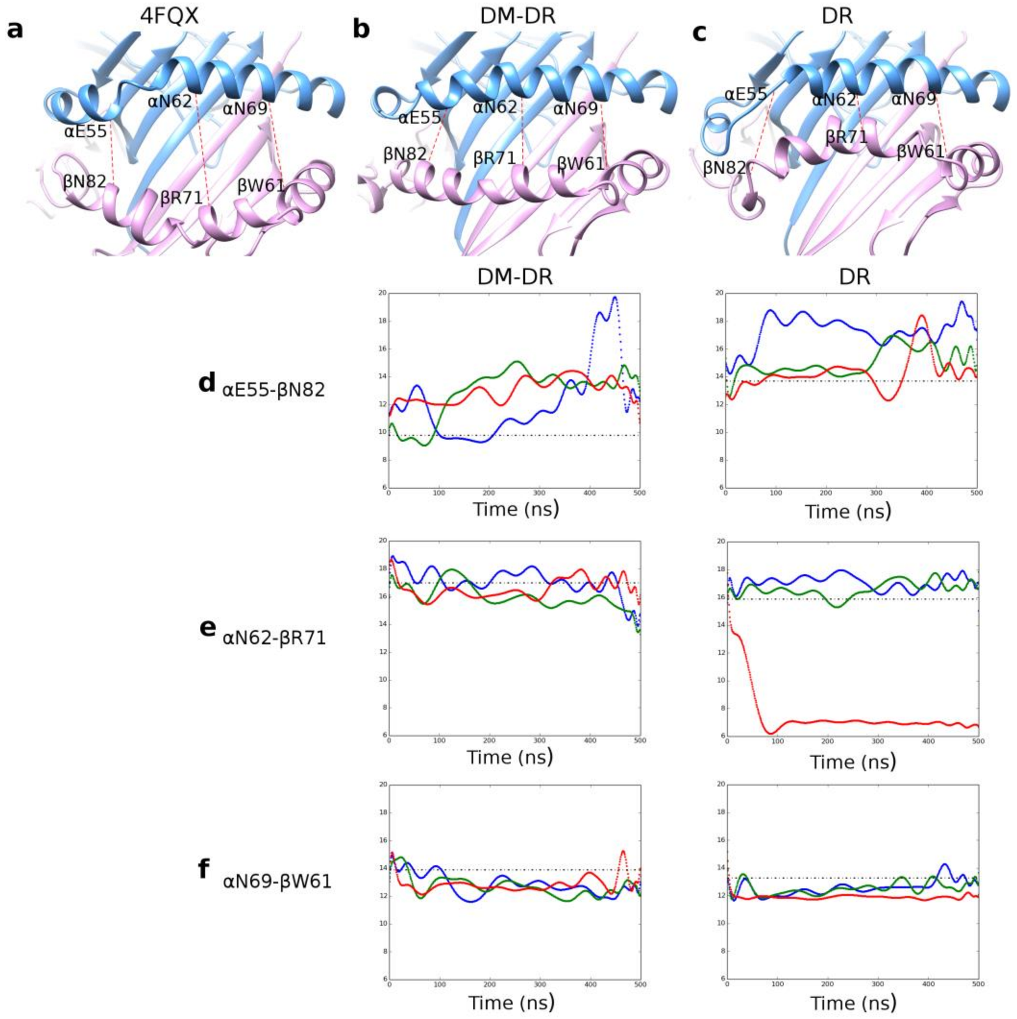


Figure 3.14: The conformational change of α-helices forming peptide binding groove. The change in (a) crystal DR–HA–DM (PDB ID: 4FQX), (b) DR–HA at 500 ns, and (c) DR at 500 ns. The α chain is in blue and the β chain is in pink. The C^{α}-C^{α} distances are highlighted in red dot. The C^{α}-C^{α} distance of (d) αE55-βN82, (e) αN62-βR71, and (f) αN69-βW61 in DR–DM (left) and DR (right) MD simulations. The y axis is the distance in Å; the x axis is the time scale of MD simulations in ns. Different color lines are for triplicate trajectories. The corresponding distances in crystal structure of the DR–HA–DM complex are shown in black dotted line.

3.4.9 Conformational Changes of DM with Peptide-bound DR and Peptide-free DR

To decipher the conformational changes in DM in the DR–DM and DR–HA–DM complexes, we analyzed the secondary structures of the regions from α64Q to α77E (Figure 3.15) because this region shows high fluctuations in RMSF (Figure 3.8) and PCA (Figure 3.10) analysis.

In all triplicate MD simulations, the secondary structure at the α69-75 region has a tendency of belonging to the coiled state in the DR–HA–DM complex more than in the DR–DM complex (Figure 3.16). The representative structures for DR–HA and DR–HA–DM complexes are in Figure 3.14b and c. The α69-75 region changes from helical conformation in the crystal structure of DR–HA–DM complex to coil conformation in MD simulations of DR–HA–DM. However, this conformational change from helix to coil is not observed in DR–DM simulations. The model_5.5 has a preference to form coil at residues from α69 to 71 only in the first run. In the region α72-75, the model_5.5 and model_6.5 prefer to adopt a coil conformation. Even though model_5.5, model_6.5 and DR–HA–DM complexes have the tendency of forming the coil conformation at α69-75 region, their localizations of the coil are different. These differences are shown by the C^{α}-C^{α} distances of α71-β75 residues (Figure 3.15, third column). In the DR–HA–DM complex, the tendency of forming coil results in widening of the DM groove by 4 Å in comparison with the crystal structure (movie at http://cospi.iiserpune.ac.in/cospi/data/)). Sometimes during 1.5 µs MD simulations that distance also increases in model_5.5, but decreases in model_6.5. The common feature between the complexes having α-helix to coil tendency at α69-75 region is that all these complexes are bound to the HA peptide (either the peptide starts from P1 or P5).

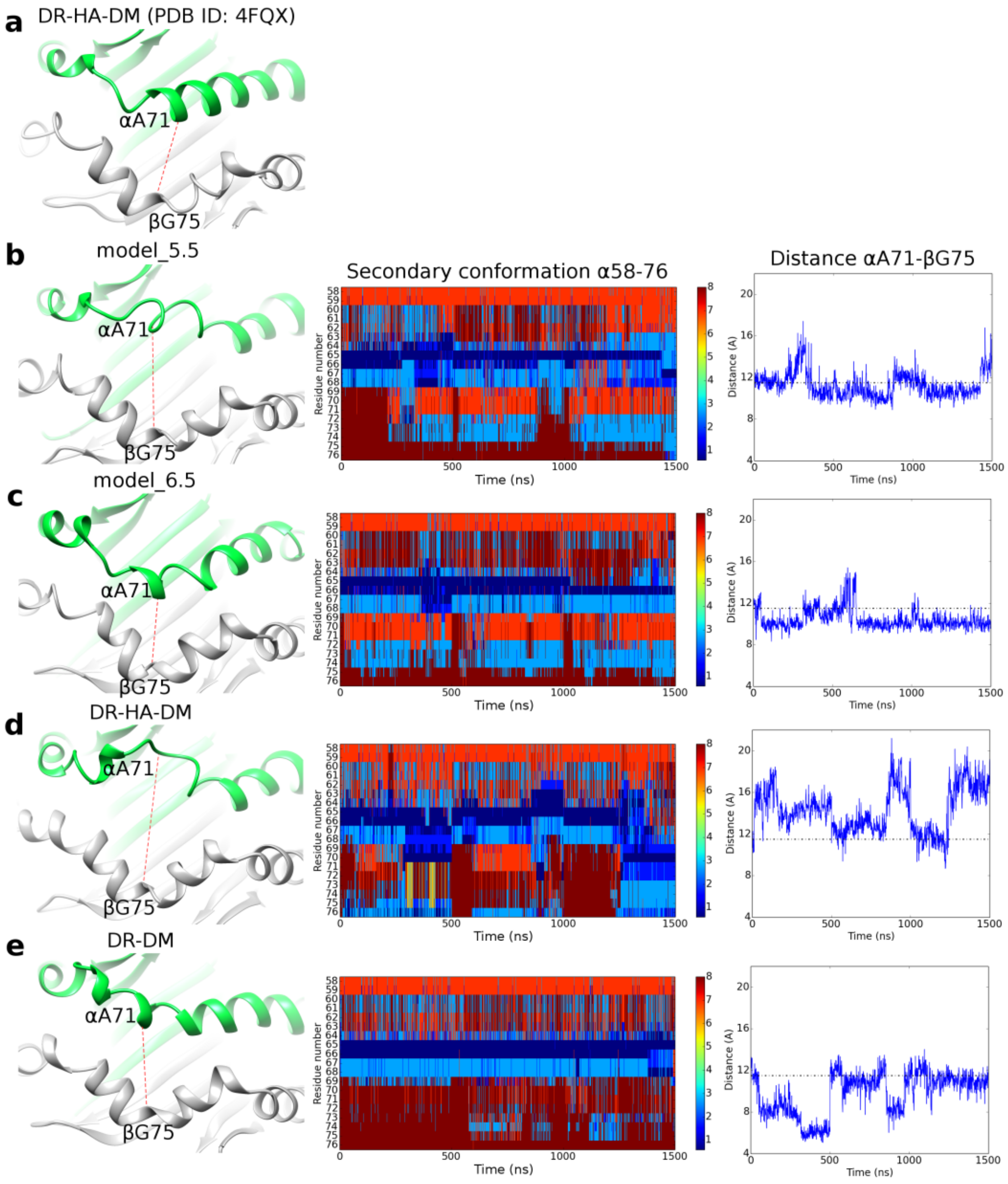


Figure 3.15: Conformation of the helical region in DM. (a) DR–HA–DM crystal structure (PDB ID: 4FQX), (b) model_5.5, (c) model_6.5, (d) DR–HA–DM, and (e) DR–DM. All the snapshots are at 500 ns. The α chain of DM is in green and the β chain is in gray. The second column is the secondary conformation from α58 to α76. The order of stability of secondary structures increases from dark blue (coil) to dark red (alpha helix). The third column shows the C^{α}-C^{α} distance between αA71 and βG75. The crystal distance is shown in black dotted line.

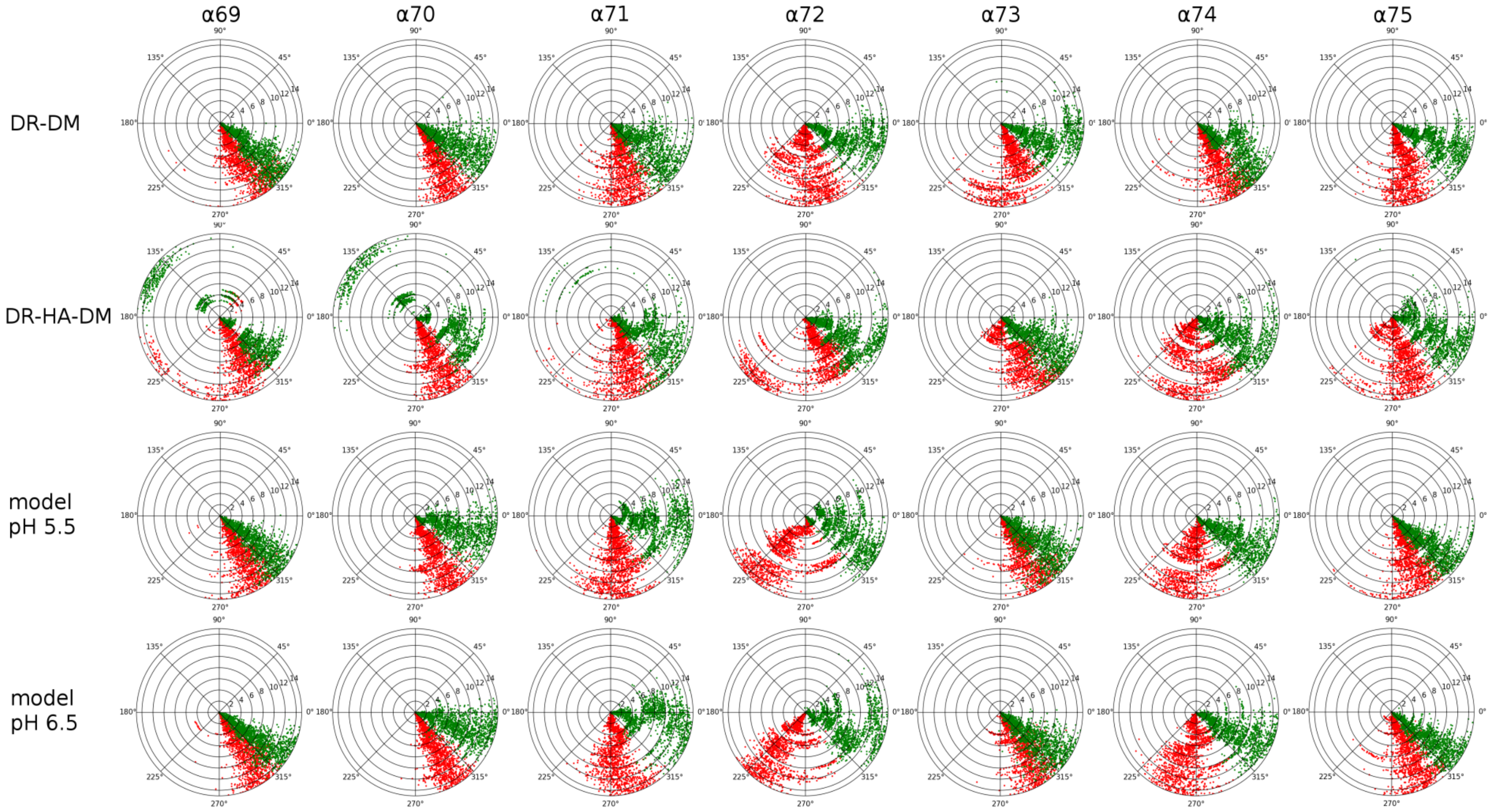


Figure 3.16: (ψ, ϕ) diheral angles residues α69-75 in DM. (a) DR–HA, (b) DR–HA–DM, (c) model_5.5 and model_6.5. The ψ, ϕ angles are in red and green, respectively. The radius is corresponding to 1.5 μs simulations. The pure α-helix region is (-54°, -45°).

3.5 Discussion

In this study, 1.5 µs MD simulations (3 x 500 ns) have been performed on six systems in order to address the following issues: (i) to determine the conformational changes in DR/DM upon interaction, (ii) to identify the important residues for such conformational changes, (iii) to reveal the effect of pH on DR–DM interaction, (iv) to decipher the mechanism of peptide release from DR, and (v) to discover the mechanism of stabilization of peptide free DR by DM. The trajectories were analyzed for RMSD, RMSF, PCA, DCCM, distance and hydrogen bond interactions in order to answer these questions.

3.5.1 Effect of pH on the DR-DM interaction

The RMSF analysis shows that the fluctuation of DR/DM in model_5.5 and model_6.5 are very similar. On the other hand, the PCA analysis shows that the β1, β2 domains (in DR) and β2 domain (in DM) have differences in the directions of the fluctuations. The distance and DSSP display different motions in model_5.5 and model_6.5 at the DM α69-75 regions. However, the connection between these differences to the DR-DM interaction is not clear and remained to be answered.

3.5.2 Mechanism of peptide editing from DR

The editing mechanism of HA peptide from DR was examined in the DR-HA-DM and DR-HA systems. None of the triplicate MD simulations showed the full removal of the peptide starting from P5 either in the presence or in the absence of DM. It is possible that the simulations were not long enough to generate the conformational states that would induce the peptide to move out.

3.5.3 Stabilization of peptide-free DR by DM

The question on how DM stabilizes DR was investigated in the peptide-free DR system in the absence and presence of the DM protein. In our study, we detected the closing of the peptide binding groove in one of the triplicate simulations of peptide-free DR system in the absence of DM. This closing was not observed in the presence of DM. Although previous studies have also shown this conformational change[79], this study shows that the closing is not a snapshot, but a stable conformation lasting more than 400 ns.

3.5.4 Conformational change upon DR-DM complex formation and important residues for the interaction.

The full conformational changes from the apo to the holo states of DR and DM were not observed during the sampling by the triplicate 500 ns simulations of model_5.5 and model_6.5. However, high fluctuation at the DR β2 domain and DM α1 domain for both models were seen during the 1.5 μs MD simulations. It could be possible that these domains are important for DR–DM interactions. This work only focused on the backbone conformational change, the side chain conformational flexibility should be paid more attention in the future work.

Currently, most DR–DM interaction studies have focused on the interactions involving the α1 and β1 domains[95-97]. Only a few random mutagenesis studies at this DR β2 domain, particularly at βD152N, βL184H, βS197N and βE187K showed reduction in the DM binding activity[61]. In our analysis, β187-197 residues were left out as they are near the C-terminal and obviously have high fluctuations. The residue β152 in the model_5.5, DR–HA–DM and DR–HA has fluctuations higher than 2 Å. In addition, experimental studies showed that some of the DR residues of the β2 domain, namely, βK98 and βR189, have different conformation in the peptide-bound and peptide-unbound conformations of DR[97]. In this study, we showed that the β2 domain in DR has a high fluctuation (more than 3 Å) in the presence of DM at DR–HA–DM and model_5.5 complexes, but not in other structures. DM β2 domain in DR–HA–DM and model_5.5 also tends to come closer to DR β2 domain. The PCA analysis also showed that the DR β2 domain fluctuated highly and correlated with the DM fluctuations. It could be possible that DR β2 domain could play important roles in the DR–DM interaction, for example, interacting with the β2 domain of DM to hold these two molecules together. The role of all residues of the β2 domain on MHCII function remains to be tested.

The DM α69-76 region that is in the DR–DM interaction has a conformational change during the 500 ns MD simulations. DSSP analysis showed that these residues have a tendency to move from helix to loop conformation in DR–HA–DM, model_5.5 and model_6.5 systems. All these three systems have the HA peptide bound. The conformational change in α69-75 residues in DM does not occur in the DR–DM complex. It could be possible that the conformational change in the α69-75 region affects the editing of HA peptide by some long-range induced interaction. This effect could be tested by mutagenesis experiment on the residues of that region. The secondary structure of DM in this region was broken from α-helix to coil conformation, and hence, we suggest the stabilization of the helix by covalent linkage with a hydrocarbon staple[98]. As the break in the secondary structure only occurs in the present of peptide, the stapled DM should have an effect on the DR-peptide interactions.

Chapter 4
Prediction of
Polyproline Type II Helix Receptors

The peptides that bind to MHCII have PPII conformation. As PPII is only the structural conformation, it could be possible that the PPII receptors should share their binding site specificity. This chapter tries to figure out these specificities and predicts whether a given protein could be the PPII receptors or not.

4.1 Background and Motivation

4.1.1 Definition and Properties of Polyproline II Helix (PPII)

Protein secondary structure is a specific local structural conformation that is classified and stabilized by the intramolecular hydrogen bond pattern of the component residues[94]. For example, α helices and β sheets are considered regular secondary structures, while random coils, loops and turns are not.

The PPII or polyproline II is another type of secondary structural conformation. Although proline residues are contained in many PPIIs, any residue can adopt this conformation, including residues with positive or negative charges, such as LYS, HIS, ARG, GLU, and ASP[99-104]. The (ϕ, ψ) backbone dihedral angles of PPII in the Ramachandran plot is roughly (-75°, +145°)[99,105]. PPII helices are left-handed and appear to have a three-fold rotational symmetry[106,107] (Figure 4.1). In comparison with α helices PPII is extended with a translation of 3.1 Å along the helical axis instead of 1.5 Å.

The number of residues per turn for α helix is 3.6 while for PPII is only 3.

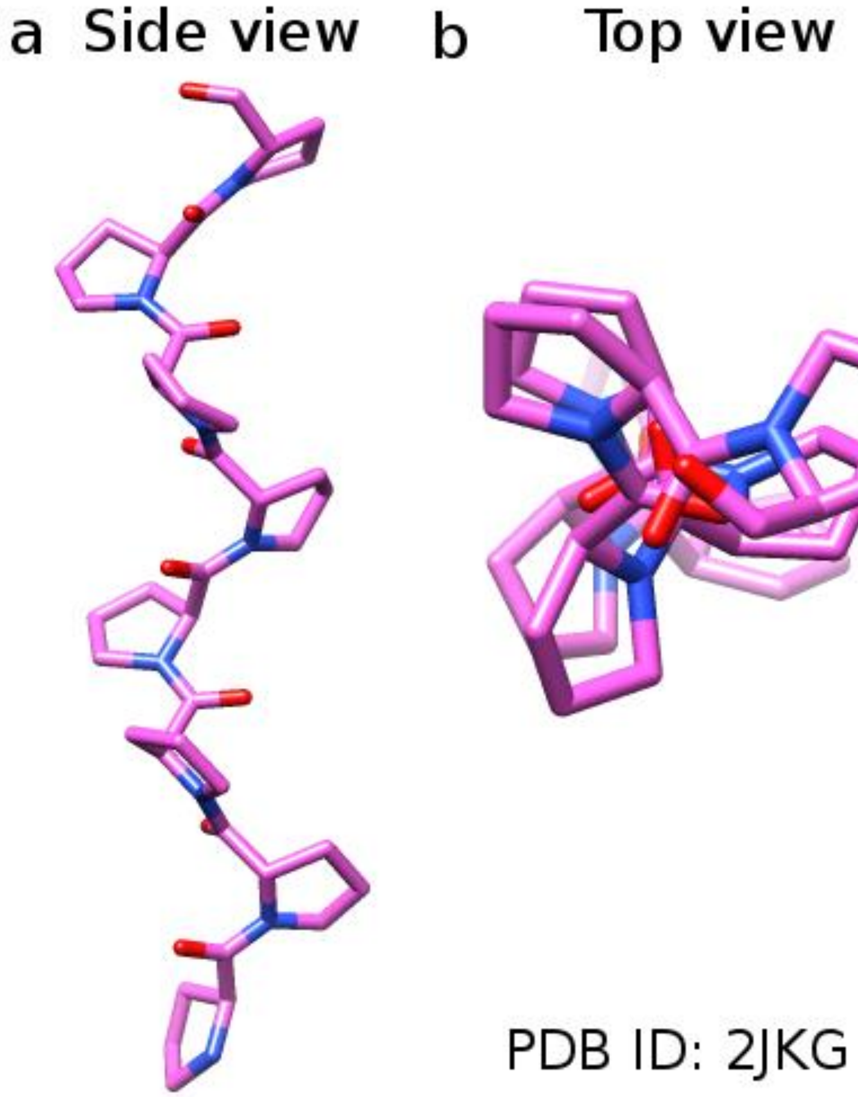


Figure 4.1: Conformation of a PPII peptide. The peptide was taken from the complex with profilin (PDB ID: 2JKG) in (a) side view and (b) top view.

4.1.2 Abundance of PPII

The PPII was thought to be a rare conformation and there is a lack of the PPII assignment in most of the commonly used secondary structure assignment methods. However, this conformation was recently shown to occur more frequently than expected. About 4% of amino acids in proteins adopts the PPII conformation with a length of three or more residues[108].

PPII helices were first found in fibrous proteins such as α-keratin and collagen[109,110]. Later, various globular proteins were shown to have this conformation[99,111]. PPIIs mediate inter-protein interactions[108,112]. This preference for PPII could be explained by the lack of intramolecular backbone hydrogen bond interaction in PPII helices. The backbone carbonyl and amide groups along the PPII helices are usually solvent exposed, so that PPII is free to make hydrogen bond with its receptor or solvent molecules. The advantage of such distinct chemical features is that it could mediate interactions even in the absence of high affinities. And hence, peptides with the PPII conformation have an ability to facilitate transient intermolecular interactions. Particularly, the PPII conformation is shown to frequently participate in protein-protein, protein-peptide or protein-nucleic acid interactions. These interactions are involved in signal transduction, transcription, antigen presenting, *etc.*[99]. There is a list of proteins (eight families) that are well-characterized to bind PPII, such as Src homology-3 (SH3) domain[113], WW domain, Enabled/VASP homology-1 (EVH1)[114], profilin[115], glycine-tyrosine-phenylalanine (GYF)[116], myeloid, Nervy, and DEAF-1 domain (MYND)[117], major histocompatibility complexes (MHC) class I[118] and class II[52].

4.1.3 Significance of Predicting the Polyproline type II Helix Receptors

The predictions of the PPII conformation have been extensively studied[99,101,106,108]. The interactions of PPII and individual PPII-binding protein families have also been investigated[114,116]. The prediction of PPII peptides that bind to particular protein, such as SH3[119] or MHCI or MHCII[120] has been examined. Programs for predicting protein-peptide interactions, such as PepSite[121], FlexPepDock[122], GalaxyPepDock[123], and CABS-dock[124] could also be used to predict protein-PPII interactions. However, none of these programs use information from the known PPII-binding proteins to understand the common requirements for the PPII receptor proteins. It is likely that the PPII-binding proteins, or the PPII receptors, could share their geometrical and biophysical features to interact with the peptides having the same conformation. And hence, our first aim is to characterize common features of the PPII-binding sites. These features were extracted from the known PPII-binding proteins. The features were then used to identify the PPII-binding site in a query protein. To do that, we compared the query protein with templates from known PPII-binding proteins using the CLICK structural alignment program[125,126]. Only the structural hits that satisfied the binding criteria were chosen. Support vector machine (SVM) classification with different kernels[127] was applied to distinguish binding and non-binding hits. The hits which have the highest absolute SVM score were used as final identification of the binding site. The protein-PPII complexes were then built using a Monte Carlo refinement simulation. Finally, we also applied our protocol to a protein dataset of more than 17, 000 structures to find the new PPII-binding proteins. The detail information of experimental procedures, results and discussion are discussed in the following sections.

4.2 Methodology

4.2.1 Dataset for Identifying Important Requirements to Bind PPII

The homologous structures of all eight known PPII-binding proteins were searched using the PSI-BLAST[128,129] program. The query sequences from SH3 domain (PDB ID: 1CKA, chain A), WW domain (PDB ID: 1JMQ, chain A), GYF domain (PDB ID: 1L2Z, chain A), profilin (PDB ID: 2V8F, chain B), EVH1 domain (PDB ID: 4WSF, chain A), MYND domain (PDB ID: 2ODD, chain A), MHCI (PDB ID: 5C0D, chain A) and MHCII (PDB ID: 3PDO, chain B) were used. The queries were used to search over the entire PDB database. Five iterations were performed using an e-value cutoff of $1e^{-5}$. The sequence identity cutoff was set to 70%. Only those proteins that were bound to a PPII peptide were chosen. For NMR structures the models that had the highest number hydrogen bond interactions between PPII and protein were taken. There were 44 homologous structures with bound-peptides for the eight families discussed above (EVH1 (5), GYF (2), MHCI (3), MHCII (3), MYND (1), Profilin (2), SH3 (22) and WW (6)).

4.2.2 Features to Characterize the PPII Binding Site

Common requirements for the PPII-binding site were learned from the 44 PPII-receptor structures. The following features were considered, namely:

(1) Number of hydrogen bond was counted where the hydrogen bond was defined as in chapter 4.2.5.2. Different types of hydrogen bond interaction between main chain (MC) and side chain (SC) atoms were analyzed, including

MC-MC, MC-SC, SC-MC and SC-SC, where the first atom name was from PPII and the second was from the receptor.

(2) Depth was defined and calculated as in chapter 5.2. Both atomic and residue depth values were investigated.

(3) Sequence entropy or conservation of each position was quantified based on multiple sequence alignment from the PSI-BLAST[128,129]. Two iterations were performed using a cutoff of $1e^{-4}$ for e-value. The absolute entropy was calculated as the Jensen-Shannon divergence formula as follows:

$$J_i = \frac{1}{2}\sum_{a=1}^{20}\left(f_{ia}^{obs}\log\left(\frac{f_{ia}^{obs}}{f_a^{exp}}\right) + f_a^{exp}\log\left(\frac{f_a^{exp}}{f_{ia}^{obs}}\right)\right) \quad (4.1)$$

where f_a^{exp}, f_{ia}^{obs} are the expected and observed frequencies of a residue type a at a position i.

We used the relative entropy, which was calculated as follows:

$$R_i = \frac{J_i - \min(J)}{\max(J) - \min(J)} \quad (4.2)$$

where J is the entropy value of all positions

4.2.3 Structural Alignments

All the structural alignments were done using the CLICK structural alignment program[125,126]. This program makes pair-wise alignment between two PDB-format structures without topology dependence.

4.2.4 Constructing the RMSD-SVM Models.

A vector of 11 features, most of which were RMSDs, was constructed. The features were:

Feature 1. RMSD of representative atoms, namely $N^{\varepsilon 1}$, C^{α}, $C^{\zeta 3}$ (from Trp) and NX, CX (from donor) in CLICK alignments, where NX is either N or O atoms from the side chain atoms of the residue, that donates its adjacent proton in a hydrogen bond interaction, and CX is the $(i+2)^{th}$ atom where the NX atom is in the i^{th} position;

Feature 2. Number of matched atoms from the CLICK alignment;

Feature 3. Relative entropy as calculated in section 4.2.2;

Feature 4. RMSD of all atoms in Trp residues;

Feature 5. RMSD of C^{α} atom in Trp residues;

Feature 6. RMSD of C^{α}, $C^{\delta 2}$, $C^{\zeta 3}$ atoms in Trp residues;

Feature 7. RMSD of $N^{\varepsilon 1}$ atom in Trp residues;

Feature 8. RMSD of NX atom in Donor residues;

Feature 9. RMSD of features 7 and 8;

Feature 10. Summary of features 7 and 8;

Feature 11. RMSD of features 5, 7 and 8.

All the possible combinations of these 11 features (2047 combinations for each kernel style (see below) and penalty parameter C) were trained/tested on the dataset of 44 structures.

The RMSD of the closest oxygen atoms in CO groups to $N^{\varepsilon 1}$ and NX atoms between template and target structures (or $\mathrm{RMSD}_{\mathrm{CO}}$) was chosen as identification for a binding and a non-binding binary classification. If $\mathrm{RMSD}_{\mathrm{CO}} < 4$ Å, the variable Y was set to 1 (binding), otherwise, it was set to -1 (non-binding).

The RMSD matrix was then trained by SVM implemented in the scikit-learn package[130].

Three different kernel types, including radial basis function (rbf), linear and polynomial kernels were used in the SVM. The penalty parameter C of the error term in SVM was set to 0.25, 0.5 and 0.75 in different trials. Nine different combinations of kernel types and penalty parameter were applied on the three different datasets (Table 4.1). The class weight for SVM was set as "balanced" to reduce the bias in the binding/non-binding frequency of the input data.

As the dataset was small, a leave-one-family-out (LOFO) cross-validation was applied in the RMSD-SVM models. LOFO means that all the structures belonging to one families were completely left out and used as a testing set. Only the templates from nonhomologous structures were used. The RMSD-SVM models were trained on the remaining data of other families and then tested on the testing set. For each structure, the alignment that had the highest SVM absolute score was chosen as a final prediction. The output from all the testing set then combined and reported.

We also used nonhomologous and homologous dataset where all the 44 structures were trained and tested against them. On the nonhomologous dataset, only nonhomologous templates were used for structural alignment, while both homologous (without using itself as a template) and nonhomologous templates were used on the homologous set.

4.2.5 Modeling the PPII Peptide into the Predicted Location of the Query Protein

Monte Carlo simulations were used to build the PPII in the predicted location on the receptor (query) protein. The PPII from the template of the best alignment in section 4.2.4 was used in the model building. This PPII was transferred from the CLICK superimpose structure onto the query protein. In this study, we only built the main chain of the peptide but ignored the side chain atoms. The Monte Carlo simulations were carried out as follows:

4.2.5.1 Monte Carlo Move Set

From any given PPII-protein conformation, the following rigid body translation and rotation were performed on all PPII peptides coordinates

a. Translation was performed along a random direction vector passing through the mean coordinates of the PPII with random amplitude ranging between -5 Å and +5 Å.

b. Rotation was performed by a random angle ranging between -20° to 20°, around a randomly chosen direction vector passing through the mean coordinates of the PPII.

4.2.5.2 Energy Calculations

a. Number of Hydrogen Bonds Score ($NHBS$) was calculated as follows:

$$NHBS = hb^2 \quad (4.3)$$

where hb is the total number of hydrogen bonds.

Square of the number of hydrogen bonds was used to harmonically weight the score. A hydrogen bond is assumed to form when the acceptor-donor distance is less than 3.5 Å and the donor-acceptor-acceptor_antecedent angle is greater than 100°. NHBS score maximizes the total number of hydrogen bonds at any given point of the simulation.

b. Restrained Hydrogen Bonds Score ($RHBS$) is the sum of distance score ($D(x)$) and angle score ($A(\theta)$) calculated as follows.

$$D(x) = \begin{matrix} 6400.(x-2.5).(x-4.2) & \forall\, x < 2.5 \; or \; x > 4.2 \\ 0 & \forall\, 2.5 \le x \le 4.2 \end{matrix} \quad (4.4)$$

where x is the distance between donor and acceptor. A factor of 6400 was multiplied to give more weight to distance score.

$$A(\theta) = \begin{matrix} (\theta-80).(\theta-180) & \forall\; \theta < 80 \\ 0 & \forall\, 80 \le \theta \le 180 \end{matrix} \quad (4.5)$$

where θ is the angle between donor, acceptor and acceptor antecedent. This angle has a maximum at 180° when three atoms, namely donor, acceptor and acceptor antecedent are in a line.

$$RHBS = D(x) + A(\theta) \quad (4.6)$$

$RHBS$ ensures the formation of hydrogen bond between TRP or Donor atoms of receptors and PPII peptide.

c. Clash Score (CS): Any two atoms within 2.8 Å distance were considered clashing. Depending upon the atom type, these clashes were categorized as main chain-main chain clash ($MCMC$), main chain-side chain clash ($MCSC$) or side chain-side chain clash ($SCSC$). Clash score was calculated as follows:

$$CS = 100.MCMC^2 + 5.MCSC^2 + 5.SCSC \quad (4.7)$$

Square terms were used for $MCMC$ and $MCSC$ to harmonically increase the score compared to the $SCSC$ clashes that could be tolerated up to some extent. The scaling factors for each term were chosen approximately such that their scores satisfied the following inequalities:

$$MCMCscore \gg MCSCscore \gg SCSCscore \quad (4.8)$$

d. Pseudo van der Waals Score ($PVWS$) was the total number of atom pairs whose distances were in the interval of 2.8 to 5.0 Å. This score was used to maximize the van der Waals energy term by increasing the density of protein atoms around PPII.

e. Total energy (E) of the system at any given point was calculated as follows:

$$E = -NHBS + RHBS + CS - PVWS \quad (4.9)$$

4.2.5.3 Temperature and Gas Constant

Temperature of the simulation was kept constant at 400K (T) and Gas constant (R) as 2.10^{-3} kcal/(mol.degree)

4.2.5.4 Selection/Rejection Criteria.

Metropolis criteria were used with a P probability,

$$P = \exp(-beta.\Delta E) \quad (4.10)$$

$$\text{where } beta = \frac{1}{R.T}$$

4.2.5.5 Number of Independent Runs

Total 40 independent runs (initialized by different random seeds) of Monte Carlo simulations were performed with 4000 Monte Carlo steps in each simulation. Each independent run leads to a unique model as none of the final models when compared to one another has an RMSD value equal to 0.

4.2.5.6 Clustering the Models

All 40 models were clustered into two sets using scipy hierarchical clustering “fcluster” module with “maxclust” criterion. The linkage matrix was created with “average” method. The model that had minimum RMSD with other models in the largest cluster was chosen as the representative model.

All scripts were written in Python

4.2.6 Dataset for Searching New PPII-binding Proteins

The non-redundant dataset for homologous searching included structures determined by both X-ray and NMR methods. This dataset was taken from PISCES database[131,132] with a pairwise sequence identity cutoff of 30%. Only protein structures with length between 40 and 500 amino acids were chosen. The dataset contained 17005 protein chains.

4.3 Results

4.3.1 Features to Characterize the PPII-binding Site

4.3.1.1 Hydrogen Bond Number

We analyzed the numbers of hydrogen bond between PPII and its receptor for 44 structures (Figure 4.2). The MC-MC hydrogen bond interactions are not present in the interactions between PPII and receptor. At the least, two MC-SC (in blue, the MC atom for this case is the carbonyl Oxygen) hydrogen bonds are observed in each structure of the eight known PPII-binding families. This trend in the MC-SC type is expected because each family in those eight PPII-binding families could bind different sequence PPII peptides. One important observation is that in all 44 structures one of the receptor residues making the MC-SC hydrogen bond with PPII is always Trp. These Trp residues are also conserved in each family (see the entropy analysis in the subsection 4.3.1.4). And hence, in the prediction we required the PPII-binding site should have at least two residues that could make the side chain hydrogen bonds. One of these two residues should be Trp, and the donor side chain atoms of those residues should not have any intramolecular hydrogen bond interaction.

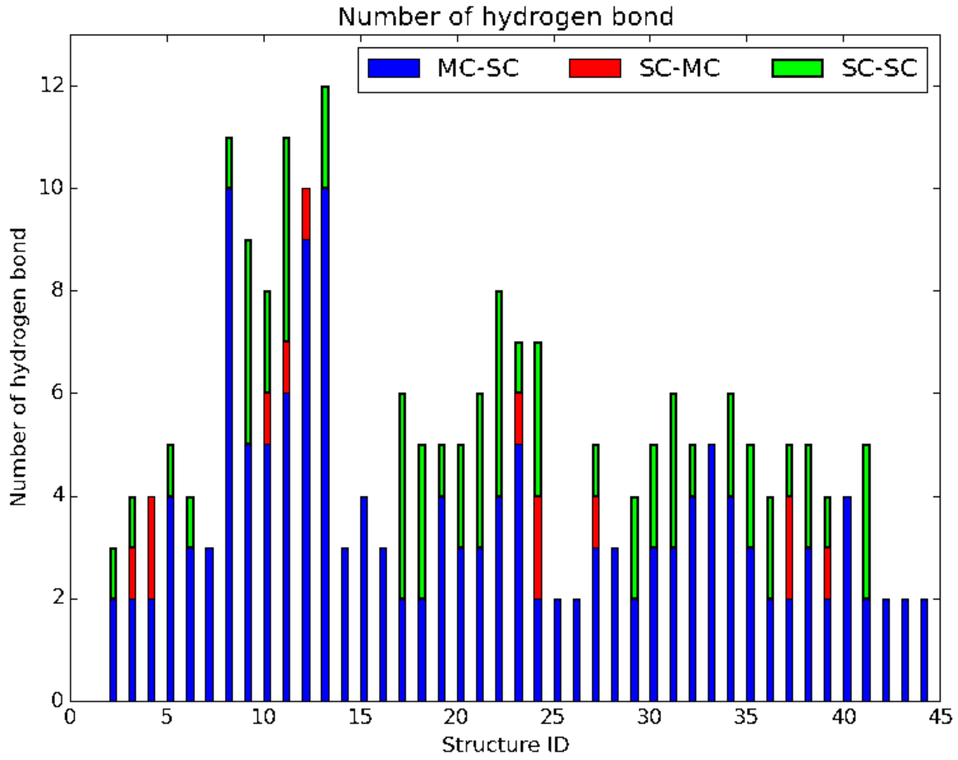


Figure 4.2: The number of hydrogen bond interactions between PPII and the receptor for the 44 structures. The MC-SC, SC-MC and SC-SC hydrogen bond are in blue, red and green, respectively. The structure ID is the same as Table 4.2. The family structure ID is following EVH1 – (1-5), GYF – (6-7), MHCI – (8-10), MHCII – (11-13), MYND – (14), Profilin – (15-16), SH3 – (17-38) and WW – (39-44).

4.3.1.2 Depth Values of Hydrogen-bond-making Residues in the PPII-binding Site

We analyzed both the side chain atomic and residue depth for Trp and donor residues that make hydrogen bond interactions with PPII peptide. Trp residue depth (grey) values range from 3.72 to 5.15 Å, while the Trp $N^{\varepsilon 1}$ atomic depth (blue) has a smaller interval from 2.96 to 3.55 Å (Figure 4.3). The residue depth values (red) of donor residues range from 3.25 to 6.36 Å, while the atomic depth values (green) range from 2.89 to 4.37 Å. All the $N^{\varepsilon 1}$ atomic depth values of Trp residues in the PPII-binding sites are below 3.60 Å. The NX atomic depth values of donor residues are below 4.00 Å in almost all the structures in the eight PPII-binding families with the exception of 4 cases in the MHCI and MHCII families. The depth values imply that both Trp and donor residues should be on the protein surfaces. This finding is consistent with the fact that in order to make hydrogen bond interactions with PPII, Trp should be exposed. And hence, we choose atomic depth threshold at 3.60 Å and 4.50 Å for $N^{\varepsilon 1}$ atom of Trp and NX atom of other donor residues.

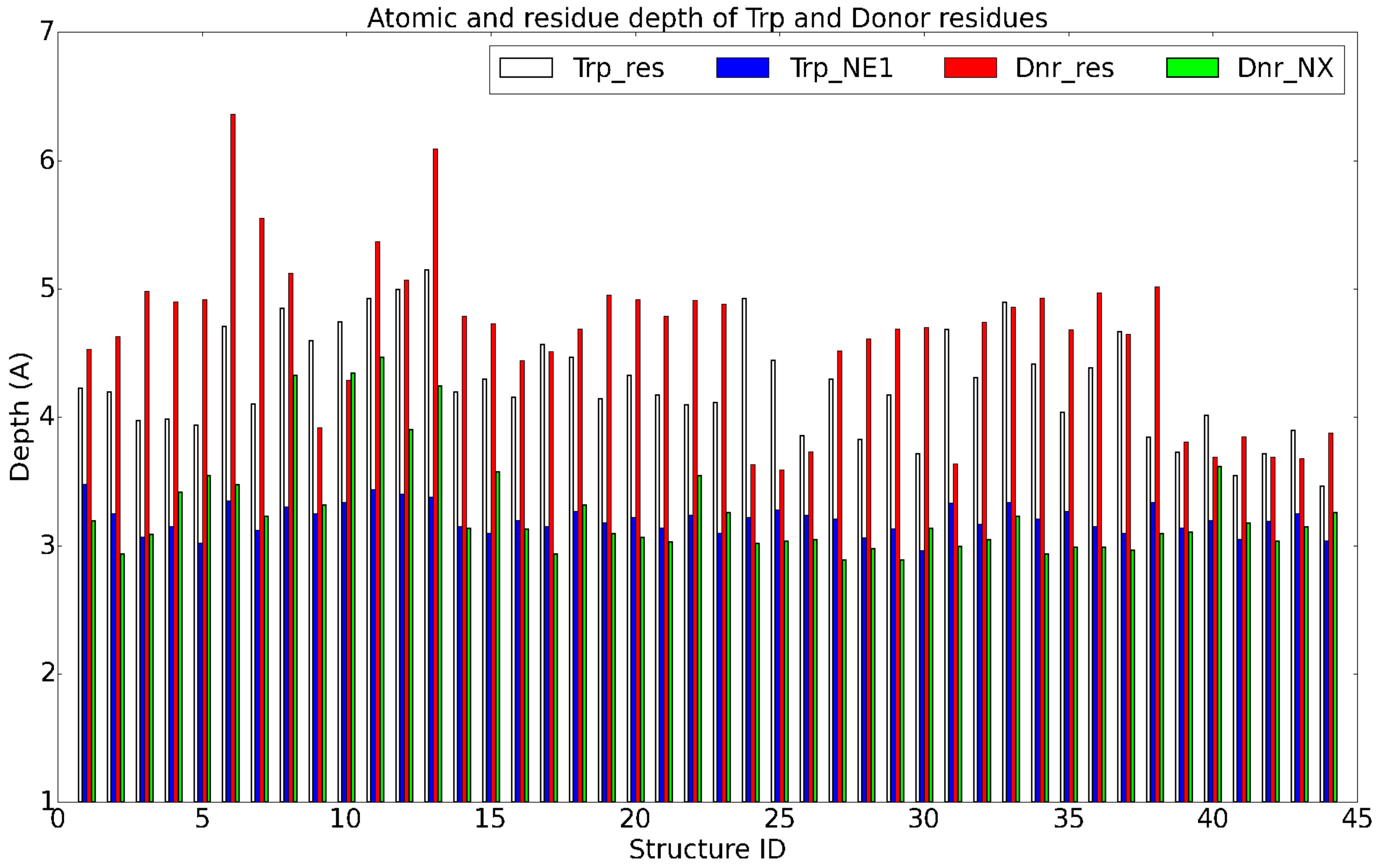


Figure 4.3: The atomic and residue depth of Trp and Donor residues for 44 structures. The $N^{\varepsilon 1}$-Trp depth is in grey, residue Trp depth in blue, NX-Donor depth is in red and residue Donor is in green. The structure ID is the same as Table 4.2. The family structure ID is following EVH1 – (1-5), GYF – (6-7), MHCI – (8-10), MHCII – (11-13), MYND – (14), Profilin – (15-16), SH3 – (17-38) and WW – (39-44).

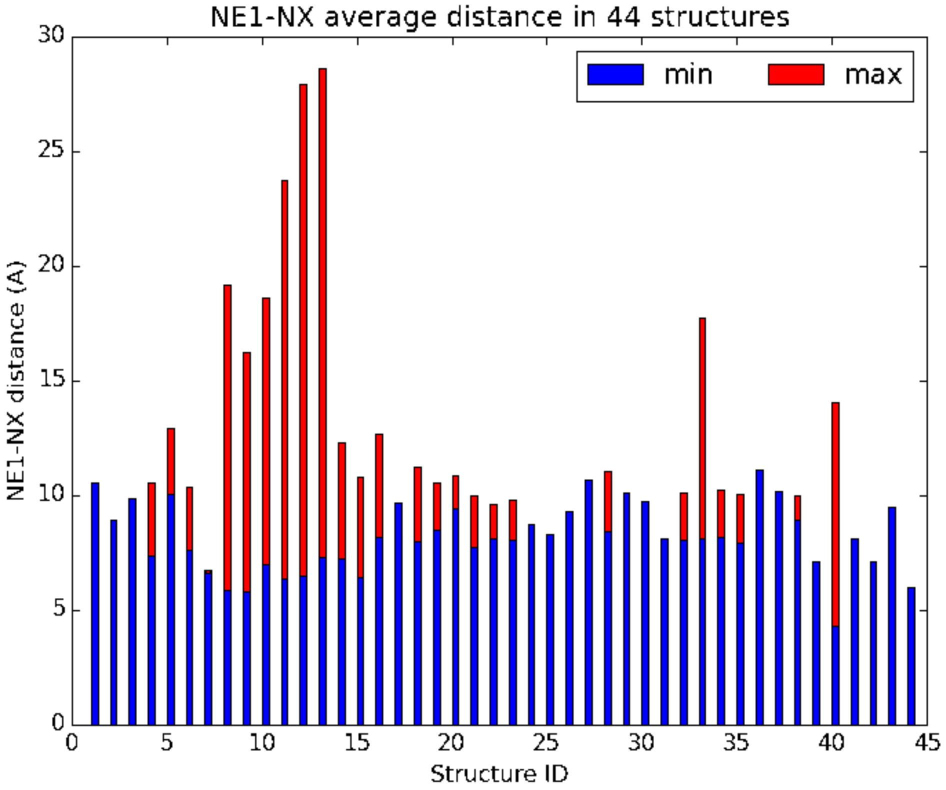


Figure 4.4: The N$^{\varepsilon1}$-NX distance in Å for 44 structures. The minimum and maximum distances are in blue and red, respectively. The structure ID is the same as Table 4.2. The family structure ID is following EVH1 – (1-5), GYF – (6-7), MHCI – (8-10), MHCII – (11-13), MYND – (14), Profilin – (15-16), SH3 – (17-38) and WW – (39-44).

4.3.1.3 Distance between Trp and Other Donor Residues.

The distances between N$^{\varepsilon1}$ and NX atoms from the hydrogen-bond-making Trp residues and their neighbor donor residues, respectively were also investigated (Figure 4.4). It is clearly demonstrated that six out of eight families, except MHCI and MHCII have similar trend for the distance between N$^{\varepsilon1}$ and NX atoms. However, all the eight families have the Trp-Donor pairs whose distances are below 12 Å. It means that it is possible to identify crucial residue pair in the interaction site using non-homologous structures. A cutoff threshold of N$^{\varepsilon1}$ and NX distance of 12 Å was applied in the prediction protocol.

4.3.1.4 Entropy Conservation

The relative entropy (formula 4.2, page 72) of the hydrogen bond making Trp is calculated using Jensen-Shannon divergence (Figure 4.5). The minimal entropy value is 0.78, while the maximal entropy value is 1. This data show that the hydrogen bond-making Trp residues should be highly conserved. And hence, we used a cutoff of entropy of 0.7 for predicting the PPII-binding site.

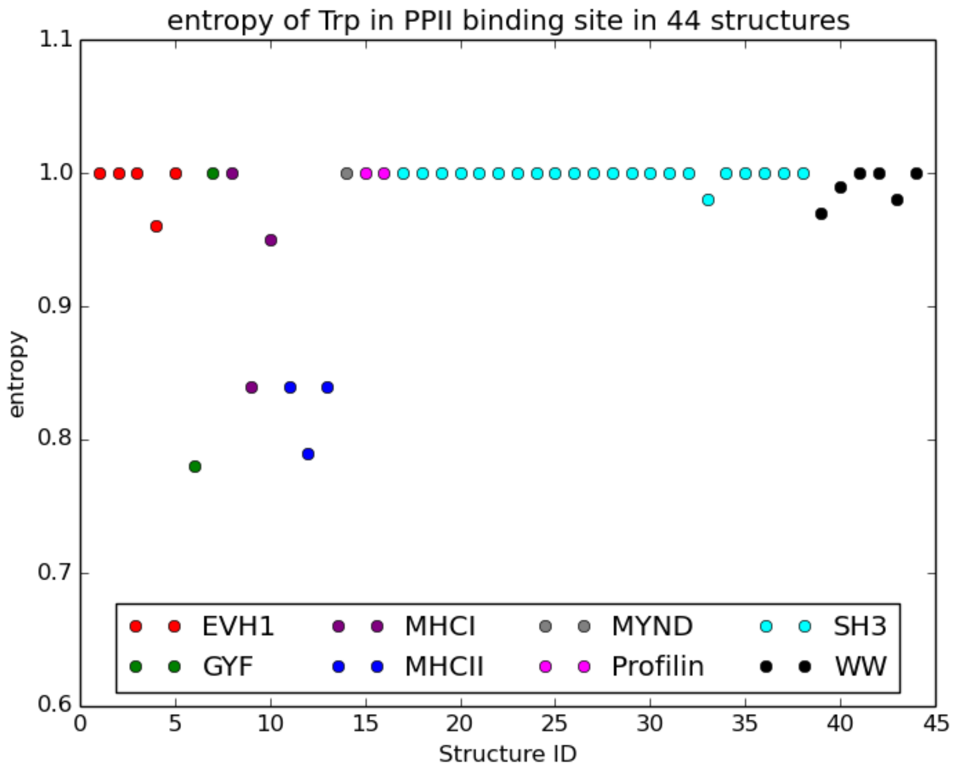


Figure 4.5: Entropy values of the PPII-binding Trp residues in 44 structures. The color is according to the family, particularly, EVH1 – red, GYF – green, MHCI – purple, MHCII – blue, MYND – grey, Profilin – pink, SH3 – cyan and WW – black. The structure ID is the same as in Table 4.2.

4.3.2 Prediction Accuracy

All the requirements for the PPII-binding site learning from previous sections were used for predicting the PPII-binding site. Even though the number of Trp-Donor pairs that satisfy the PPII-binding criteria is smaller than the total number of Trp-Donor pairs in the known PPII binding site, not all those pairs are actually PPII-binding sites. And hence, we used support vector machine to differentiate the actual binding sites.

To identify the PPII-binding region, we first located all the Trp residues and its neighbor donor residues that satisfy the requirement for hydrogen bond interactions, depth value, conservation entropy, and the $N^{\varepsilon 1}$-NX distance. Then these structures were aligned with all templates of Trp-Donor residues using the CLICK program. The representative atoms are $N^{\varepsilon 1}$ (Trp), C^{α} (Trp), $C^{\zeta 3}$ (Trp), NX (donor residues) and CX (donor residues). Only the alignments that have number of aligned atoms equal to 4 and CLICK RMSD value lower than 0.6 or number of aligned atoms higher than 4 and CLICK RMSD value lower than 1 were chosen. The RMSD-SVM models for three different datasets (LOFO, Nonhomolog and Homolog), kernel, and penalty constant C as described in the method section, were constructed. The average RMSD values between the transferred PPII (taken from the PPII of the template structure in the CLICK alignment) and native PPII (taken from crystal structure of the target protein) were calculated (Table 4.1). In comparison of the three datasets (LOFO, Nonhomolog, and Homolog), the average RMSD values on LOFO dataset are highest, while those values on the Homolog dataset are lowest. The lowest RMSD values (3.80 Å) for the LOFO dataset is when C was equal to 0.5 and the rbf kernel was used. The lowest RMSD values (3.51 Å) for the Nonhomolog dataset is when C was equal to 0.25 and the rbf kernel was used. The lowest average RMSD value (2.14 Å) was obtained on the Homolog dataset when C was set to 0.75 and the rbf kernel was used.

After modeling the PPII using Monte Carlo simulations, the average RMSD values have a maximum at 4.92 Å and a minimum at 1.88 Å (Table 4.2). The RMSD has a mean of 3.01 Å and a standard deviation of 0.85 Å. The RMSD values before (2.14 Å) and after (3.01 Å) building the PPII using Monte Carlo simulations are not significant difference, but the usage of Monte Carlo simulations in modeling the peptide is to reduce the clashes between PPII peptide and the query protein.

Table 4.1: Statistics of RMSD between the transfer PPII and the native PPII in different SVM models.

Kernel	C	Dataset	Feature	Average RMSD (Å)
rbf	0.25	LOFO	2, 5	3.86
		Non-homolog	2, 5, 6, 8	3.51*
		Homolog	2, 4, 8	2.23
	0.5	LOFO	3, 5, 9	3.80*
		Non-homolog	2, 5, 6, 8	3.56
		Homolog	2, 4, 5, 8	2.15
	0.75	LOFO	1, 4, 5, 6, 7	4.06
		Non-homolog	1, 2, 5, 11	3.81
		Homolog	2, 4, 11	2.14*
linear	0.25	LOFO	1, 3, 4, 10	4.12
		Non-homolog	1, 6, 9	4.21
		Homolog	2, 6, 7, 8, 10, 11	2.54
	0.5	LOFO	1, 3, 4, 10	4.15
		Non-homolog	1, 4, 7, 10	4.22
		Homolog	2, 6, 7, 8, 10, 11	2.54
	0.75	LOFO	1, 4, 9	4.1
		Non-homolog	1, 4, 8, 9	4.22
		Homolog	1, 2, 4, 5, 9, 11	2.48
poly	0.25	LOFO	6, 11	3.9
		Non-homolog	1, 2, 3, 4, 6, 7, 8	3.81
		Homolog	1, 2, 6, 8, 9, 11	2.25
	0.5	LOFO	6	4.02
		Non-homolog	1, 2, 3, 4, 6, 7, 10	3.8
		Homolog	2, 3, 5, 6, 8, 10, 11	2.19
	0.75	LOFO	6	4.02
		Non-homolog	1, 2, 3, 4, 6, 8, 10	3.83
		Homolog	2, 3, 5, 6, 8, 10, 11	2.19

*is for the best RMSD values

4.3.3 Benchmarking with Other Protein-Peptide Interaction Prediction Methods

We also compared our prediction with state-of-the-art methods in predicting protein-peptide interactions, namely CABS-dock[124] and GalaxyPepDock[123] (Table 4.2). The CABS-dock method applies Monte Carlo simulations to search the binding site of the fully flexible given peptide in the receptor with small fluctuations of its backbone. In GalaxyPepDock a template from database which is homologous to the given receptor is determined and then models are built using energy-based optimization. In comparison, our method on homologous template prediction has the RMSD of 2.14 Å and 3.01 Å for transfer peptides and peptides built by Monte Carlo simulations, respectively. The average RMSD of all built models by our method (3.01 Å) are lower than both CABS-dock (9.60 Å) and GalaxyPepDock (3.70 Å) methods.

Table 4.2: RMSD benchmarking for CABS-dock, GalaxyPepDock and our methods.

Structure ID	Family	PDB ID	transfer	Monte Carlo	CABS	Galaxy
1	EVH1	1K5D	1.7	1.88	20.64	4.82
2	EVH1	1RRP*$	4.78	3.49	23.48	3.96
3	EVH1	4B6H	2.82	3.34	10.76	4.1
4	EVH1	4WSF$	4.53	4.08	3.76	2.74
5	EVH1	5J3T	0.43	2.04	11.4	2.78
6	GYF	1L2Z	3.07	2.05	6.67	3.12
7	GYF	3FMA	3.07	4.92	18.43	22.84
8	MHCI	2QRT*	2.65	2.5	8.99	1.36
9	MHCI	4CW1*$	4.8	4.7	12.19	2.64
10	MHCI	5C0D	1.3	3.11	9.38	21.29
11	MHCII	1JK8	1.08	1.99	-	2.32
12	MHCII	3PDO*	3.4	2.34	-	1.8
13	MHCII	4P57	1.44	2.31	-	1.38
14	MYND	2ODD*	2.41	2.83	10.36	2.5
15	Profilin	2PBD	0.72	2.27	10.98	1.85
16	Profilin	2V8F	0.72	2.39	15.5	1.82
17	SH3	1CKA	1.92	2.64	7.29	1.09
18	SH3	1GBQ	0.84	1.92	6.11	1.28
19	SH3	1SEM	1.62	1.98	9.6	1.86
20	SH3	1UTI	1.51	2.48	8.28	2.87
21	SH3	1YWO	1.55	2.57	9.33	2.1
22	SH3	2DF6	2.08	4.3	5.61	3.08
23	SH3	2DRM	1.61	2.15	4.14	1.07
24	SH3	2J6F	1.18	2.48	6.67	2.07

25	SH3	2LCS	2.79	3.18	6.39	2.67
26	SH3	2ROL	3.54	3.73	9.53	3.09
27	SH3	2RPN	1.4	4.27	3.78	3.19
28	SH3	2VKN	0.98	2.29	6.53	2.48
29	SH3	2VWF	2.21	2.49	11.27	2.57
30	SH3	3I5R	0.82	3.28	13.67	1.14
31	SH3	3U23	1.74	4.8	4.08	1.82
32	SH3	3ULR	1.73	2.63	7.7	11.79
33	SH3	4CC2	3.76	2.34	9.17	1.54
34	SH3	4F14	0.86	2.86	7.24	2.56
35	SH3	4HVW	2.35	3.37	11.39	1.74
36	SH3	4J9C	1.61	2.78	7.58	1.5
37	SH3	4LNP	2.27	2.43	9.36	1.37
38	SH3	4U5W	2.27	4.48	12.17	4.06
39	WW	1JMQ*	2.43	3.55	12.64	3.98
40	WW	2EZ5*	2.41	2.62	11.68	2.74
41	WW	2JO9	0.88	3	6.75	2.87
42	WW	2LAJ*	2.93	3.88	8.3	4.12
43	WW	2LAW*	3.13	4.02	7.63	4.28
44	WW	2LAZ*	2.62	3.56	7.06	6.36
RMSD	-	-	2.14	3.01	9.60	3.70
Standard deviation	-	-	1.09	0.85	4.19	4.38

* Cases where template and target structure are nonhomologous

$ Cases where the predictions of transfer RMSD higher than 4 Å

"transfer" refers to the PPII peptide from template structure was transferred into the target structure.

CABS and Galaxy refer to CABS-dock and GalaxyPepDock methods, respectively

“Average” refers to the average values of RMSD among 10 models built by protein-peptide prediction methods (either by CABS-dock or GalaxyPepDock) or among 40 models by Monte Carlo simulations.

4.3.4 Searching possible PPII receptors in the PDB

We have used our protocol to predict possible PPII receptors among 17, 000 non-redundant (30% sequence identity) proteins from the PDB. 138 structures were predicted as having the PPII-binding sites (Table 4.3). 13 cases (of the 138) belong to the eight known PPII receptor families. The other 125 structures are from different families, some of which have the function in signaling network and immune response. Three interesting examples are NADPH oxidase (PDB ID: 1OEY) (ubiquitin-like), clathrin adaptor (PDB ID: 1KYF) and secretion chaperon-like (PDB ID: 1JYA). In these cases, the location of Trp and Donor residues that we predicted to bind PPII actually has a bound-peptide (Figure 4.6b) or a part of the partner protein (Figure 4.6a,c). The RMSD values between the transferred peptides from the templates and the native peptides range from 3.36 to 3.69 Å. In addition to these three cases, the other 18 cases, in which the predicted location of Trp-Donor residues are in the homo-oligomer interaction sites (Table 4.3). The homo-oligomers do not always have the peptide-mediate interactions and hence RMSD values are not calculated.

Table 4.3: List of the predicted PPII-receptors (*cases where the predicted PPII-binding sites are in the interaction sites with its homo-oligomers)

PDB ID	Chain	Dnr	Trp	Scop Fold[133]	Pfam Description[134]
12AS	A	84	76	Class II aaRS and biotin synthetases	Aspartate-ammonia ligase
1A0T	P	361	337	Transmembrane beta-barrels	LamB porin
1A8D	A	10	7	beta-Trefoil, Concanavalin A-like lectins/glucanases	Clostridium neurotoxin, receptor binding (C-terminal)
1ACF	A	5	2	Profilin-like	Profilin
1B3T	A	497	503	Ferredoxin-like	Epstein Barr virus nuclear antigen-1, DNA-binding domain
1B5Q	A	413	285	FAD/NAD(P)-binding domain	Flavin containing amine oxidoreductase
1B8K	A	18	20	Cystine-knot cytokines	Nerve growth factor family
1B9M*	A	183	186	OB-fold, OB-fold, DNA/RNA-binding 3-helical bundle	Bacterial regulatory helix-turn-helix protein, lysR family, TOBE domain
1BGF	A	67	37	Transcription factor STAT-4 N-domain	STAT protein, protein interaction domain
1BIA	A	42	46	Class II aaRS and biotin synthetases, SH3-like barrel	HTH domain, Biotin/lipoate A/B protein ligase family
1BM8	A	81	33	Mlu1-box binding protein MBP1	KilA-N domain
1BQU*	A	138	192	Immunoglobulin-like beta-sandwich	Interleukin-6 receptor alpha chain, binding, Fibronectin type III domain
1BS0	A	6	3	PLP-dependent transferase-like (DNA-binding domain)	Aminotransferase class I and II
1BVY	F	536	574	Flavodoxin-like	Flavodoxin
1CFB	A	750	762	Immunoglobulin-like beta-sandwich	Fibronectin type III domain

1CNT*	1	174	64	4-helical cytokines	Ciliary neurotrophic factor
1COZ	A	9	74	Adenine nucleotide alpha hydrolase-like	Cytidylyltransferase
1CV8	A	165	143	Cysteine proteinases	Staphopain peptidase C47
1CXQ	A	97	76	Ribonuclease H-like motif	Integrase core domain
1D02	A	194	12	Restriction endonuclease-like	Type II restriction enzyme MunI
1D0D	A	25	37	BPTI-like	Kunitz/Bovine pancreatic trypsin inhibitor domain
1D2S	A	47	100	Concanavalin A-like lectins/glucanases	Laminin G domain
1DDW	A	76	24	PH domain-like barrel	WH1 domain
1DI2	A	153	126	dsRBD-like	Double-stranded RNA binding motif
1DMG	A	100	97	Ribosomal protein L4	Ribosomal protein L4/L1 family
1DS1	A	285	288	Double-stranded beta-helix	Taurine catabolism dioxygenase TauD, TfdA family
1E2K*	A	306	310	P-loop containing nucleoside triphosphate hydrolases	Thymidine kinase from herpesvirus
1E6U	A	276	202	NAD(P)-binding Rossmann-fold domains	NAD dependent epimerase/dehydratase family
1EDZ	A	11	56	NAD(P)-binding Rossmann-fold domains	Tetrahydrofolate dehydrogenase/cyclohydrolase, NAD(P)-binding domain
1EG3	A	81	83	WW domain-like, EF Hand-like, EF Hand-like	WW domain, EF hand, EF-hand
1EQ2*	A	81	84	NAD(P)-binding Rossmann-fold domains	NAD dependent epimerase/dehydratase family
1F0K	A	136	137	UDP-Glycosyltransferase/glycogen phosphorylase	Glycosyltransferase family 28 N-terminal domain

1FSU	A	479	438	Alkaline phosphatase-like	Sulfatase
1GCQ*	C	654	636	SH3-like barrel	Variant SH3 domain
1GSA	A	265	130	ATP-grasp, PreATP-grasp domain	Prokaryotic glutathione synthetase, ATP-grasp domain (N-terminal)
1H4X	A	93	98	SpoIIaa-like	STAS domain
1H8A	C	179	166	DNA/RNA-binding 3-helical bundle	Myb-like DNA-binding domain
1HDK	A	75	72	Concanavalin A-like lectins/glucanases	Galactoside-binding lectin
1HQI	A	64	61	Monooxygenase (hydroxylase) regulatory protein	MmoB/DmpM family
1HT6	A	212	299	Glycosyl hydrolase domain, TIM beta/alpha-barrel	Alpha-amylase C-terminal beta-sheet domain
1HYO	A	371	367	SH3-like barrel, FAH	Fumarylacetoacetate (FAA) hydrolase family (N-terminal)
1HZ4	A	208	211	alpha-alpha superhelix	Transcription factor MalT domain III
1I71	A	35	70	Kringle-like	Kringle domain
1I8A	A	68	71	Immunoglobulin-like beta-sandwich	Domain of unknown function (DUF1083)
1IA9	A	1581	1714	Protein kinase-like (PK-like)	Alpha-kinase family
1IAR	B	193	190	Immunoglobulin-like beta-sandwich	Interleukin-4 receptor alpha chain, N-terminal
1IG0*	A	287	270	Thiamin pyrophosphokinase	Thiamin pyrophosphokinase, vitamin B1 binding domain
1IOJ	A	48	41	Fragments of the apolipoproteins	The apolipoprotein C-I (The apoC-1)
1J2R*	A	143	115	Isochorismatase-like hydrolases	Isochorismatase family
1J58	A	183	171	Double-stranded beta-helix	Cupin, Cupin
1J5W	A	119	115	Class II aaRS and biotin synthetases	Glycyl-tRNA synthetase alpha subunit

1JL1	A	72	120	Ribonuclease H-like motif	RNase H
1JOV	A	242	229	Supersandwich	Aldose 1-epimerase
1JY1	A	506	392	Phospholipase D/nuclease	Tyrosyl-DNA phosphodiesterase
1JYA	A	41	90	Secretion chaperone-like	Tir chaperone protein (CesT) family
1JYH	A	134	144	Probable bacterial effector-binding domain	GyrI-like small molecule binding domain
1JYK	A	196	136	Nucleotide-diphospho-sugar transferases	Nucleotidyl transferase
1K0H	A	77	89	Phage tail proteins	Phage Head-Tail Attachment
1K5N	A	146	147	Immunoglobulin-like beta-sandwich, MHC	Class I Histocompatibility antigen, domains alpha 1 and 2
1K77	A	226	257	TIM beta/alpha-barrel	Xylose isomerase-like TIM barrel
1KAE	A	248	353	ALDH-like	Histidinol dehydrogenase
1KFT	A	54	74	SAM domain-like	Helix-hairpin-helix motif
1KJQ	A	139	178	Barrel-sandwich hybrid, PreATP-grasp domain	ATP-grasp domain
1KKO	A	224	217	TIM beta/alpha-barrel, Enolase N-terminal domain-like	Methylaspartate ammonia-lyase N-terminus
1KNZ	A	134	87	NSP3 homodimer	Rotavirus non-structural protein NSP3
1KOL	A	156	353	GroES-like, NAD(P)-binding Rossmann-fold domains	Alcohol dehydrogenase GroES-like domain, Alanine dehydrogenase/PNT
1KT6	A	54	91	Lipocalins	Lipocalin / cytosolic fatty-acid binding protein family
1KYF	A	841	840	Immunoglobulin-like beta-sandwich	Adaptin C-terminal domain, Alpha adaptin AP2, C-terminal domain
1LBA	A	46	41	N-acetylmuramoyl-L-alanine amidase-like	N-acetylmuramoyl-L-alanine amidase
1LG4	A	82	83	Prion-like	Prion/Doppel alpha-helical domain
1LO7	A	24	23	Thioesterase/thiol ester dehydrase-isomerase	Thioesterase superfamily

1MAI	A	43	52	PH domain-like barrel	PH domain
1MIJ	A	1257	1291	DNA/RNA-binding 3-helical bundle	Homeo-prospero domain
1MIW	A	400	392	Poly A polymerase, Nucleotidyltransferase	tRNA nucleotidyltransferase domain 2 putative
1MUN	A	142	154	DNA-glycosylase	Iron-sulfur binding domain of endonuclease III
1N4Q	B	235	188	alpha/alpha toroid	Prenyltransferase and squalene oxidase repeat
1NEI	A	48	60	Hypothetical protein YoaG	Domain of unknown function (DUF1869)
1NF1	A	1487	1491	GTPase activation domain, GAP	GTPase-activator protein for Ras-like GTPase
1NG2	A	173	193	SH3-like barrel	SH3 domain
1NO1	A	63	33	Replisome organizer (g39p helicase loader/inhibitor)	Loader and inhibitor of phage G40P
1O59	A	326	247	Galactose-binding domain-like	Allantoicase repeat, Allantoicase repeat
1O6W	A	15	26	WW domain-like	WW domain
1O9I*	A	132	33	Ferritin-like	Manganese containing catalase
1OEY	A	395	425	beta-Grasp (ubiquitin-like)	PB1 domain
1OPC	A	224	226	DNA/RNA-binding 3-helical bundle	Transcriptional regulatory protein, C terminal
1ORR	A	241	335	NAD(P)-binding Rossmann-fold domains	NAD dependent epimerase/dehydratase family
1OUW	A	87	12	beta-Prism I	Jacalin-like lectin domain
1OV2	A	50	29	RAP domain-like	Alpha-2-macroglobulin RAP, N-terminal domain

1OWW	A	40	46	Immunoglobulin-like beta-sandwich	Fibronectin type III domain
1OZJ	A	125	94	SMAD MH1 domain	MH1 domain
1PKH	A	115	133	beta-clip	dUTPase
1PM4	A	95	72	Superantigen (mitogen) Ypm	Yersinia pseudotuberculosis mitogen
1PMI	A	109	18	Double-stranded beta-helix	Phosphomannose isomerase type I
1PUJ	A	162	160	P-loop containing nucleoside triphosphate hydrolases	50S ribosome-binding GTPase
1Q9C	A	143	85	Histone-fold	Core histone H2A/H2B/H3/H4
1QO0	D	19	43	Flavodoxin-like	ANTAR domain
1QQE	A	192	196	alpha-alpha superhelix	Soluble NSF attachment protein, SNAP
1QRV	A	10	43	HMG-box	HMG (high mobility group) box
1QWD	A	77	86	Lipocalins	Lipocalin-like domain
1R6X	A	353	374	Adenine nucleotide alpha hydrolase-like (PUA-like)	PUA-like domain, ATP-sulfurylase
1RA0	A	109	146	metallo-dependent hydrolases,TIM beta/alpha-barrel	Amidohydrolase family
1RGX	A	39	82	Resistin	Resistin
1RHS	A	6	14	Rhodanese/Cell cycle control phosphatase	Rhodanese-like domain
1RP0*	A	176	160	FAD/NAD(P)-binding domain	Thi4 family
1RXD	A	78	67	(Phosphotyrosine protein) phosphatases II	Protein-tyrosine phosphatase
1RXQ*	A	46	177	DinB/YfiT-like putative metalloenzymes	DinB superfamily
1RY3	A	25	18	Leucocin-like bacteriocin	Class II bacteriocin
1RY9*	A	34	55	Secretion chaperone-like	Invasion protein B family
1S1D	A	186	128	5-bladed beta-propeller	Apyrase
1S5D	A	84	127	ADP-ribosylation	Heat-labile enterotoxin alpha chain
1SE8	A	97	88	OB-fold	Single-strand binding protein family

1SG4	A	200	201	ClpP/crotonase	Enoyl-CoA hydratase/isomerase family
1SPK	A	62	44	SH3-like barrel	Variant SH3 domain
1SQ4	A	229	231	Double-stranded beta-helix	Cupin domain, Cupin domain
1SQG	A	152	148	NusB-like, methyltransferases	NusB family, NOL1/NOP2/sun family
1SYX	B	30	52	GYF/BRK domain-like	GYF domain
1T0B*	A	146	131	Flavodoxin-like	Trehalose utilisation
1T1G	A	305	230	Subtilisin-like	Subtilase family
1T33	A	73	65	DNA/RNA-binding, Tetracyclin repressor-like	Bacterial regulatory proteins, tetR family, Domain of unknown function
1TG0	A	60	40	SH3 domain	SH3 domain
1TJO*	A	45	53	Ferritin-like	Ferritin-like domain
1TLY	A	216	262	Transmembrane beta-barrels	Nucleoside-specific channel-forming protein, Tsx
1TWD	A	9	202	TIM beta/alpha-barrel	CutC family
1U0T*	A	290	304	NAD kinase/diacylglycerol kinase-like	ATP-NAD kinase
1U3E	M	18	3	DNA-binding domain (intron-encoded endonucleases)	NUMOD4 motif, HNH endonuclease
1U7K*	A	18	23	Retrovirus capsid protein, N-terminal core domain	Gag P30 core shell protein
1UDD	A	28	212	Heme oxygenase-like	TENA/THI-4/PQQC family
1UII	A	98	99	Parallel coiled-coil	Geminin
1UTE	A	171	168	Metallo-dependent phosphatases	Calcineurin-like phosphoesterase
1UUJ*	A	18	55	Lissencephaly-1 protein (Lis-1, PAF-AH alpha)	LisH
1UV7	A	114	116	RRF/tRNA synthetase additional domain-like	Type II secretion system (T2SS), protein M
1UXY	A	254	267	FAD-binding/transporter-associated domain-	FAD binding domain

				like	
1V2X	A	71	73	alpha/beta knot	SpoU rRNA Methylase family
1V4A	A	140	133	Nucleotidyltransferase	Glutamate-ammonia ligase adenylyltransferase
1V4P*	A	25	29	RRF/tRNA synthetase additional domain-like	Threonyl and Alanyl tRNA synthetase second additional domain
1V64	A	26	52	HMG-box	HMG-box domain
1V88	A	17	23	PH domain-like barrel	PH domain
1V9Y	A	108	110	Profilin-like	PAS domain

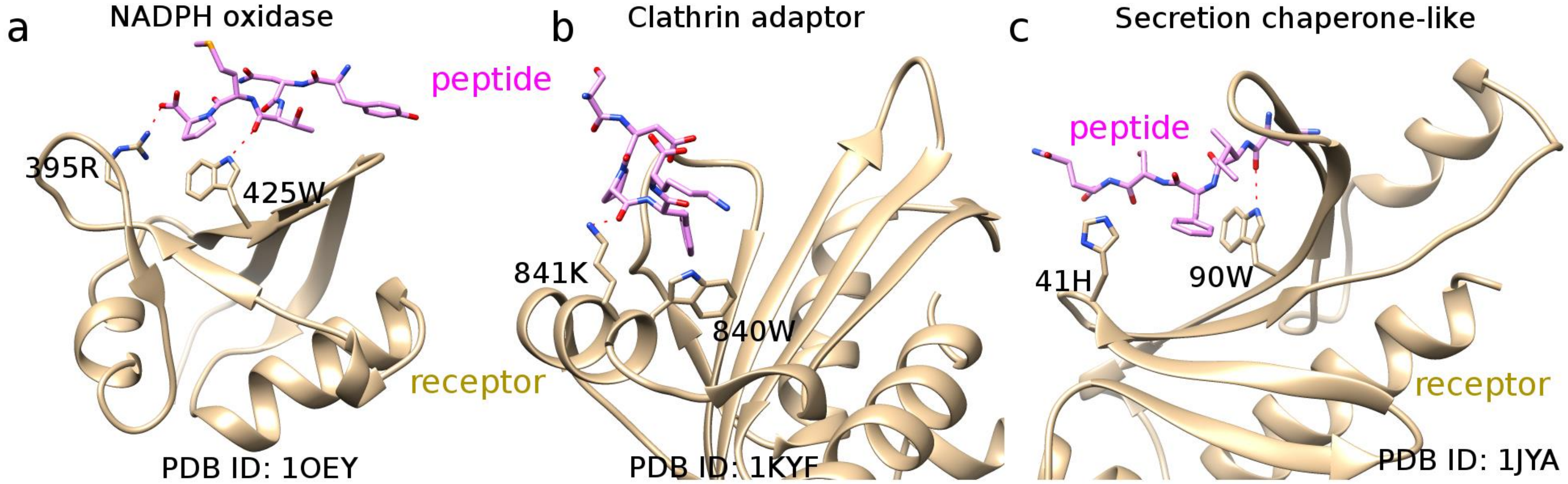


Figure 4.6: New cases of the PPII-binding proteins. (a) NADPH oxidase (PDB ID: 1OEY), (b) Clathrin adaptor (PDB ID: 1KYF) and (c) Secretion chaperon-like (PDB ID: 1JYA). The peptides or the parts of the partner proteins are shown in pink and the receptors are shown in light brown. The donor, Trp residues, and peptides are in stick representation. The hydrogen bond interactions between peptide and receptor are shown in red dotted lines.

4.4 Discussion

The aim of this study is to (i) reveal the key features of the PPII receptor sites, (ii) accurately model the PPII-protein interactions and (iii) identify new PPII-binding proteins from the PDB. The results show that the specificities in the number of hydrogen bond, depth values, entropy conservation and $N^{\varepsilon 1}$-NX distance could be used as the signals for PPII receptor sites. The Trp residues that are highly conserved in each PPII-binding site are shown to be important in the PPII recognition. The importance of Trp residues could be explained by their possibility of making hydrogen bond from their side chain atoms, their potentials on making hydrophobic and van der Waals interactions. Recently, it is shown that in the SH3 domain of CASKIN2 scaffolding protein, an Arg residue has replaced the PPII recognizing Trp, and hence, CASKIN2 lacks the ability to interact with the calmodulin kinase domain. The mutagenesis from Arg to Trp restores the interaction[135]. Another example showing the importance of Trp is in the case of smurf2 WW3 domain. When Phe is presented at the hydrogen bond making Trp position, the binding affinity to the PPII peptide decreases in comparison with the canonical WW3 domain[136].

The average RMSD of the PPII modeled by Monte Carlo simulations and the native PPII has a mean of 3.01 Å and a standard deviation at 0.85 Å. In comparison with other available protein-peptide interaction prediction methods, particularly CABS-dock and GalaxyPepDock, our method predicts the PPII with lower RMSD values. The CABS-dock method, in addition, only works in the single chain proteins, and hence, for the heterogous structures such as MHCII, the method returns no result. The GalaxyPepDock, on the other hand, allows users to submit only three jobs at a time. The GalaxyPepDock also uses the homologous to find the location of peptide. The other protein-peptide prediction programs, such as PepSite or FlexPepDock are not user friendly and are out of the benchmarking. Using only two residues as the templates, the RMSD of the modeled and the native PPII peptides in our method is better than both GalaxyPepDock and CABS-dock methods.

When applying our method on a data set of 17005 structures with the sequence identity lower than 30%, we detected 125 new PPII-binding proteins, which have not been trained in our model. Three of these structures have the peptides, or stretches of protein that are bound exactly to the Trp and donor residues we predicted. A web-server for this prediction will be available soon. The user gets not only the key residues for binding PPII, but also the location of the PPII backbone atoms.

Our assumption in this study is that the conformation of the PPII receptor sites in the PPII bound and unbound conditions would not dramatically change and hence, it cannot account for the conformational flexibility. Another limitation is that our method requires at least two hydrogen bond interactions between the PPII and its receptors, and hence, some of the receptors, which make only one hydrogen bond interactions or lost the Trp residue, could not be predicted. For example, the collagen-bound protein[137-139] does not have hydrogen bond making Trp residue, and could not be predicted by our method. This limitation could be avoided if this protocol is generalized to two hydrogen bond requirement, without preferences on Trp residues. In addition, while we consider only the hydrogen bond interactions between the receptors and the PPII peptides, the shape of the interface is not taken into consideration because of the difficulty of this feature. This PPII prediction protocol is only concerned about the conformation of proteins, as well as, peptides, but not their sequence. The optimal sequences of the protein and peptide could be dealt with in future as an extension of this study. However, this first PPII receptor prediction model could be beneficial for detecting the PPII receptor in signaling network and immune system. The protocol in this study could be generalized for other conformations, such as alpha helical conformation. As the PPII peptide could be a potential source of new peptide-based novel drugs[140,141], a search for the receptor of the PPII could benefit the understanding of pathways, as well as, the side effect of synthetic peptides. As this study only predicts whether or not the protein could bind the PPII, the specificity of the peptide sequence has not been considered. The further extension of this work could be on predicting the affinity of a particular PPII peptide to a PPII-binding protein.

Chapter 5

pK_a Prediction of Ionizable Amino Acid Residues in Proteins

The interaction or a function of a protein highly correlates with the charge of its ionizable residues. The charges are monitored by the protonation state of these residues. One measurement can be used to identify the protonation state is pKa, and in this chapter, we explained how we can predict the pKa of ionizable residues in protein

5.1 Background and Motivations

5.1.1 Significance of the pK_a Prediction

The acid dissociation constant, or more commonly its negative logarithm (base 10), pK_a, measures the protonation strength of an acid in solution. In proteins, this pK_a term is also used to quantify the protonation state of ionizable amino acid residues including ASP, GLU, HIS, LYS, and ARG. The importance of pK_a in the protein structure and the protein function is illustrated by the fact that about 25% of protein residues and 65% of residues in the active sites[142] are ionizable. The change of pH can induce the shift of the ionizations, which affects electrostatic interactions, as well as, the molecular structure and the function. Particularly, the protonation states of residues modulate many protein properties, such as folding[143], stability[144,145], solubility[146], dynamics[147], interactions[148] and other functions[149-152]. And hence, estimating or predicting pK$_a$ is a powerful means of investigating the protein function. To decipher these pH-dependent processes, it is important to correctly estimate the pK$_a$ values; which could reveal the underlying physical principles guiding these processes.

pK$_a$s of ionizable amino acid residues depend on their immediate protein/solvent environment. In bulk solvent, the environmental physicochemical properties of a chemical group are simply the properties of a homogeneous aqueous solution of water. However, proteins have inhomogeneous environment and their physicochemical properties could drastically change across different regions in proteins. For example, the average dielectric constant of a polar chemical group in bulk solvent could be about ~80, while its value in protein is around ~20 - 30 on the protein surface and ~6 - 7 in the protein interior[153]. Therefore, it is not surprising that the same chemical group or amino acid residue would behave differently depending on their location in proteins.

A popular experimental method to measure the pK$_a$ of protein residues is nuclear magnetic resonance (NMR)-monitored pH titration[154-156]. Other techniques, such as isothermal titration calorimetry, enzymatic pH-activity profiles[157,158], potentiometric titration and site-directed mutagenesis[159-161] are less commonly used. However, all these techniques are time consuming and expensive. The size of proteins is also another limitation of experimental methods, particularly, NMR. Hence, validated or calibrated computational pK$_a$ predictions could be a useful way to estimate the pK$_a$ values, especially when experimental measurements are difficult or not possible[162].

5.1.2 Review of Previous Works

The standard $\mathrm{pK_a}$ value of an ionizable amino acid residue can be determined in aqueous solution of the isolated form of this residue. The common term used in textbook for standard $\mathrm{pK_a}$ is model $\mathrm{pK_a}$, which will be used throughout the rest of this thesis. This $\mathrm{pK_a}$ value correlates with the standard-state Gibbs free energy (ΔG^o) as follows:

$$\Delta G^o = 2.303.\mathrm{R.T.pK_a}\ (5.1)$$

where R is ideal gas constant, $R = 8.314\, J/(mol.K)$,

T is temperature (in K)

In proteins, this $\mathrm{pK_a}$ value of an amino acid could shift from its model value by an additional energy term when this residue is transferred from the solvent to the protein environment. This energy value is determined by the electrostatics and other energy terms of immediate surroundings or microenvironment of the residue. Several approaches have been used to predict $\mathrm{pK_a}$, namely (1) macroscopic approaches, (2) microscopic approaches, and (3) empirical approaches.

(1) Macroscopic approaches

The transferred energy in macroscopic approaches can be directly calculated from the macroscopic electrostatics equations or Poisson-Boltzmann equation (PBE). Techniques classified under macroscopic approaches include PBE[163-166], PBE and conformational flexibility[167,168], or Generalized Born[169-171] methods. The limitations of these methods are their underestimation of hydrogen-bonding and desolvation effects[172] and overestimation the intra-protein charge–charge interactions[173,174] in calculating pK_a shifts.

(2) Microscopic approaches

Microscopic approaches quantify all interactions at the atomic resolution. These approaches do not include any macroscopic physical features. These are the most desirable approaches because of their accuracy. However, their disadvantages are intensive computational complexity and time requirement.

In these approaches, the quantities of electrostatic, and other physical interactions can be obtained by solving the Schrodinger equation. However, the current computing power is not sufficient for exactly solving the equation, and hence, some levels of approximation are applied. Several methods are in this category, including, quantum mechanical/molecular mechanics (QM/MM) based methods[175-177], molecular dynamics (MD) based methods[178-180], or continuum solvent models from the microscopic description[181-185].

(3) Empirical approaches

Empirical approaches use a statistical analysis over a large database of experimentally determined pK_a values. This method has the advantage in speed; however, the physical meanings of the determinants contributing to the pK_a value are not clearly understood. PROPKA[186,187] and MoKaBio[187] are classified as empirical methods. Among these methods, PROPKA is the most widely used because of its small root-mean-square deviation (reported as less than 1 pH unit).

Our pK_a prediction method falls into this category. As this empirical method requires abundant data for training and testing purpose, it only has the ability to predict ionizable amino acid residue types with sufficient number of experimentally available pK$_a$ values, particularly, ASP, GLU, HIS and LYS. The next section explained the definition of priority features which are used to predict the pK$_a$.

5.2 Residue Depth

Atom/residue depth is a measurement of the atomic or residue distance to the nearest surface bulk water[188,189]. A water molecule is a bulk solvent if it is surrounded by more than three neighbour waters within a sphere of 4.2 Å radius. Depth has been shown to correlate with a various physical and chemical properties in protein structures, including structural stability[188], hydrogen/deuterium amide proton exchange rates[188,190], sizes of globular domains[188,191], hydrophobicity[188,191,192], residue conservation[192], protein activity and 3D structural model accuracy[193]. In the context of proteins, pK$_a$ values are correlated with their immediate environments and could differ from the model pK$_a$ values. We used depth and other features to predict these shifts by characterizing the environment of ionizable groups.

In the next two sections, 5.3 and 5.4, two different methods to predict the pK$_a$ are explained in detail.

5.3 DEPTH-based pK$_a$ Prediction

The predicted pK$_a$, pK_a^{pred}, is computed as follows:

$$pK_a^{pred} = pK_a^{model} + c_1.depth^{MC} + c_2.depth^{polarSC} + c_3.HB + c_4.EE_R + c_5.ASA^{SC} + c_0 \quad (5.2)$$

where pK_a^{model} is the model pK$_a$ (Table 5.1).

$c_0 - c_5$ are coefficients of the individual features.

The values of the coefficients were optimized over a training set of residues.

Table 5.1: RMSD of predicted pK_a (in pH units) from experimentally determined values

Residue type	model pK_a	c_0	c_1	c_2	c_3	c_4	c_5	RMSD (pH units)	
								Training set (size)	Testing set (size)
ASP	3.8	-2.18	0.29	0.47	-0.61	0.16	-0.15	1.02 (112)	0.71 (15)
GLU	4.5	-1.91	-0.1	0.79	-0.19	0.26	-0.09	0.83 (125)	1.07 (15)
HIS	6.5	3.13	-0.04	-0.54	0.28	-1.12	-0.83	1.14 (60)	1.26 (15)
LYS	10.5	4.22	-0.21	-0.19	-0.01	-7.65	-1.81	0.86 (70)	0.8 (15)
Total								0.94 (367)	0.96 (60)

5.3.1 Features Constructing the DEPTH Model

We used the following features to describe the environment, namely (1) depth, (2) electrostatic energy, (3) number of hydrogen bond, and (4) solvent accessible surface area.

(1) Depth

To accurately describe the solvent effects on an ionizable group, two complementary measures of depth are used in our predictor, particularly, average depth of main chain atoms ($depth^{MC}$), and average depth of polar side chain atoms ($depth^{polarSC}$)

(2) Electrostatic energy (EE_R):

All hydrogen atoms were explicitly added using the program Reduce[194] for the electrostatics energy calculation. This energy term is calculated as follows:

$$EE_R = \sum_{i \in R} \sum_{j \in R_b} \frac{Q_i \cdot Q_j}{r_{ij}} \text{ (5.3)}$$

where: Q_i is the partial charge of an atom i in a residue R.

Q_j is the partial charge on an atom j in a residue R_b of the surrounding microenvironments (within a cut-off distance of 12 Å from the atom i).

r_{ij} is the atomic distance between i and j atoms.

We assumed that all acidic groups of ASP and GLU residues were deprotonated, whereas the basic groups of HIS and LYS residues were protonated. The values of partial charges Q_i and Q_j were obtained from the gromos43a1 force field[195].

(3) Hydrogen Bond (HB):

If the distance between donor-acceptor atom pairs was less than 3.5 Å and the donor-acceptor-acceptor antecedent angle was greater or equal to 100°[196,197], the bond was identified as a hydrogen bond.

(4) Solvent accessible surface area:

The Shrake–Rupley algorithm[198] was used to compute solvent accessible surface area of side chain atoms (ASA^{SC}).

5.3.2 Dataset of experimental values of pK$_a$ used in DEPTH Prediction

The coefficients $c_0 - c_5$ of separate amino acid reside types in equation 5.2 were obtained by optimizing the predictions on the training set. The number of training residues for ASP, GLU, LYS and HIS are 112, 125, 70 and 60 respectively (Table 5.2). The prediction formula was then tested on a set of 15 residues for each amino acid type (Table 5.2 and Table 5.3). The data on testing and training sets did not overlap with each other.

In the cases where the pK$_a$s were determined for mutant residues of proteins, to construct homology models we used the mutate_residue command of MODELLER[40]. In the cases where more than one alternative conformation for residues were reported, the first listed conformation was always chosen.

Table 5.2: Listing of experimentally determined pK_a values of ionizable residues and their sources. 367 of the values are used for training (number 1 in bracket at Method (set) column) of the predictor, and 60 are used on testing (number 2 in bracket at Method (set) column).

protein name	Reference	PDB code	Residue number (chain)	Residue name	$pK_a^{\exp}$	Method (set)	pK_a^{pred}	Error = $(pK_a^{pred} - pK_a^{\exp})$
The apo E2	199	1LE2	69 (A)	LYS	10.1	X-ray (1)	10.59	0.49
The apo E2	199	1LE2	72 (A)	LYS	10	X-ray (1)	9.72	-0.28
The apo E2	199	1LE2	75 (A)	LYS	10	X-ray (1)	9.99	-0.01
The apo E2	199	1LE2	95 (A)	LYS	10.2	X-ray (1)	10.4	0.2
The apo E2	199	1LE2	143 (A)	LYS	9.4	X-ray (1)	9.43	0.03
The apo E2	199	1LE2	146 (A)	LYS	9.9	X-ray (2)	9.89	-0.01
The apo E2	199	1LE2	157 (A)	LYS	10.9	X-ray (1)	10.77	-0.13
The apo E3	200	1NFN	69 (A)	LYS	10.4	X-ray (1)	10.38	-0.02
The apo E3	200	1NFN	72 (A)	LYS	10	X-ray (1)	9.93	-0.07
The apo E3	200	1NFN	75 (A)	LYS	10.1	X-ray (1)	10.4	0.3
The apo E3	200	1NFN	95 (A)	LYS	10.1	X-ray (1)	10.4	0.3
The apo E3	200	1NFN	143 (A)	LYS	9.5	X-ray (2)	9.88	0.38
The apo E3	200	1NFN	146 (A)	LYS	9.2	X-ray (2)	9.81	0.61
The apo E3	200	1NFN	157 (A)	LYS	11.1	X-ray (1)	10.7	-0.4
The apo E4	199	1GS9	69 (A)	LYS	10.1	X-ray (1)	10.32	0.22
The apo E4	199	1GS9	72 (A)	LYS	10	X-ray (1)	10.23	0.23
The apo E4	199	1GS9	75 (A)	LYS	10.1	X-ray (1)	10.06	-0.04
The apo E4	199	1GS9	95 (A)	LYS	10.1	X-ray (1)	10.36	0.26
The apo E4	199	1GS9	143 (A)	LYS	9.9	X-ray (1)	9.73	-0.17
The apo E4	199	1GS9	146 (A)	LYS	9.4	X-ray (2)	9.83	0.43
The apo E4	199	1GS9	157 (A)	LYS	10.9	X-ray (1)	10.38	-0.52
ATP synthase	201	1A91	7 (A)	ASP	5.6	NMR	4.21	-1.39
ATP synthase	201	1A91	61 (A)	ASP	7	NMR (1)	3.57	-3.43
Bacterial nuclease mutant	202	2SNM	66 (A)	LYS	6.4	X-ray (2)	7.47	1.07
Bacterial MutT	203	1MUT	39 (A)	LYS	8.4	NMR (1)	10.61	2.21
Bacterial phosphonoacetaldehyde hydrolase	204	1RQL	53 (A)	LYS	9.3	X-ray (1)	9.06	-0.24
Barnase	205	1A2P	8 (A)	ASP	3.3	X-ray (1)	2.67	-0.63
Barnase	205	1A2P	12 (A)	ASP	3.8	X-ray (1)	3.39	-0.41
Barnase	206	1A2P	18 (A)	HIS	7.73	X-ray (1)	6.87	-0.86
Barnase	205	1A2P	22 (A)	ASP	3.3	X-ray (1)	3.19	-0.11
Barnase	205	1A2P	29 (A)	GLU	3.75	X-ray (1)	3.87	0.12
Barnase	205	1A2P	44 (A)	ASP	3.35	X-ray (1)	3.71	0.36
Barnase	205	1A2P	54 (A)	ASP	2.2	X-ray (1)	3.14	0.94
Barnase	205	1A2P	60 (A)	GLU	3.2	X-ray (1)	4.31	1.11
Barnase	205	1A2P	73 (A)	GLU	2.2	X-ray (1)	4.14	1.94
Barnase	205	1A2P	75 (A)	ASP	3.1	X-ray (1)	5.06	1.96
Barnase	205	1A2P	86 (A)	ASP	4.2	X-ray (1)	4.12	-0.08
Barnase	205	1A2P	93 (A)	ASP	2	X-ray (1)	2.18	0.18
Barnase	205	1A2P	101 (A)	ASP	2	X-ray (1)	1.9	-0.1

Barnase	205	1A2P	102 (A)	HIS	6.3	X-ray (1)	6.69	0.39
B1 domain of protein G	207	1PGB	4 (A)	LYS	11	X-ray (1)	10.63	-0.37
B1 domain of protein G	207	1PGB	10 (A)	LYS	11	X-ray (1)	10.78	-0.22
B1 domain of protein G	207	1PGB	13 (A)	LYS	11	X-ray (1)	10.67	-0.33
B1 domain of protein G	207	1PGB	15 (A)	GLU	4.4	X-ray (1)	4.16	-0.24
B1 domain of protein G	207	1PGB	19 (A)	GLU	3.7	X-ray (1)	4.28	0.58
B1 domain of protein G	207	1PGB	22 (A)	ASP	2.9	X-ray (2)	2.98	0.08
B1 domain of protein G	207	1PGB	27 (A)	GLU	4.5	X-ray (2)	3.76	-0.74
B1 domain of protein G	207	1PGB	28 (A)	LYS	10.9	X-ray (2)	10.53	-0.37
B1 domain of protein G	207	1PGB	36 (A)	ASP	3.8	X-ray (2)	3.92	0.12
B1 domain of protein G	207	1PGB	40 (A)	ASP	4	X-ray (1)	3.89	-0.11
B1 domain of protein G	207	1PGB	42 (A)	GLU	4.4	X-ray (1)	4.29	-0.11
B1 domain of protein G	207	1PGB	46 (A)	ASP	3.6	X-ray (1)	2.87	-0.73
B1 domain of protein G	207	1PGB	47 (A)	ASP	3.4	X-ray (2)	3.26	-0.14
B1 domain of protein G	207	1PGB	56 (A)	GLU	4	X-ray (1)	4.51	0.51
B2 domain of protein G	207	1IGD	9 (A)	LYS	11	X-ray (1)	10.57	-0.43
B2 domain of protein G	207	1IGD	15 (A)	LYS	11	X-ray (1)	10.96	-0.04
B2 domain of protein G	207	1IGD	18 (A)	LYS	11	X-ray (1)	10.81	-0.19
B2 domain of protein G	207	1IGD	20 (A)	GLU	4.3	X-ray (1)	4.22	-0.08
B2 domain of protein G	207	1IGD	24 (A)	LYS	10.7	X-ray (1)	10.65	-0.05
B2 domain of protein G	207	1IGD	27 (A)	ASP	2.9	X-ray (1)	2.5	-0.4
B2 domain of protein G	207	1IGD	29 (A)	GLU	4.2	X-ray (1)	4.08	-0.12
B2 domain of protein G	207	1IGD	32 (A)	GLU	4.6	X-ray (1)	3.66	-0.94
B2 domain of protein G	207	1IGD	33 (A)	LYS	11	X-ray (1)	10.73	-0.27
B2 domain of protein G	207	1IGD	41 (A)	ASP	3.9	X-ray (1)	3.28	-0.62
B2 domain of protein G	207	1IGD	45 (A)	ASP	4	X-ray (1)	3.86	-0.14
B2 domain of protein G	207	1IGD	51 (A)	ASP	3.6	X-ray (1)	2.28	-1.32
B2 domain of protein G	207	1IGD	52 (A)	ASP	3.4	X-ray (1)	2.71	-0.69
B2 domain of protein G	207	1IGD	61 (A)	GLU	4.2	X-ray (1)	4.75	0.55
Bull	208	2BUS	9 (A)	GLU	4.3	NMR (1)	4.38	0.08

seminal inhibitor IIA								
Bull seminal inhibitor IIA	208	2BUS	20 (A)	GLU	4.1	NMR (1)	4.53	0.43
Bull seminal inhibitor IIA	208	2BUS	6 (A)	ASP	4	NMR (1)	3.9	-0.1
Bull seminal inhibitor IIA	208	2BUS	12 (A)	ASP	3.6	NMR (1)	4.01	0.41
Bacterial proteinase inhibitor Ssi	209	2SIC	43 (I)	HIS	3.2	X-ray (1)	4.45	1.25
Bacterial proteinase inhibitor Ssi	209	2SIC	106 (I)	HIS	6	X-ray (1)	5.68	-0.32
Calbindin D9k	210	1IG5	1 (A)	LYS	10.6	X-ray (1)	11.11	0.51
Calbindin D9k	210	1IG5	4 (A)	GLU	3.8	X-ray (1)	4.26	0.46
Calbindin D9k	210	1IG5	5 (A)	GLU	3.4	X-ray (1)	4.11	0.71
Calbindin D9k	210	1IG5	7 (A)	LYS	11.2	X-ray (2)	10.46	-0.74
Calbindin D9k	210	1IG5	11 (A)	GLU	4.7	X-ray (1)	4.04	-0.66
Calbindin D9k	210	1IG5	12 (A)	LYS	11.1	X-ray (1)	10.46	-0.64
Calbindin D9k	210	1IG5	16 (A)	LYS	10.9	X-ray (2)	10.89	-0.01
Calbindin D9k	210	1IG5	17 (A)	GLU	3.62	X-ray (1)	4.42	0.8
Calbindin D9k	210	1IG5	25 (A)	LYS	11.7	X-ray (1)	10.65	-1.05
Calbindin D9k	210	1IG5	26 (A)	GLU	4.1	X-ray (1)	4.23	0.13
Calbindin D9k	210	1IG5	29 (A)	LYS	11.4	X-ray (1)	10.51	-0.89
Calbindin D9k	210	1IG5	41 (A)	LYS	10.8	X-ray (2)	10.4	-0.4
Calbindin D9k	210	1IG5	47 (A)	ASP	3	X-ray (1)	3.35	0.35
Calbindin D9k	210	1IG5	48 (A)	GLU	4.6	X-ray (1)	4.4	-0.2
Calbindin D9k	210	1IG5	55 (A)	LYS	11.8	X-ray (2)	10.52	-1.28
Calbindin D9k	210	1IG5	64 (A)	GLU	3.8	X-ray (1)	4.19	0.39
Calbindin D9k	210	1IG5	71 (A)	LYS	10.7	X-ray (1)	10.16	-0.54
Calbindin D9k	210	1IG5	72 (A)	LYS	11.3	X-ray (1)	10.8	-0.5
Cardiotoxin	211	1KXI	4 (A)	HIS	5.6	X-ray (1)	6.5	0.9

A5								
Cardiotoxin A5	211	1KXI	17 (A)	GLU	4	X-ray (1)	4.28	0.28
Cardiotoxin A5	211	1KXI	42 (A)	ASP	3.2	X-ray (1)	3.29	0.09
Cardiotoxin A5	211	1KXI	59 (A)	ASP	2.3	X-ray (1)	3.4	1.1
CD2d1	212	1HNG	2 (A)	ASP	3.5	X-ray (1)	3.22	-0.28
CD2d1	212	1HNG	25 (A)	ASP	3.53	X-ray (1)	3.84	0.31
CD2d1	212	1HNG	26 (A)	ASP	3.58	X-ray (1)	3.97	0.39
CD2d1	212	1HNG	28 (A)	ASP	3.57	X-ray (1)	4.07	0.5
CD2d1	212	1HNG	29 (A)	GLU	4.51	X-ray (1)	3.81	-0.7
CD2d1	212	1HNG	33 (A)	GLU	4.2	X-ray (1)	3.95	-0.25
CD2d1	212	1HNG	41 (A)	GLU	6.7	X-ray (2)	4.26	-2.44
CD2d1	212	1HNG	56 (A)	GLU	3.95	X-ray (1)	3.87	-0.08
CD2d1	212	1HNG	62 (A)	ASP	4.18	X-ray (1)	4.49	0.31
CD2d1	212	1HNG	71 (A)	ASP	3.2	X-ray (1)	3.87	0.67
CD2d1	212	1HNG	72 (A)	ASP	4.14	X-ray (1)	3.3	-0.84
CD2d1	212	1HNG	94 (A)	ASP	3.83	X-ray (1)	4.45	0.62
CD2d1	212	1HNG	99 (A)	GLU	4.1	X-ray (1)	3.85	-0.25
Chymotrypsinogen	213	2TGA	40 (A)	HIS	4.6	X-ray (1)	5.37	0.77
Chymotrypsinogen	213	2TGA	57 (A)	HIS	7.3	X-ray (1)	5.62	-1.68
Cyclophilin	214	2CPL	54 (A)	HIS	4.2	X-ray (1)	4.9	0.7
Cyclophilin	214	2CPL	70 (A)	HIS	5.8	X-ray (1)	5.87	0.07
Cyclophilin	214	2CPL	92 (A)	HIS	4.2	X-ray (1)	3.92	-0.28
Cyclophilin	214	2CPL	126 (A)	HIS	6.3	X-ray (1)	5.89	-0.41
Epidermal growth factor (mouse: EGF)	215	1EGF	11 (A)	ASP	3.9	NMR (1)	3.86	-0.04
Epidermal growth factor (mouse: EGF)	215	1EGF	24 (A)	GLU	4.1	NMR (1)	4.31	0.21
Epidermal growth factor (mouse: EGF)	215	1EGF	27 (A)	ASP	4	NMR (1)	3.94	-0.06
Epidermal growth factor (mouse: EGF)	215	1EGF	40 (A)	ASP	3.6	NMR (1)	3.64	0.04
Epidermal growth factor (mouse: EGF)	215	1EGF	46 (A)	ASP	3.8	NMR (1)	4.06	0.26
Epidermal growth factor (mouse: EGF)	215	1EGF	51 (A)	GLU	4	NMR (1)	4.48	0.48

FKBP	214	1FKS	25 (A)	HIS	3.6	X-ray (1)	6.25	2.65
FKBP	214	1FKS	87 (A)	HIS	6.5	X-ray (1)	6.7	0.2
FKBP	214	1FKS	94 (A)	HIS	5.8	X-ray (1)	6.29	0.49
Fungal beta cryptogein	216	1BEO	21 (A)	ASP	2.5	X-ray (1)	2.25	-0.25
Fungal beta cryptogein	216	1BEO	30 (A)	ASP	2.51	X-ray (1)	2.76	0.25
Fungal beta cryptogein	216	1BEO	61 (A)	LYS	10.1	X-ray (1)	9.91	-0.19
Fungal beta cryptogein	216	1BEO	72 (A)	ASP	2.61	X-ray (1)	3.56	0.95
Fungal beta cryptogein	216	1BEO	94 (A)	LYS	9.4	X-ray (1)	10	0.6
Fungal Ribonuclease alpha-sarcin	217	1DE3	9 (A)	ASP	3.9	NMR (1)	3.7	-0.2
Fungal Ribonuclease alpha-sarcin	217	1DE3	19 (A)	GLU	4.6	NMR (1)	4.36	-0.24
Fungal Ribonuclease alpha-sarcin	217	1DE3	31 (A)	GLU	4.6	NMR (1)	3.91	-0.69
Fungal Ribonuclease alpha-sarcin	217	1DE3	36 (A)	HIS	6.8	NMR (2)	6.36	-0.44
Fungal Ribonuclease alpha-sarcin	217	1DE3	41 (A)	ASP	3	NMR (1)	3.55	0.55
Fungal Ribonuclease alpha-sarcin	217	1DE3	50 (A)	HIS	7.7	NMR (1)	6.06	-1.64
Fungal Ribonuclease alpha-sarcin	217	1DE3	57 (A)	ASP	4.3	NMR (2)	4.04	-0.26
Fungal Ribonuclease alpha-sarcin	217	1DE3	59 (A)	ASP	4.1	NMR (2)	3.74	-0.36
Fungal Ribonuclease alpha-sarcin	217	1DE3	75 (A)	ASP	3.9	NMR (1)	4.66	0.76
Fungal Ribonuclease alpha-sarcin	217	1DE3	77 (A)	ASP	3	NMR (1)	3.66	0.66
Fungal Ribonuclease alpha-sarcin	217	1DE3	85 (A)	ASP	3.8	NMR (1)	4.44	0.64
Fungal Ribonucleas	217	1DE3	91 (A)	ASP	3	NMR (1)	3.43	0.43

e alpha-sarcin								
Fungal Ribonuclease alpha-sarcin	217	1DE3	96 (A)	GLU	5.1	NMR (1)	6.29	1.19
Fungal Ribonuclease alpha-sarcin	217	1DE3	102 (A)	ASP	3	NMR (1)	3.96	0.96
Fungal Ribonuclease alpha-sarcin	217	1DE3	104 (A)	HIS	6.5	NMR (2)	6.08	-0.42
Fungal Ribonuclease alpha-sarcin	217	1DE3	105 (A)	ASP	3	NMR (1)	4.01	1.01
Fungal Ribonuclease alpha-sarcin	217	1DE3	109 (A)	ASP	3.7	NMR (1)	4.01	0.31
Fungal Ribonuclease alpha-sarcin	217	1DE3	115 (A)	GLU	4.9	NMR (1)	4.11	-0.79
Fungal Ribonuclease alpha-sarcin	217	1DE3	137 (A)	HIS	5.8	NMR (2)	5.15	-0.65
Fungal Ribonuclease alpha-sarcin	217	1DE3	140 (A)	GLU	4.3	NMR (1)	4.01	-0.29
Fungal Ribonuclease alpha-sarcin	217	1DE3	144 (A)	GLU	4.3	NMR (1)	4.01	-0.29
Hen egg white lysozome	218	4LZT	1 (A)	LYS	10.8	X-ray (1)	10.4	-0.4
Hen egg white lysozome	150	4LZT	7 (A)	GLU	2.9	X-ray (2)	4.07	1.17
Hen egg white lysozome	218	4LZT	13 (A)	LYS	10.5	X-ray (1)	10.71	0.21
Hen egg white lysozome	150	4LZT	15 (A)	HIS	5.4	X-ray (2)	6.49	1.09
Hen egg white lysozome	150	4LZT	18 (A)	ASP	2.7	X-ray (2)	3.33	0.63
Hen egg white lysozome	218	4LZT	33 (A)	LYS	10.4	X-ray (2)	10.56	0.16
Hen egg white lysozome	150	4LZT	35 (A)	GLU	6.2	X-ray (2)	4.24	-1.96

Hen egg white lysozome	150	4LZT	48 (A)	ASP	1.6	X-ray (1)	2.1	0.5
Hen egg white lysozome	150	4LZT	52 (A)	ASP	3.7	X-ray (2)	3.56	-0.14
Hen egg white lysozome	150	4LZT	66 (A)	ASP	0.9	X-ray (1)	1.75	0.85
Hen egg white lysozome	150	4LZT	87 (A)	ASP	2.1	X-ray (2)	3.32	1.22
Hen egg white lysozome	218	4LZT	96 (A)	LYS	10.8	X-ray (2)	10.15	-0.65
Hen egg white lysozome	218	4LZT	97 (A)	LYS	10.3	X-ray (1)	10.69	0.39
Hen egg white lysozome	150	4LZT	101 (A)	ASP	4.08	X-ray (1)	3.91	-0.17
Hen egg white lysozome	218	4LZT	116 (A)	LYS	10.2	X-ray (1)	10.41	0.21
Hen egg white lysozome	150	4LZT	119 (A)	ASP	3.2	X-ray (1)	2.65	-0.55
Hirudin	219	1HIC	5 (A)	ASP	4.3	NMR (1)	4.21	-0.09
Hirudin	219	1HIC	8 (A)	GLU	4.3	NMR (1)	4.08	-0.22
Hirudin	219	1HIC	17 (A)	GLU	3.8	NMR (1)	4.12	0.32
Hirudin	219	1HIC	35 (A)	GLU	4.3	NMR (1)	4.29	-0.01
Hirudin	219	1HIC	43 (A)	GLU	4.2	NMR (1)	4.27	0.07
HIV-1 protease	220	1HPX	25 (A)	ASP	6.2	X-ray (1)	2.89	-3.31
HIV-1 protease	220	1HPX	29 (A)	ASP	3.2	X-ray (1)	3.24	0.04
HIV-1 protease	220	1HPX	30 (A)	ASP	3.9	X-ray (1)	4.11	0.21
HIV-1 protease	220	1HPX	60 (A)	ASP	3	X-ray (1)	2.84	-0.16
Human DNA polymerase lambdalyase domain	221	1NZP	312 (A)	LYS	9.5	X-ray (1)	9.3	-0.2
Human insulin	222	1MHI	13 (A)	GLU	2.2	NMR (1)	3.98	1.78
Human thioredoxin (ox)	223	1ERU	6 (A)	GLU	4.9	X-ray (1)	4.31	-0.59
Human thioredoxin (ox)	223	1ERU	13 (A)	GLU	4.4	X-ray (1)	4.17	-0.23
Human thioredoxin (ox)	223	1ERU	16 (A)	ASP	4.2	X-ray (1)	3.81	-0.39
Human thioredoxin	223	1ERU	20 (A)	ASP	3.8	X-ray (1)	3.87	0.07

(ox)								
Human thioredoxin (ox)	[223]	1ERU	26 (A)	ASP	8.1	X-ray (1)	7.45	-0.65
Human thioredoxin (ox)	[223]	1ERU	47 (A)	GLU	4.3	X-ray (1)	4.14	-0.16
Human thioredoxin (ox)	[223]	1ERU	56 (A)	GLU	3.2	X-ray (1)	4.57	1.37
Human thioredoxin (ox)	[223]	1ERU	58 (A)	ASP	2.7	X-ray (1)	4.33	1.63
Human thioredoxin (ox)	[223]	1ERU	60 (A)	ASP	3.9	X-ray (1)	4.11	0.21
Human thioredoxin (ox)	[223]	1ERU	61 (A)	ASP	5.2	X-ray (1)	4.01	-1.19
Human thioredoxin (ox)	[223]	1ERU	64 (A)	ASP	3.2	X-ray (1)	4.06	0.86
Human thioredoxin (ox)	[223]	1ERU	68 (A)	GLU	5.1	X-ray (1)	4.38	-0.72
Human thioredoxin (ox)	[223]	1ERU	70 (A)	GLU	4.8	X-ray (1)	4.27	-0.53
Human thioredoxin (ox)	[223]	1ERU	88 (A)	GLU	3.6	X-ray (1)	4.33	0.73
Human thioredoxin (ox)	[223]	1ERU	95 (A)	GLU	4.1	X-ray (1)	4.24	0.14
Human thioredoxin (ox)	[223]	1ERU	98 (A)	GLU	3.9	X-ray (1)	4.05	0.15
Human thioredoxin (ox)	[223]	1ERU	103 (A)	GLU	4.5	X-ray (1)	4.23	-0.27
Human thioredoxin (red)	[223]	1ERT	6 (A)	GLU	4.8	X-ray (1)	4.37	-0.43
Human thioredoxin (red)	[223]	1ERT	13 (A)	GLU	4.4	X-ray (1)	4.15	-0.25
Human thioredoxin (red)	[223]	1ERT	16 (A)	ASP	4	X-ray (1)	3.81	-0.19
Human thioredoxin (red)	[223]	1ERT	20 (A)	ASP	3.8	X-ray (1)	3.8	0
Human thioredoxin (red)	[223]	1ERT	26 (A)	ASP	9.9	X-ray (1)	7.21	-2.69
Human thioredoxin (red)	[223]	1ERT	43 (A)	HIS	5.5	X-ray (1)	6.77	1.27
Human	[223]	1ERT	47 (A)	GLU	4.1	X-ray (1)	4.08	-0.02

thioredoxin (red)								
Human thioredoxin (red)	223	1ERT	56 (A)	GLU	3.1	X-ray (1)	4.19	1.09
Human thioredoxin (red)	223	1ERT	58 (A)	ASP	2.8	X-ray (1)	4.43	1.63
Human thioredoxin (red)	223	1ERT	60 (A)	ASP	4.2	X-ray (1)	4.13	-0.07
Human thioredoxin (red)	223	1ERT	61 (A)	ASP	5.3	X-ray (1)	4.07	-1.23
Human thioredoxin (red)	223	1ERT	64 (A)	ASP	3.2	X-ray (1)	4.04	0.84
Human thioredoxin (red)	223	1ERT	68 (A)	GLU	4.9	X-ray (1)	4.33	-0.57
Human thioredoxin (red)	223	1ERT	70 (A)	GLU	4.6	X-ray (1)	4.27	-0.33
Human thioredoxin (red)	223	1ERT	88 (A)	GLU	3.7	X-ray (1)	4.09	0.39
Human thioredoxin (red)	223	1ERT	95 (A)	GLU	4.1	X-ray (1)	4.12	0.02
Human thioredoxin (red)	223	1ERT	98 (A)	GLU	3.9	X-ray (1)	3.85	-0.05
Human thioredoxin (red)	223	1ERT	103 (A)	GLU	4.4	X-ray (1)	4.35	-0.05
Myoglobin horse	224	1DWR	24 (A)	HIS	4.8	X-ray (1)	5.44	0.64
Myoglobin horse	224	1DWR	36 (A)	HIS	7.8	X-ray (1)	6.19	-1.61
Myoglobin horse	224	1DWR	48 (A)	HIS	5.62	X-ray (1)	6.48	0.86
Myoglobin horse	224	1DWR	81 (A)	HIS	6.94	X-ray (1)	6.38	-0.56
Myoglobin horse	224	1DWR	113 (A)	HIS	5.87	X-ray (1)	6.16	0.29
Myoglobin horse	224	1DWR	116 (A)	HIS	6.79	X-ray (1)	6.02	-0.77
Myoglobin horse	224	1DWR	119 (A)	HIS	6.56	X-ray (1)	6.29	-0.27
Myoglobin sperm whale	224	1A6K	12 (A)	HIS	6.5	X-ray (1)	6.81	0.31
Myoglobin sperm whale	224	1A6K	24 (A)	HIS	5	X-ray (1)	5.41	0.41
Myoglobin sperm whale	224	1A6K	36 (A)	HIS	8	X-ray (1)	6.28	-1.72

Myoglobin sperm whale	224	1A6K	48 (A)	HIS	5.6	X-ray (1)	6.42	0.82
Myoglobin sperm whale	225	1A6K	64 (A)	HIS	5	X-ray (1)	5.71	0.71
Myoglobin sperm whale	224	1A6K	81 (A)	HIS	6.9	X-ray (1)	6.71	-0.19
Myoglobin sperm whale	225	1A6K	82 (A)	HIS	5	X-ray (1)	5.05	0.05
Myoglobin sperm whale	225	1A6K	97 (A)	HIS	5.6	X-ray (1)	6.15	0.55
Myoglobin sperm whale	224	1A6K	113 (A)	HIS	5.4	X-ray (1)	6.08	0.68
Myoglobin sperm whale	224	1A6K	116 (A)	HIS	6.7	X-ray (1)	6.15	-0.55
Myoglobin sperm whale	224	1A6K	119 (A)	HIS	6.2	X-ray (1)	6.02	-0.18
Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	3 (A)	ASP	3.55	X-ray (1)	3.82	0.27
Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	7 (A)	GLU	3.85	X-ray (1)	4.5	0.65
Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	15 (A)	LYS	10.4	X-ray (1)	10.42	0.02
Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	26 (A)	LYS	10.4	X-ray (1)	10.39	-0.01
Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	41 (A)	LYS	10.8	X-ray (1)	10.85	0.05
Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	46 (A)	LYS	10.4	X-ray (1)	10.49	0.09
Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	49 (A)	GLU	3.91	X-ray (1)	4.24	0.33

Pancreatic trypsin inhibitor precursor (BPTI)	226	4PTI	50 (A)	ASP	3.2	X-ray (1)	3.46	0.26
Phage T4 lysozyme	227	2LZM	31 (A)	HIS	9.1	X-ray (2)	6.29	-2.81
Phage T4 lysozyme mutant	228	1L54	102 (A)	LYS	6.6	X-ray (2)	8.54	1.94
Phosphocarrier protein	94	1POH	76 (A)	HIS	6	X-ray (1)	6.43	0.43
Phosphatidylinositol	229	1GYM	32 (A)	HIS	7.6	X-ray (1)	4.84	-2.76
Phosphatidylinositol	229	1GYM	61 (A)	HIS	3	X-ray (1)	6.6	3.6
Phosphatidylinositol	229	1GYM	81 (A)	HIS	3	X-ray (1)	5.32	2.32
Phosphatidylinositol	229	1GYM	82 (A)	HIS	6.9	X-ray (1)	5.67	-1.23
Phosphatidylinositol	94	1GYM	92 (A)	HIS	5.4	X-ray (1)	5.88	0.48
Phosphatidylinositol	94	1GYM	227 (A)	HIS	6.9	X-ray (1)	6.24	-0.66
Ribonuclease H1	151	2RN2	6 (A)	GLU	4.5	X-ray (1)	3.7	-0.8
Ribonuclease H1	151	2RN2	10 (A)	ASP	6.1	X-ray (2)	4.51	-1.59
Ribonuclease H1	151	2RN2	32 (A)	GLU	3.6	X-ray (1)	3.84	0.24
Ribonuclease H1	151	2RN2	48 (A)	GLU	4.4	X-ray (1)	3.58	-0.82
Ribonuclease H1	151	2RN2	57 (A)	GLU	3.2	X-ray (2)	3.77	0.57
Ribonuclease H1	151	2RN2	61 (A)	GLU	3.9	X-ray (2)	3.83	-0.07
Ribonuclease H1	151	2RN2	62 (A)	HIS	7	X-ray (2)	6.99	-0.01
Ribonuclease H1	151	2RN2	64 (A)	GLU	4.4	X-ray (1)	3.98	-0.42
Ribonuclease H1	151	2RN2	70 (A)	ASP	2.6	X-ray (1)	3.63	1.03
Ribonuclease H1	151	2RN2	83 (A)	HIS	5.5	X-ray (1)	6.38	0.88
Ribonuclease H1	151	2RN2	94 (A)	ASP	3.2	X-ray (2)	2.61	-0.59
Ribonuclease H1	151	2RN2	102 (A)	ASP	2	X-ray (1)	2.91	0.91
Ribonuclease H1	151	2RN2	108 (A)	ASP	3.2	X-ray (1)	3.54	0.34
Ribonuclease H1	151	2RN2	114 (A)	HIS	5	X-ray (1)	5.14	0.14
Ribonuclease H1	151	2RN2	119 (A)	GLU	4.1	X-ray (2)	4.03	-0.07
Ribonuclease H1	151	2RN2	124 (A)	HIS	7.1	X-ray (2)	6.7	-0.4
Ribonuclease H1	151	2RN2	127 (A)	HIS	7.9	X-ray (2)	6.98	-0.92

Ribonuclease H1	151	2RN2	129 (A)	GLU	3.6	X-ray (2)	3.8	0.2
Ribonuclease H1	151	2RN2	131 (A)	GLU	4.3	X-ray (1)	4.27	-0.03
Ribonuclease H1	151	2RN2	134 (A)	ASP	4.1	X-ray (1)	4.46	0.36
Ribonuclease H1	151	2RN2	135 (A)	GLU	4.3	X-ray (1)	4.17	-0.13
Ribonuclease H1	151	2RN2	147 (A)	GLU	4.2	X-ray (1)	4.25	0.05
Ribonuclease H1	151	2RN2	148 (A)	ASP	2	X-ray (1)	1.3	-0.7
Ribonuclease H1	151	2RN2	154 (A)	GLU	4.4	X-ray (1)	4.06	-0.34
Ribonuclease A	230	3RN3	2 (A)	GLU	2.6	X-ray (2)	3.9	1.3
Ribonuclease A	-36	3RN3	9 (A)	GLU	4	X-ray (2)	4.06	0.06
Ribonuclease A	230	3RN3	12 (A)	HIS	6	X-ray (2)	5.41	-0.59
Ribonuclease A	230	3RN3	14 (A)	ASP	1.8	X-ray (2)	1.53	-0.27
Ribonuclease A	230	3RN3	38 (A)	ASP	3.5	X-ray (1)	3.8	0.3
Ribonuclease A	230	3RN3	48 (A)	HIS	6.1	X-ray (2)	5.24	-0.86
Ribonuclease A	230	3RN3	49 (A)	GLU	4.7	X-ray (1)	3.96	-0.74
Ribonuclease A	230	3RN3	53 (A)	ASP	3.7	X-ray (2)	4.07	0.37
Ribonuclease A	230	3RN3	83 (A)	ASP	3.3	X-ray (1)	3.33	0.03
Ribonuclease A	230	3RN3	86 (A)	GLU	4	X-ray (1)	4.09	0.09
Ribonuclease A	230	3RN3	105 (A)	HIS	6.5	X-ray (1)	6.39	-0.11
Ribonuclease A	231	3RN3	111 (A)	GLU	3.5	X-ray (2)	4.01	0.51
Ribonuclease A	230	3RN3	119 (A)	HIS	6.5	X-ray (2)	6.25	-0.25
Ribonuclease A	230	3RN3	121 (A)	ASP	3.1	X-ray (2)	3.32	0.22
Ribonuclease SA	172	1RGG	1 (A)	ASP	3.44	X-ray (1)	3.89	0.45
Ribonuclease SA	172	1RGG	14 (A)	GLU	5.05	X-ray (1)	4.5	-0.55
Ribonuclease SA	172	1RGG	17 (A)	ASP	3.72	X-ray (1)	4.19	0.47
Ribonuclease SA	172	1RGG	25 (A)	ASP	4.87	X-ray (1)	4.02	-0.85
Ribonuclease SA	172	1RGG	33 (A)	ASP	2.39	X-ray (1)	4.16	1.77
Ribonuclease SA	172	1RGG	41 (A)	GLU	4.14	X-ray (1)	4.47	0.33
Ribonuclease SA	172	1RGG	53 (A)	HIS	8.27	X-ray (1)	6.11	-2.16
Ribonuclease SA	172	1RGG	54 (A)	GLU	3.42	X-ray (1)	3.42	0

Ribonuclease SA	172	1RGG	74 (A)	GLU	3.47	X-ray (1)	4.28	0.81
Ribonuclease SA	172	1RGG	78 (A)	GLU	3.13	X-ray (1)	4.14	1.01
Ribonuclease SA	172	1RGG	79 (A)	ASP	7.37	X-ray (1)	5.36	-2.01
Ribonuclease SA	172	1RGG	84 (A)	ASP	3.01	X-ray (1)	1.94	-1.07
Ribonuclease SA	172	1RGG	85 (A)	HIS	6.35	X-ray (1)	6.68	0.33
Ribonuclease SA	172	1RGG	93 (A)	ASP	3.09	X-ray (1)	2.72	-0.37
Ribonuclease T1	232	1I0V	15 (A)	ASP	3.52	X-ray (1)	3.9	0.38
Ribonuclease T1	232	1I0V	27 (A)	HIS	7	X-ray (1)	6.8	-0.2
Ribonuclease T1	232	1I0V	28 (A)	GLU	5.9	X-ray (2)	4.49	-1.41
Ribonuclease T1	232	1I0V	29 (A)	ASP	4.26	X-ray (1)	3.99	-0.27
Ribonuclease T1	232	1I0V	31 (A)	GLU	5.36	X-ray (1)	4.33	-1.03
Ribonuclease T1	233	1I0V	40 (A)	HIS	7.9	X-ray (2)	6.47	-1.43
Ribonuclease T1	232	1I0V	46 (A)	GLU	3.62	X-ray (1)	4.51	0.89
Ribonuclease T1	232	1I0V	58 (A)	GLU	3.96	X-ray (1)	4.01	0.05
Ribonuclease T1	232	1I0V	66 (A)	ASP	3.9	X-ray (1)	4.16	0.26
Ribonuclease T1	157	1I0V	76 (A)	ASP	0.5	X-ray (1)	3.71	3.21
Ribonuclease T1	232	1I0V	82 (A)	GLU	3.27	X-ray (1)	3.9	0.63
Ribonuclease T1	233	1I0V	92 (A)	HIS	7.8	X-ray (2)	6.23	-1.57
Ribonuclease T1	232	1I0V	102 (A)	GLU	5.3	X-ray (1)	4.27	-1.03
Sea anemone neurotoxin	234	1ANS	20 (A)	GLU	5.4	NMR (1)	4.34	-1.06
Staph nuclease variant Delta+PHS	149	3BDC	10 (A)	GLU	2.8	X-ray (1)	4.33	1.53
Staph nuclease variant Delta+PHS	149	3BDC	19 (A)	ASP	2.2	X-ray (1)	2.81	0.61
Staph nuclease variant Delta+PHS	235	3BDC	20 (A)	GLU *	4.5	X-ray (1)	5.86	1.36
Staph nuclease variant Delta+PHS	236	3BDC	20 (A)	LYS *	10.4	X-ray (1)	8.74	-1.66
Staph	149	3BDC	21 (A)	ASP	6.5	X-ray (1)	3.79	-2.71

nuclease variant Delta+PHS								
Staph nuclease variant Delta+PHS	235	3BDC	23 (A)	GLU *	7.1	X-ray (1)	8.86	1.76
Staph nuclease variant Delta+PHS	236	3BDC	23 (A)	LYS *	7.3	X-ray (1)	7.36	0.06
Staph nuclease variant Delta+PHS	235	3BDC	34 (A)	GLU *	7.3	X-ray (1)	6.65	-0.65
Staph nuclease variant Delta+PHS	235	3BDC	36 (A)	GLU *	8.7	X-ray (1)	9.16	0.46
Staph nuclease variant Delta+PHS	235	3BDC	37 (A)	GLU *	5.2	X-ray (1)	5.67	0.47
Staph nuclease variant Delta+PHS	236	3BDC	37 (A)	LYS *	10.4	X-ray (1)	8.48	-1.92
Staph nuclease variant Delta+PHS	235	3BDC	39 (A)	GLU *	8.2	X-ray (1)	8.68	0.48
Staph nuclease variant Delta+PHS	236	3BDC	39 (A)	LYS *	9	X-ray (1)	8.28	-0.72
Staph nuclease variant Delta+PHS	149	3BDC	40 (A)	ASP	3.9	X-ray (1)	3.96	0.06
Staph nuclease variant Delta+PHS	235	3BDC	41 (A)	GLU *	6.8	X-ray (1)	6.5	-0.3
Staph nuclease variant Delta+PHS	236	3BDC	41 (A)	LYS *	9.3	X-ray (1)	9.67	0.37
Staph nuclease variant Delta+PHS	149	3BDC	43 (A)	GLU	4.3	X-ray (1)	4.13	-0.17
Staph nuclease variant Delta+PHS	149	3BDC	52 (A)	GLU	3.9	X-ray (1)	4.19	0.29
Staph nuclease variant Delta+PHS	149	3BDC	57 (A)	GLU	3.5	X-ray (1)	4.29	0.79

Staph nuclease variant Delta+PHS	235	3BDC	58 (A)	GLU *	7.7	X-ray (1)	6.56	-1.14
Staph nuclease variant Delta+PHS	236	3BDC	58 (A)	LYS *	10.4	X-ray (1)	8.99	-1.41
Staph nuclease variant Delta+PHS	235	3BDC	62 (A)	GLU *	7.7	X-ray (1)	7.51	-0.19
Staph nuclease variant Delta+PHS	149	3BDC	67 (A)	GLU	3.8	X-ray (1)	4.05	0.25
Staph nuclease variant Delta+PHS	149	3BDC	73 (A)	GLU	3.3	X-ray (1)	4.06	0.76
Staph nuclease variant Delta+PHS	235	3BDC	74 (A)	GLU *	7.8	X-ray (1)	6.7	-1.1
Staph nuclease variant Delta+PHS	149	3BDC	75 (A)	GLU	3.3	X-ray (1)	4.04	0.74
Staph nuclease variant Delta+PHS	149	3BDC	77 (A)	ASP	2.2	X-ray (1)	2.08	-0.12
Staph nuclease variant Delta+PHS	149	3BDC	83 (A)	ASP	2.2	X-ray (1)	1.6	-0.6
Staph nuclease variant Delta+PHS	236	3BDC	90 (A)	LYS *	8.6	X-ray (1)	8.81	0.21
Staph nuclease variant Delta+PHS	235	3BDC	90 (A)	GLU *	6.4	X-ray (1)	6.82	0.42
Staph nuclease variant Delta+PHS	236	3BDC	91 (A)	LYS *	9	X-ray (1)	8.16	-0.84
Staph nuclease variant Delta+PHS	149	3BDC	95 (A)	ASP	2.2	X-ray (1)	2.75	0.55
Staph nuclease variant Delta+PHS	235	3BDC	99 (A)	GLU *	8.4	X-ray (1)	7.11	-1.29
Staph nuclease variant	235	3BDC	100 (A)	GLU *	7.6	X-ray (1)	8.33	0.73

Delta+PHS								
Staph nuclease variant Delta+PHS	236	3BDC	100 (A)	LYS *	8.6	X-ray (1)	7.73	-0.87
Staph nuclease variant Delta+PHS	149	3BDC	101 (A)	GLU	3.8	X-ray (1)	3.76	-0.04
Staph nuclease variant Delta+PHS	235	3BDC	103 (A)	GLU *	8.9	X-ray (1)	6.97	-1.93
Staph nuclease variant Delta+PHS	235	3BDC	109 (A)	GLU *	7.9	X-ray (1)	8.46	0.56
Staph nuclease variant Delta+PHS	236	3BDC	109 (A)	LYS *	9.2	X-ray (1)	7.99	-1.21
Staph nuclease variant Delta+PHS	236	3BDC	118 (A)	LYS *	10.4	X-ray (1)	9.78	-0.62
Staph nuclease variant Delta+PHS	235	3BDC	118 (A)	GLU *	4.5	X-ray (1)	5.27	0.77
Staph nuclease variant Delta+PHS	149	3BDC	122 (A)	GLU	3.9	X-ray (1)	3.75	-0.15
Staph nuclease variant Delta+PHS	235	3BDC	125 (A)	GLU *	9.1	X-ray (1)	7.91	-1.19
Staph nuclease variant Delta+PHS	149	3BDC	129 (A)	GLU	3.8	X-ray (1)	3.83	0.03
Staph nuclease variant Delta+PHS	235	3BDC	132 (A)	GLU	7	X-ray (1)	6.48	-0.52
Staph nuclease variant Delta+PHS	236	3BDC	132 (A)	LYS	10.4	X-ray (1)	8.53	-1.87
Staph nuclease variant Delta+PHS	149	3BDC	135 (A)	GLU	3.8	X-ray (1)	3.87	0.07
Staph nuclease variant Delta+PHS	235	3EVQ	25 (A)	GLU	7.5	X-ray (1)	7.91	0.41
Staph nuclease	184	3ERQ	25 (A)	LYS	6.2	X-ray (1)	7.53	1.33

variant Delta+PHS								
Staph nuclease variant Delta+PHS	237	3ITP	34 (A)	LYS	7.1	X-ray (1)	8.87	1.77
Staph nuclease variant Delta+PHS	237	3EJI	36 (A)	LYS	7.2	X-ray (1)	7.3	0.1
Staph nuclease variant Delta+PHS	238	3D6C	38 (A)	GLU	7.2	X-ray (1)	6.04	-1.16
Staph nuclease variant Delta+PHS	239	2RKS	38 (A)	LYS	10.4	X-ray (1)	9.2	-1.2
Staph nuclease variant Delta+PHS	237	3DMU	62 (A)	LYS	8.1	X-ray (1)	7.65	-0.45
Staph nuclease variant Delta+PHS	202	1U9R	66 (A)	GLU	8.5	X-ray (1)	8.1	-0.4
Staph nuclease variant Delta+PHS	240	2OXP	66 (A)	ASP	8.8	X-ray (1)	6.62	-2.18
Staph nuclease variant Delta+PHS	236	2RBM	72 (A)	LYS	8.6	X-ray (1)	10.29	1.69
Staph nuclease variant Delta+PHS	235	3ERO	72 (A)	GLU	7.3	X-ray (1)	4.4	-2.9
Staph nuclease variant Delta+PHS	236	3RUZ	74 (A)	LYS	7.4	X-ray (1)	9.3	1.9
Staph nuclease variant Delta+PHS	235	3D4D	91 (A)	GLU	7.1	X-ray (1)	6.65	-0.45
Staph nuclease variant Delta+PHS	241	1TT2	92 (A)	LYS	5.6	X-ray (1)	7.2	1.6
Staph nuclease variant Delta+PHS	235	2OEO	92 (A)	ASP	7.5	X-ray (1)	9.45	1.95
Staph nuclease variant Delta+PHS	235	1TQO	92 (A)	GLU	9	X-ray (1)	8.76	-0.24
Staph	236	4HMI	99 (A)	LYS	6.5	X-ray (1)	7.98	1.48

nuclease variant Delta+PHS								
Staph nuclease variant Delta+PHS	236	3E5S	103 (A)	LYS	8.2	X-ray (1)	8.39	0.19
Staph nuclease variant Delta+PHS	237	3P75	104 (A)	ASP	9.7	X-ray (1)	6.54	-3.16
Staph nuclease variant Delta+PHS	236	3C1F	104 (A)	LYS	7.7	X-ray (1)	8.57	0.87
Staph nuclease variant Delta+PHS	235	3H6M	104 (A)	GLU	9.4	X-ray (1)	6.4	-3
Staph nuclease variant Delta+PHS	236	3C1E	125 (A)	LYS	6.2	X-ray (1)	8.38	2.18
Staph. Nuclease	242	1STY	8 (A)	HIS	6.52	X-ray (1)	6.38	-0.14
Staph. Nuclease	242	1STY	46 (A)	HIS	5.86	X-ray (1)	6.8	0.94
Staph. Nuclease	242	1STY	121 (A)	HIS	5.3	X-ray (2)	6.2	0.9
Staph. Nuclease	242	1STY	124 (A)	HIS	5.73	X-ray (1)	5.93	0.2
Snake erabutoxin b	243	3EBX	6 (A)	HIS	2.8	X-ray (2)	5.31	2.51
Snake erabutoxin b	243	3EBX	26 (A)	HIS	5.8	X-ray (1)	5.89	0.09
Tyrosine phosphotase	244	1DG9	66 (A)	HIS	8.3	X-ray (1)	6.37	-1.93
Tyrosine phosphotase	244	1DG9	72 (A)	HIS	9.2	X-ray (1)	6.75	-2.45
Turkey ovomucoid inhibitor	245	1PPF	7 (A)	ASP	2.6	X-ray (1)	3.35	0.75
Turkey ovomucoid inhibitor	245	1PPF	10 (A)	GLU	4.1	X-ray (2)	4.3	0.2
Turkey ovomucoid inhibitor	245	1PPF	13 (A)	LYS	9.9	X-ray (1)	11.1	1.2
Turkey ovomucoid inhibitor	245	1PPF	19 (A)	GLU	3.2	X-ray (2)	4.1	0.9
Turkey ovomucoid inhibitor	245	1PPF	27 (A)	ASP	2.2	X-ray (2)	2.34	0.14
Turkey ovomucoid inhibitor	246	1PPF	29 (A)	LYS	11.1	X-ray (1)	10.85	-0.25

Turkey ovomucoid inhibitor	245	1PPF	34 (A)	LYS	10.1	X-ray (2)	10.76	0.66
Turkey ovomucoid inhibitor	245	1PPF	43 (A)	GLU	4.8	X-ray (2)	4.18	-0.62
Turkey ovomucoid inhibitor	247	1PPF	52 (A)	HIS	7.5	X-ray (1)	6.52	-0.98
Turkey ovomucoid inhibitor	246	1PPF	55 (A)	LYS	11.1	X-ray (2)	10.37	-0.73
Xylanase BA	248	1H4G	5 (A)	ASP	3.84	X-ray (1)	3.76	-0.08
Xylanase BA	248	1H4G	11 (A)	HIS	6.5	X-ray (1)	6.03	-0.47
Xylanase BA	248	1H4G	12 (A)	ASP	3.94	X-ray (1)	3.1	-0.84
Xylanase BA	248	1H4G	15 (A)	ASP	3.35	X-ray (1)	3.52	0.17
Xylanase BA	248	1H4G	17 (A)	GLU	4.31	X-ray (1)	3.97	-0.34
Xylanase BA	248	1H4G	21 (A)	ASP	3.46	X-ray (1)	3.11	-0.35
Xylanase BA	248	1H4G	32 (A)	HIS	6.7	X-ray (1)	6.12	-0.58
Xylanase BA	248	1H4G	60 (A)	HIS	4	X-ray (1)	5.24	1.24
Xylanase BA	248	1H4G	90 (A)	ASP	3.88	X-ray (1)	3.73	-0.15
Xylanase BA	248	1H4G	94 (A)	GLU	3.94	X-ray (1)	6.06	2.12
Xylanase BA	248	1H4G	99 (A)	ASP	2.7	X-ray (1)	4.67	1.97
Xylanase BA	248	1H4G	118 (A)	ASP	2.7	X-ray (1)	2.94	0.24
Xylanase BA	248	1H4G	123 (A)	ASP	2.7	X-ray (1)	2.34	-0.36
Xylanase BA	248	1H4G	126 (A)	GLU	4.51	X-ray (1)	4.11	-0.4
Xylanase BA	248	1H4G	162 (A)	HIS	2.7	X-ray (1)	2.23	-0.47
Xylanase BA	248	1H4G	167 (A)	GLU	3.58	X-ray (1)	3.83	0.25
Xylanase BA	248	1H4G	178 (A)	GLU	4.1	X-ray (1)	5.91	1.81
Xylanase BA	248	1H4G	184 (A)	GLU	6.5	X-ray (1)	5.08	-1.42
Xylanase BC	249	1XNB	11 (A)	ASP	2.5	X-ray (1)	2.79	0.29
Xylanase BC	249	1XNB	78 (A)	GLU	4.6	X-ray (1)	5.6	1
Xylanase BC	249	1XNB	101 (A)	ASP	2	X-ray (1)	2.3	0.3
Xylanase BC	249	1XNB	106 (A)	ASP	2.7	X-ray (2)	4.19	1.49
Xylanase BC	249	1XNB	119 (A)	ASP	3.2	X-ray (1)	2.66	-0.54

Xylanase BC	249	1XNB	121 (A)	ASP	3.6	X-ray (1)	3.83	0.23
Xylanase BC	249	1XNB	149 (A)	HIS	2.3	X-ray (1)	2.75	0.45
Xylanase BC	249	1XNB	156 (A)	HIS	6.5	X-ray (1)	6.54	0.04
Xylanase BC	249	1XNB	172 (A)	GLU	6.7	X-ray (1)	4.8	-1.9

* indicates a mutant residues, modeled using MODELLER.

Table 5.3: Benchmarking of $\mathrm{pK_a}$ prediction using DEPTH and other methods on a testing set of 60 ionizable residues.

PDB Code	Residue	pK_a^{exp}	Error = $pK_a^{exp} - pK_a^{pred}$								
			MD/GB/TI w/ waters	MD/GB/TI w/o waters	PROPKA3.0	GDDM	MM-SCP	EGAD	MCCE	QM/MM	DEPTH
3RN3	ASP14	-2.2	1.2	1.9	0.4	0.4	0.6	NA	1.2	NA	0.1
4LZT	ASP87	-1.9	NA	1.1	1.4	0.9	1.1	0.8	-0.9	NA	1.2
1PPF	ASP27	-1.8	2.0	2.4	0.5	1.2	1.7	0.8	1.1	-0.3	0.2
1XNB	ASP11	-1.5	NA	1.4	0.4	0.7	NA	1.1	NA	NA	0.3
1BEO	ASP21	-1.5	NA	0.4	-1.0	NA	NA	NA	2.6	0.0	-0.2
4LZT	ASP18	-1.3	NA	0.5	0.8	1.1	0.8	0.8	0.3	NA	0.5
1XNB	ASP106	-1.3	NA	0.1	1.1	0.8	NA	0.8	NA	NA	1.7
1PGA	ASP22	-1.1	NA	0.7	-0.2	1.3	0.0	0.6	-0.7	NA	0.2
3RN3	ASP121	-0.9	1.9	1.9	0.0	-0.9	0.8	NA	0.1	NA	0.7
1A2P	ASP75	-0.9	NA	1.2	1.0	-0.7	NA	3.2	1.4	NA	2.3
2RN2	ASP94	-0.8	NA	1.1	-0.4	-0.5	0.3	NA	0.6	NA	-0.6
1PGA	ASP47	-0.6	NA	0.5	-1.2	0.5	-0.8	-0.1	-1.1	NA	-0.2
3RN3	ASP53	-0.3	NA	1.3	0.3	-0.2	0.3	NA	0.0	NA	0.3
4LZT	ASP52	-0.3	2.4	2.3	-0.1	1.1	-0.2	-0.1	0.2	NA	0.1
1PGA	ASP36	-0.2	NA	0.8	0.3	0.1	0.6	0.7	1.2	NA	0.0

2TRX	ASP20	-0.2	NA	0.2	0.2	NA	NA	NA	NA	NA	-0.1
1DE3	ASP59	0.1	NA	-0.2	-0.4	0.7	NA	-1.7	NA	NA	-0.2
1DE3	ASP57	0.3	NA	-0.4	-0.3	-0.8	NA	-0.9	NA	NA	-0.3
2RN2	ASP10	2.1	NA	0.6	1.0	NA	-0.2	NA	4.3	NA	-1.4
2TRX	ASP26	4.1	NA	-0.4	-1.3	-2.2	NA	NA	NA	NA	-0.5
RMSD (N)			1.9 (4)	1.2 (20)	0.7 (20)	1.0 (17)	0.8 (12)	1.2 (13)	1.6 (14)	0.2 (2)	0.8 (20)
MAD			1.9	1.0	0.6	0.8	0.6	1.0	1.0	0.3	0.6
MAX			2.4	2.4	1.4	2.2	1.7	3.2	4.3	0.3	2.3

PDB Code	Residue	pK_a^{exp}	Error = $pK_a^{exp} - pK_a^{pred}$								
			MD/GB/TI w/ waters	MD/GB/TI w/o waters	PROPKA3.0	GDDM	MM-SCP	EGAD	MCCE	QM/MM	DEPTH
3RN3	GLU2	-1.8	NA	-0.2	0.5	1.3	1.2	NA	-1.3	NA	1.3
4LZT	GLU7	-1.5	2.1	1.5	1.1	0.4	0.6	-0.3	0.6	-0.2	1.2
1PPF	GLU19	-1.2	NA	0.9	2.2	1.0	0.9	0.5	-1.6	-0.5	0.9
2RN2	GLU57	-1.2	NA	1.4	0.3	1.8	-0.5	NA	-0.7	NA	0.5
1A2P	GLU60	-1.2	1.6	1.9	0.2	-0.1	NA	0.0	-1.4	NA	1.1

3RN3	GLU111	-0.9	NA	1.4	1.2	0.5	0.9	NA	0.4	NA	0.5
2RN2	GLU129	-0.8	NA	0.3	-0.3	-0.2	-0.6	NA	-0.8	NA	0.2
2RN2	GLU61	-0.5	NA	1.0	-0.1	-0.4	-0.3	NA	-1.0	NA	-0.1
3RN3	GLU9	-0.4	2.0	1.7	1.0	-0.2	0.6	NA	1.4	NA	0.1
2BCA	GLU26	-0.3	NA	0.7	1.1	NA	NA	NA	-1.4	NA	0.1
1PPF	GLU10	-0.3	NA	1.0	0.3	-0.3	0.2	0.0	-0.6	0.2	0.2
2RN2	GLU119	-0.3	1.6	1.9	-0.6	-0.7	-0.3	NA	-1.0	NA	0.1
1PGA	GLU27	0.1	NA	1.2	-1.7	-0.8	-1.4	-0.2	-0.7	NA	-0.8
1PPF	GLU43	0.4	NA	-0.1	-0.2	-0.5	-0.4	0.6	-0.3	-0.3	-0.6
1DE3	GLU96	0.7	0.0	1.7	2.4	-0.8	NA	-1.0	NA	NA	1.1
1ANS	GLU20	1	NA	-0.3	-0.8	-1.1	NA	NA	NA	NA	-1.0
1RGA	GLU28	1.5	NA	-0.3	-1.6	-1.6	NA	-0.2	NA	NA	-1.4
4LZT	GLU35	1.8	NA	0.0	0.2	-1.0	0.1	0.0	0.0	NA	-2.0
1HNG	GLU41	2.3	NA	0.0	-1.1	NA	NA	-3.3	-0.9	NA	-2.5
1XNB	GLU172	2.3	NA	0.4	0.6	-2.0	NA	0.9	NA	NA	-2.0
RMSD (N)			1.6 (5)	1.1 (20)	1.1 (20)	1.0 (18)	0.7 (13)	1.1 (11)	1.0 (16)	0.3 (4)	1.1 (20)
MAD			1.5	0.9	0.9	0.8	0.6	0.6	0.9	0.3	0.9
MAX			2.1	1.9	2.4	2.0	1.4	3.3	1.6	0.5	2.5

PDB Code	Residue	pK_a^{exp}	Error = $pK_a^{exp} - pK_a^{pred}$								
			MD/GB/TI w/ waters	MD/GB/TI w/o waters	PROPKA3.0	GDDM	MM-SCP	EGAD	MCCE	QM/MM	DEPTH
2SNM	LYS66	-4.1	NA	1.0	0.6	1.5	NA	NA	NA	NA	1.0
1L54	LYS102	-3.9	2.5	3.0	0.2	1.8	NA	NA	NA	NA	1.9
1MUT	LYS39	-2.1	2.5	2.4	2.0	NA	NA	NA	NA	NA	2.6
1NFN	LYS146	-1.3	NA	0.5	1.0	NA	NA	NA	0.2	NA	0.5
1FEZ	LYS53	-1.2	1.4	3.0	-1.0	NA	NA	NA	NA	NA	-0.3
1GS9	LYS146	-1.1	NA	0.3	0.9	NA	NA	NA	NA	NA	0.3
1LE2	LYS143	-1.1	NA	0.9	0.2	NA	NA	NA	NA	NA	-0.2
1NFN	LYS143	-1.0	NA	0.7	0.9	NA	NA	NA	-1.4	NA	0.3
1NZP	LYS312	-1.0	1.0	1.7	1.4	NA	NA	NA	NA	NA	-0.3
1GS9	LYS143	-0.6	NA	0.3	0.5	NA	NA	NA	NA	NA	-0.4
1LE2	LYS146	-0.6	NA	0.1	0.3	NA	NA	NA	NA	NA	-0.2
1PPF	LYS34	-0.4	NA	0.2	0.0	NA	0.9	NA	-2.9	NA	0.8
4LZT	LYS33	-0.1	NA	-0.3	-0.5	0.0	1.0	NA	-0.6	NA	0.3
2BCA	LYS41	0.3	NA	-0.1	-0.5	0.2	-0.3	NA	-0.2	NA	-0.3
4LZT	LYS96	0.3	NA	-0.2	-0.7	-0.2	-0.1	NA	0.5	NA	-0.8

1PGA	LYS28	0.4	NA	-0.2	0.2	0.3	0.5	NA	0.8	NA	-0.3
2BCA	LYS16	0.4	NA	0.3	-0.7	0.6	-0.2	NA	0.3	NA	0.2
1PPF	LYS55	0.6	NA	-0.4	-0.8	NA	-0.6	NA	-0.9	NA	-0.7
2BCA	LYS7	0.7	NA	0.2	-0.7	0.1	-0.3	NA	-0.3	NA	-0.6
2BCA	LYS55	1.3	NA	0.0	-1.3	-0.4	-0.6	NA	-0.1	NA	-1.3
RMSD (N)			2.0 (4)	1.2 (20)	0.9 (20)	0.8 (9)	0.6 (9)	NA	1.1 (11)	NA	0.9 (20)
MAD			1.9	0.8	0.7	0.6	0.5	NA	0.7	NA	0.7
MAX			2.5	3.0	2.0	1.8	1.0	NA	2.9	NA	2.6

PDB Code	Residue	pK_a^{exp}	Error = $pK_a^{exp} - pK_a^{pred}$								
			MD/GB/TI w/ waters	MD/GB/TI w/o waters	PROPKA3.0	GDDM	MM-SCP	EGAD	MCCE	QM/MM	DEPTH
3EBX	HIS6	-3.5	NA	-1.1	2.9	3.2	NA	NA	NA	NA	2.7
3SSI	HIS43	-3.1	NA	0.0	2.9	2.5	NA	NA	NA	NA	0.9
1STN	HIS121	-1	1.9	2.4	1.1	2.2	NA	2.8	NA	NA	1.0
4LZT	HIS15	-0.9	2.2	2.8	1.2	0.7	0.3	1.3	1.1	NA	0.8
1ERT	HIS43	-0.8	NA	1.0	0.3	1.2	NA	NA	NA	NA	1.2
1DE3	HIS137	-0.5	2.6	1.8	-0.9	-0.8	NA	1.2	NA	NA	-0.8

3RN3	HIS48	-0.2	1.1	5.1	-0.4	0.3	0.3	NA	2.7	NA	-1.6
3RN3	HIS119	0.2	NA	-0.7	-0.1	0.9	-0.3	NA	-1.1	NA	-0.5
3RN3	HIS12	-0.3	NA	0.8	-2.3	0.0	-0.2	NA	-1.8	NA	-1.2
1DE3	HIS104	0.2	0.7	2.1	-0.1	-0.6	NA	1.1	NA	NA	0.2
1DE3	HIS36	0.5	NA	0.2	-0.3	-0.2	NA	1.1	NA	NA	-0.7
2RN2	HIS62	0.7	NA	0.0	-0.3	-0.1	0.0	0.0	-0.3	NA	-0.2
2RN2	HIS124	0.8	-1.6	-1.9	-0.8	-0.5	-1.3	NA	-2.6	NA	-0.2
1DE3	HIS50	1.4	2.1	1.7	-3.9	-1.0	NA	0.7	NA	NA	-1.3
1RGA	HIS92	1.5	NA	-0.3	-3.4	-1.1	-0.4	NA	-0.7	NA	-1.3
1RGA	HIS40	1.6	NA	0.0	-2.5	-1.5	-0.5	NA	1.1	NA	-1.4
2RN2	HIS127	1.6	NA	0.0	-0.5	-0.5	-0.3	NA	-0.9	NA	-0.7
1DG9	HIS66	2	NA	0.7	-1.7	-0.9	NA	NA	NA	NA	-1.4
2LZM	HIS31	2.8	-1.3	-1.9	-2.5	-1.9	NA	1.1	NA	NA	-3.0
1DG9	HIS72	2.9	NA	0.3	-3.5	-2.2	NA	NA	NA	NA	-3.3
RMSD (N)			1.9 (8)	1.8 (20)	2 (20)	1.4 (20)	0.5 (9)	1.4 (8)	1.6 (9)	NA	1.5 (20)
MAD			1.8	1.3	1.6	1.1	0.4	1.2	1.4	NA	1.2
MAX			2.4	5.1	3.9	3.2	1.3	2.8	2.7	NA	3.3
Total			1.9 (21)	1.4 (80)	1.3 (80)	1.1 (64)	0.7 (43)	1.2	1.4 (50)	0.3 (6)	1.1 (80)

RMSD (N)								(32)			
TotalMAD			1.8	1.0	0.9	0.8	0.5	0.9	1.0	0.3	0.8
TotalMAX			2.5	5.1	3.9	3.2	1.7	3.3	4.3	0.5	3.3

Note: In some cases, we have used structures of the same protein under different PDB code than those as listed in [250]. They are: 2TRZ = 1 ERT, 2BCA = 1IG5, 1PGA = 1PGB, 1FEZ = 1RQL, 1NZP = 1XSN, 3SSI = 2 SIC, 1STN = 1STY.

RMSD (Root mean squared deviations), MAD (mean absolute deviation), MAX (maximum absolute deviation) for predicted pK_a values are shown for each residue type. The number of pK_a values used to calculate RMSD, MAD, MAX is in parentheses.

5.3.3 Results and Benchmarking

A large number of features were arranged in a linear combination in an attempt to describe microenvironment of a residue as shown in equation 5.2 (Table 5.4). Among these features, depth of polar side chain atoms and depth of main chain atoms were the most informative of the environmental features.

Using the training set of 367 residues with experimentally determined pK_a values, the coefficients of the linear combination of microenvironment features (equation (5.2) were optimized (Table 5.1). These optimized values were applied to make the pK_a predictions on 60 residues on the testing set (Table 5.2). On average, RMSDs of DEPTH-based pK_a predictions were 0.96 pH units in comparison with that of the experimentally determined values. The best performance of pK_a prediction by DEPTH-based method is for ASP with the RMSD of 0.71 pH units, whereas the worst performance is for HIS with the RMSD of 1.26 pH units.

We benchmarked our predictions with those made by other methods, including (i) Molecular dynamics/ generalized-Born/thermodynamic integration (MD/GB/TI), with and without water[251], (ii) PROPKA[186], (iii) Geometry-dependent dielectric method (GDDM)[185], (iv) MM-SCP[182], (v) Egad! A Genetic Algorithm for Protein Design! (EGAD)[252], (vi) Monte Carlo sampling with continuum electrostatics (MCCE)[253] and a Quantum mechanics/molecular mechanics (QM/MM) method[175] (Table 5.3). The predicted pK_a values from the methods listed earlier in the text were obtained from literature[250], except PROPKA. PROPKA 3.0 using default parameters was run over the web server (http://propka.ki.ku.dk/).

In terms of the predicted pK_a error, our predictions were significantly better (using a Wilcoxon paired sign rank test at 95% confidence) than the predictions of EGAD, MD/GB/TI, and GDDM (Table 5.3). Our results were on par with the PROPKA 3.0 and MCCE methods. Only pK_a predicted values from QM/MM (0.30 pH units over five predictions) and MM-SCP (0.70 pH units over 43 predictions) methods have lower errors than our method (0.96 pH units). Though the MM-SCP method is statistically superior to our simple empirical method, we are closer to the experimentally determined value in 18 and worse in 21 of the 43 common predictions. In four cases the results were identical between our method and MM-SCP.

Table 5.4: Physical features tested (individually) for correlation with pK_a.

Number	Feature
1	Main chain atom depth
2	Polar main chain atom depth
3	Side chain atom depth
4	Polar side chain atom depth
5	Residue depth
6	Polar residue depth
7	Number of neighbor atoms
8	Main chain atom depth of neighbor atoms
9	Polar main chain atom depth of neighbor atoms
10	Side chain atom depth of neighbor atoms
11	Polar side chain atom depth of neighbor atoms
12	Residue depth of neighbor atoms
13	Polar residue depth of neighbor atoms
14	Number of neighbor charged atoms
15	Number of hydrogen bonds
16	Electrostatic energy between the ionizable groups and their environments if ionizable groups in charged
17	Electrostatic energy between the ionizable groups and their environments if ionizable groups in neutral
18	Delta of two above electrostatic energies
19	All atom solvent accessible surface area
20	Percentage all atom solvent accessible surface area
21	Non-polar side chain solvent accessible surface area
22	Percentage non-polar side chain solvent accessible surface area
23	Polar side chain solvent accessible surface area
24	Percentage polar side chain solvent accessible surface area
25	Side chain solvent accessible surface area
26	Percentage side chain solvent accessible surface area
27	Main chain solvent accessible surface area
28	Percentage main chain solvent accessible surface area
29	Chi torsion angle of ionizable groups
30	B factor of ionizable groups
31	B factor of ionizable group neighbor

Table 5.5: Optimized coefficients of linear recombination for the different ionizable amino acid.

RES	Model pK_a	c_1	c_2	c_3	c_4	c_5	c_6	c_0
ASP	3.8	0.22	-0.07	-0.21	0.10	-0.02	0.66	-0.88
GLU	4.5	0.00	0.06	-0.01	0.16	0.06	0.14	-0.50
HIS	6.5	-0.20	-0.05	0.06	1.76	-1.14	0.44	1.84
LYS	10.5	-0.02	0.01	-0.01	0.83	0.66	0.69	-0.15

5.4 Meta-algorithm DEMM for pK_a Prediction

We discovered that our predictor synergizes (a modest correlation coefficient of 0.66) with another physics-based method of MM-SCP (Figure 5.1). Hence a meta-predictor DEMM combining these two methods are constructed. MM-SCP attributes the shift of pK_a from the model value solely to electrostatic interaction among ionizable groups in proteins. This method improved the calculation of electrostatic interactions by explicitly considering the screening of the Coulombic potential. Contributions to the Coulombic screening come from the amino acid residues in the surrounding microenvironment of the ionizable group. To model this screening effect the MM-SCP method uses the hydrophobicity and accessibilities of chemical groups constituting the amino acids.

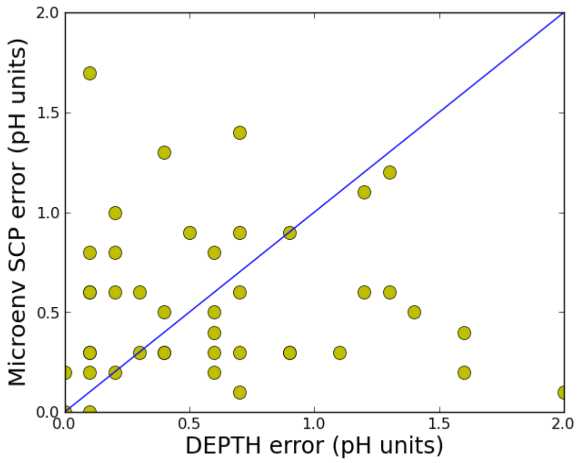


Figure 5.1: Complementarity between DEPTH and MMSCP in pK_a prediction.

5.4.1 Improvement in the Electrostatic Calculation.

In DEMM, we made a change in the calculation of the electrostatic term, which improved the pK$_a$ calculation according to the formula

$$EE_R = \sum_{i \in R} \sum_{j \in R_b} \frac{\Delta Q_i . Q_j}{r_{ij}} \quad (5.4)$$

where ΔQ_i is the difference in partial charge of an atom i in a residue R between its protonated/deprotonated forms.

Q_i, Q_j and r_{ij} are as in section 5.3.1

If R_b is an ionizable residue, the partial charge Q_j was chosen to correspond to the protonation state of residue R_b at a pH equivalent to the model pK$_a$ of residue R.

The predicted pK$_a$, $pK_a^{pred}$is computed as

$$pK_a^{pred} = pK_a^{model} + c_1.depth^{MC} + c_2.depth^{polarSC} + c_3.HB$$

$$+ c_4.EE_R + c_5.ASA^{SC} + c_6.(pK_a^{MM-SCP} - pK_a^{model}) + c_0 \quad (5.5)$$

where pK_a^{model} and pK_a^{MM-SCP} are the model pK$_a$ and pK$_a$ predicted by MM-SCP method respectively.

$c_0 - c_6$ are coefficients of the individual features.

The values of the coefficients were optimized over a training set of residues.

5.4.2 Dataset of experimental values of pK$_a$ used in DEMM Prediction

A dataset of 222 amino acid residues with their pK$_a$ values experimentally measured[254] were used to train (175) and test (47) our algorithm. The entire dataset consists of 58 ASP, 57 GLU, 71 HIS and 36 LYS residues from 54 X-ray structures (resolution ranging from 1.2Å to 3.2Å).

5.4.3 Results and Performance Benchmark

The coefficients of the linear combination, $c_0 - c_6$ (equation 5.5) for each of the residue types ASP, GLU, HIS and LYS were optimized separately. On the training dataset of 175 residues a conjugate gradient optimization was done (Table 5.5). First, the coefficient of the MM-SCP method contribution, c_6, was set to 0.5. The coefficients of the features contributing to original DEPTH algorithm, $c_0 - c_5$ were taken as one half of their values from the original algorithm[88] in section 5.3. The optimization was performed until convergence was reached with a tolerance value of $1e^{-8}$.

Our prediction method was tested on a set of 47 residues. The error rates for different amino acids were slightly different from each other. Our predictions for LYS were closest to the experimentally determined values (Mean error = 0.33 pH units, RMSD = 0.40 pH units), whereas predictions for HIS were the farthest (Mean error = 0.65 pH units, RMSD = 0.88 pH units). Overall, the prediction error is about 0.49 pH units (or RMSD 0.67 pH units) (Table 5.6). Using a Wilcoxon paired sign rank test, our method was shown to be statistically significantly superior to its individual component methods, particularly DEPTH and MM-SCP (Table 5.6).

Detailed information on the pK_a predictions for the 222 ionizable residues of the training and testing set are shown in Table 5.7.

5.5 Case Study and Web-server

5.5.1 Case Study

The residue ASP 148 E-coli ribonuclease H (PDB ID: 2RN2) (Figure 5.2) is an example of complementarity between pK_a prediction by DEPTH and MM-SCP. The experimental determined pK_a value of this residue is 2. This residue has a polar side chain depth of 5.6 Å and 6 hydrogen bonding interactions. As depth and hydrogen bonding effect are not considered in MM-SCP, this method overpredicted the pK_a value by 1.25 pH units. In the case DEPTH had a smaller error but underpredicted the value by -0.69 pH units. DEMM predicts a pK_a of 1.84, only -0.16 pH units from the experimental value. There are improvement of 0.53 and 1.09 pH units in DEMM prediction over the DEPTH and MM-SCP methods, respectively.

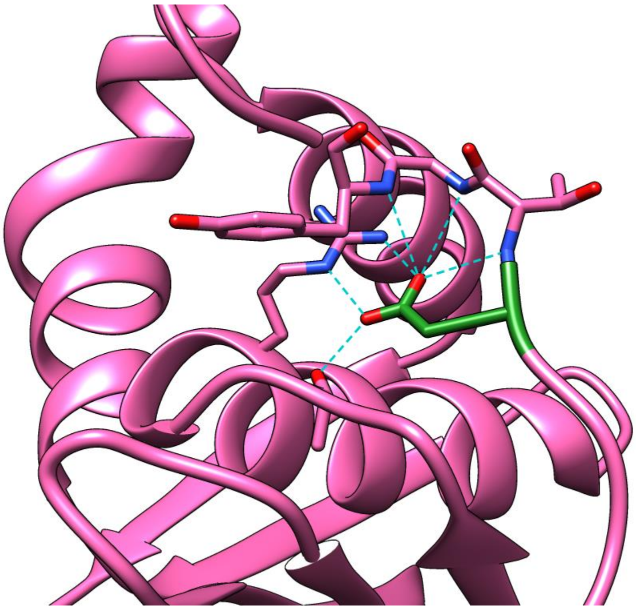


Figure 5.2: A ribbon representation of the ribonuclease H (PDB ID: 2RN2). The pK_a predicted 148D residue is shown in green sticks. The four residues (46, 149-151) that make hydrogen bonds (cyan lines) with 148D are also shown in pink stick representation. The figure was generated using Chimera[42].

Table 5.6: Performance benchmark of DEPTH, MMSCP and DEMM over 47 ionizable groups on the testing set. The mean absolute errors of predictions by the different methods and the corresponding RMSD (in brackets) are recorded in pH units. The improvement is the difference between DEMM and the more accurate prediction between DEPTH and MM-SCP.

Residue type (N)	DEPTH	MM-SCP	DEMM	Improvement
ASP (12)	0.72 (1.04)	0.67 (0.79)	0.53 (0.67)	0.14 (0.12)
GLU (12)	0.37 (0.48)	0.45 (0.58)	0.37 (0.5)	0 (-0.02)
HIS (15)	1.14 (1.45)	1.15 (1.54)	0.65 (0.88)	0.49 (0.57)
LYS (7)	0.43 (0.53)	0.48 (0.66)	0.33 (0.4)	0.1 (0.13)
Total (47)	0.71 (1.03)	0.73 (1.04)	0.49 (0.67)	0.22 (0.36)
P-value (1-tailed)	0.000*	0.004*		
P-value (2-tailed)	0.001*	0.012*		

* statistical significance

5.5.2 Web-server

The pK_a prediction of ionizable amino acid residues is available with other prediction tools which use depth as a determinant feature. The server http://mspc.bii.a-star.edu.sg/depth is freely accessible with no login requirements. Either four-letter PDB code or protein structure in PDB format is acceptable in our server. The information about the program, as well as its parameters, is available at help pages.

The results of the pK_a prediction are pictorially viewed represented with figure legends. The results in tab-delimited format are also available for downloading. All results are stored up to 30 days. Users can download the standalone version of the pK_a prediction program for local use.

5.6 Chapter Summary and Discussions

DEPTH-based pK_a prediction method empirically describes the microenvironment using both physical and chemical features. MM-SCP pK_a prediction method calculates the screened electrostatic interactions of ionizable groups by considering the hydrophobicity of the surrounding amino acid residues. These two methods complement each other. On the one hand DEPTH has considered both the residue environments using residue depth and a coarse treatment of the electrostatics within a radius of up to 12 Å. On the other hand, MM-SCP has sophisticated electrostatics, describing the screening effect in the 4.25 Å vicinity of an ionizable group. However, MMSCP uses only solvent accessible area to describe residue environment.

The DEMM pK_a prediction method integrated the two complementary methods, DEPTH and MM-SCP. Overall, DEMM has a mean error of 0.49 pH units and an RMSD of 0.67 pH units. This model improved the prediction of all four ionizable amino acid residue types, including ASP, GLU, HIS and LYS. One significant improvement is in the case of HIS, which is usually the most difficult to predict. With a deviation of 0.49 pH units from experimental values, it was better predicted than the previous methods.

In comparison with other empirical models, quantum mechanics/molecular mechanics (QM/MM)[175] method still remains as the most accurate pK_a prediction method with an RMSD of 0.3 pH units. The high performance of QM/MM could be because of its flexibility and adaptive assignment of partial charge, as well as its explicit consideration of protein dynamics. However, the QM/MM approach is computationally expensive and not possible for large-scale studies. In contrast, our empirical method is fast, relatively accurate and can readily be applied to proteins/protein complexes without a size limitation. One disadvantage of DEMM pK_a prediction is the absence of pK_a prediction of residues Cys, Tyr and Arg. This problem could be overcome when the number of the experimentally determined pK_a values of those residues is sufficient to train our algorithm.

Table 5.7: pK$_a$ dataset for DEMM method.

PDB code	Res name	Res number	pK_a^{exp}	Error = $pK_a^{exp} - pK_a^{pred}$		
				MM-SCP	DEPTH	DEMM
1A2P	ASP	12	3.65	0.56	-0.21	0.04
1A2P	ASP	22	3.3	0.45	0.22	-0.04
1A2P	ASP	44	3.35	0.66	0.46	0.18
1A2P	ASP	75	3.1	-0.36	1.85	-0.51
1A2P	ASP	8	3	0.81	-0.19	0.35
1A2P	ASP	86	4.2	-0.76	-0.57	-0.80
1A2P	GLU	29	3.75	0.32	0.29	0.07
1A2P	GLU	60	3.4	0.41	0.72	0.29
1A2P	HIS	102	6.3	-0.04	0.15	0.32
1A2P	HIS	18	7.9	-1.16	-1.42	-1.28
1AZP	ASP	16	2.89	1.08	0.56	0.52
1AZP	ASP	35	3.42	-0.20	-1.6	-0.55
1AZP	ASP	36	3.12	0.87	0.43	0.26
1AZP	ASP	49	3.55	-0.21	-0.73	-0.66
1AZP	ASP	56	3.35	-0.02	-0.51	-0.47
1AZP	GLU	11	4.19	-0.41	-0.39	-0.59
1AZP	GLU	12	4.41	-0.43	-0.33	-0.70
1AZP	GLU	14	4	-0.05	-0.01	-0.18
1AZP	GLU	47	4.21	0.02	-0.3	-0.26
1AZP	GLU	53	3.53	0.86	0.4	0.59
1AZP	GLU	62	3.99	0.73	0.21	0.30
1AZP	GLU	64	4.23	0.44	-0.08	0.01
1BCX	GLU	78	4	-0.54	1.19	-0.61
1BEO	ASP	21	2.49	0.26	-0.33	-0.03
1BEO	ASP	30	2.51	0.85	0.47	0.51
1BEO	ASP	72	2.61	0.86	0.81	0.62
1BNZ	ASP	16	2.11	1.28	2.1	1.21
1BNZ	ASP	35	2.16	0.59	1.13	0.52
1BNZ	ASP	50	2.96	0.40	0.7	0.09
1BNZ	GLU	11	3.78	-0.52	-0.01	-0.50
1BNZ	GLU	12	3.9	-0.21	0.42	-0.23
1BNZ	GLU	36	4.33	-0.52	-0.51	-0.67
1BNZ	GLU	48	3.45	0.73	0.53	0.40
1BNZ	GLU	54	3.01	0.23	0.83	0.13
1BNZ	GLU	60	3.82	0.31	0.28	0.11
1BNZ	ASP	35	2.67	0.95	0.62	0.52
1BNZ	ASP	50	3.55	0.27	0.11	-0.23
1BNZ	GLU	11	4.17	-0.25	-0.4	-0.51
1BNZ	GLU	12	4.33	-0.51	-0.01	-0.59
1BNZ	GLU	36	4.89	-0.32	-1.07	-0.78
1BNZ	GLU	48	4.03	0.37	-0.05	-0.05
1BNZ	GLU	54	3.56	0.37	0.28	-0.02
1BNZ	GLU	60	4.24	0.31	-0.14	-0.07
1BTJ	HIS	249	7.4	-1.78	-1.21	-0.73

1DG9	HIS	66	8.29	-2.03	-1.87	-1.26
1DIV	ASP	23	3.05	0.21	0.65	-0.01
1DIV	ASP	8	2.99	0.99	0.56	0.43
1DIV	GLU	17	3.57	0.37	0.16	0.17
1DIV	GLU	38	4.04	0.30	-0.05	0.03
1DIV	GLU	48	4.21	0.41	-0.15	0.03
1DIV	GLU	54	4.21	-0.30	-0.16	-0.57
1DUI	HIS	64	7.17	-0.29	-0.25	0.15
1ERT	ASP	16	3.7	0.40	0.34	-0.01
1ERT	ASP	20	3.6	0.18	0.13	-0.25
1ERT	GLU	103	4.9	-0.43	-0.47	-0.68
1ERT	GLU	13	4.8	-0.31	-0.61	-0.58
1ERT	GLU	68	4.2	0.60	0.15	0.20
1ERT	GLU	88	3.9	0.08	0.18	-0.17
1ERT	HIS	43	5.5	0.37	1.32	1.21
1FEZ	LYS	53	9.3	0.35	-0.96	3.43
1FNA	ASP	67	4.18	-0.33	-1.21	-0.87
1FNA	ASP	80	3.4	0.87	0.32	0.15
1FNA	GLU	38	3.79	-0.02	0.13	-0.18
1FNA	GLU	47	3.94	-0.22	-0.07	-0.46
1GS9	LYS	143	9.9	0.45	-0.28	-3.20
1GS9	LYS	146	9.4	0.92	0.32	-2.65
1GS9	LYS	157	10.9	-0.17	-0.42	-3.70
1GS9	LYS	69	10.1	0.57	0.4	-2.88
1GS9	LYS	72	10	0.40	0.5	-2.91
1GS9	LYS	75	10.1	1.59	-0.14	-2.20
1GS9	LYS	95	10.1	0.34	0.14	-3.12
1GYM	HIS	227	6.9	-0.64	-0.75	-0.68
1GYM	HIS	32	7.6	-2.94	-1.64	-1.44
1GYM	HIS	82	6.9	0.37	-0.94	-0.05
1GYM	HIS	92	5.4	0.55	0.13	0.37
1HHO	HIS	20	6.7	0.04	-0.06	-0.15
1HHO	HIS	45	7	-1.27	-1.5	-1.25
1HHO	HIS	50	7.5	-0.99	-0.69	-0.54
1HHO	HIS	72	6	1.23	0.83	1.03
1HHO	HIS	89	7.2	-2.02	-0.88	-0.90
1HHO	HIS	2	6.51	-1.33	-0.32	-0.40
1HHO	HIS	77	6.6	0.05	-0.06	0.30
1HNG	GLU	41	6.7	-1.34	-2.54	-1.94
1HRC	HIS	33	6.4	-0.86	0.17	-0.08
1HRC	LYS	79	9	1.70	1.47	-1.89
1HV0	GLU	78	5	-0.85	0.64	-1.15
1HV1	GLU	78	4.2	3.49	1.73	1.86
1L54	LYS	102	6.5	-0.58	1.71	-2.29
1L98	GLU	105	6	-1.62	-1.24	-2.08
1LE2	LYS	143	9.4	0.87	-0.16	-2.83
1LE2	LYS	146	9.9	0.33	-0.2	-3.32
1LE2	LYS	157	10.9	-0.33	-0.23	-3.78

1LE2	LYS	69	10.1	0.58	0.6	-2.90
1LE2	LYS	72	10	0.55	0.34	-2.92
1LE2	LYS	75	10	0.42	0.31	-2.69
1LE2	LYS	95	10.2	0.25	0.21	-3.20
1LZ1	HIS	78	7.12	-0.68	-0.65	-0.87
1NFN	LYS	143	9.5	0.85	0.4	-2.71
1NFN	LYS	146	9.2	1.11	0.45	-2.52
1NFN	LYS	157	11.1	-0.54	-0.53	-4.00
1NFN	LYS	69	10.4	0.41	0.26	-3.10
1NFN	LYS	72	10	0.69	0.64	-2.76
1NFN	LYS	75	10.1	0.88	0.1	-2.57
1NFN	LYS	95	10.1	0.29	0.24	-3.21
1PGA	LYS	10	11	-0.35	-0.11	-3.78
1PGA	GLU	15	4.4	-0.67	-0.41	-0.70
1PGA	GLU	19	3.7	0.28	0.27	0.01
1PGA	ASP	22	2.9	0.31	-0.52	-0.04
1PGA	GLU	27	4.5	-0.78	-1.09	-1.04
1PGA	LYS	28	10.9	0.31	-0.5	-3.47
1PGA	ASP	36	3.8	0.61	0.23	0.04
1PGA	ASP	40	4	0.03	-0.1	-0.42
1PGA	GLU	42	4.4	0.04	-0.02	-0.17
1PGA	ASP	46	3.6	0.31	-0.55	-0.10
1PGA	ASP	47	3.4	0.01	-0.61	-0.34
1PGA	GLU	56	4	0.51	0.45	-0.01
1PNT	HIS	66	8.29	-2.01	-1.74	-1.21
1PNT	HIS	72	9.19	-2.09	-2.47	-2.98
1POH	HIS	76	6	-0.86	0.18	0.32
1PPF	ASP	7	2.99	0.58	0.09	0.14
1PPF	GLU	10	4.1	0.00	0.09	-0.17
1PPF	LYS	13	9.9	0.50	0.93	-2.86
1PPF	GLU	19	3.2	0.50	0.68	0.21
1PPF	ASP	27	2.71	1.21	-0.22	0.31
1PPF	LYS	29	11.1	-0.64	-0.89	-3.94
1PPF	LYS	34	10.1	0.88	0.3	-2.65
1PPF	GLU	43	4.7	-0.29	-0.52	-0.57
1PPF	LYS	55	11.1	-0.59	-0.83	-4.09
1RCA	ASP	121	3.1	1.11	-0.31	0.20
1RCA	ASP	14	2	1.20	0	0.66
1RCA	ASP	38	3.1	0.75	0.12	0.12
1RCA	ASP	53	3.9	0.06	0.28	-0.24
1RCA	ASP	83	3.5	-0.24	-0.37	-0.53
1RCA	GLU	111	3.5	1.23	0.6	0.83
1RCA	GLU	2	2.8	0.92	0.47	0.37
1RCA	GLU	49	4.7	-0.35	-0.56	-0.67
1RCA	GLU	86	4.1	0.12	-0.06	-0.25
1RCA	GLU	9	4	0.55	0.08	-0.10
1RCA	HIS	105	6.7	0.30	-0.24	0.00
1RCA	HIS	119	6.1	1.17	0.23	0.06

1RCA	HIS	12	6.2	-1.03	-1.34	-0.68
1RCA	HIS	48	6	0.78	-1.07	-0.09
1RGG	ASP	1	3.44	0.67	0.72	0.28
1RGG	ASP	17	3.72	0.40	0.68	0.04
1RGG	ASP	25	4.87	-0.72	-0.46	-1.05
1RGG	ASP	33	2.39	0.47	1.58	-0.48
1RGG	ASP	79	7.37	-1.84	-2.17	-2.61
1RGG	ASP	84	3.01	0.34	-0.95	-0.29
1RGG	ASP	93	3.09	0.88	0.1	0.59
1RGG	GLU	14	5.02	0.13	-0.51	-0.36
1RGG	GLU	41	4.14	0.03	0.15	-0.14
1RGG	GLU	74	3.47	1.02	1.06	0.79
1RGG	GLU	78	3.13	1.69	1.31	0.99
1RGG	HIS	53	8.27	-1.00	-1.27	-0.60
1RGG	HIS	85	6.35	-0.18	0.07	0.21
1XNB	ASP	106	2.7	1.39	1.19	1.04
1XNB	ASP	11	2.5	1.09	0.35	0.54
1XNB	ASP	119	3.2	0.01	-0.32	-0.22
1XNB	ASP	121	3.6	0.27	0.4	-0.10
1XNB	ASP	4	3	0.35	-0.63	-0.17
1XNB	GLU	78	4.6	1.32	0.95	0.31
1XNB	HIS	156	6.5	-0.31	-0.62	-0.89
1XWW	HIS	157	7.72	-1.29	-0.36	-0.81
1XWW	HIS	66	8.22	-2.16	-2.09	-1.61
1XWW	HIS	69	6.4	-0.18	-0.47	-0.47
1XWW	HIS	72	9.18	-3.60	-2.45	-3.21
1YMB	HIS	113	5.4	-0.10	1.28	1.38
1YMB	HIS	116	6.6	-0.71	-0.64	-0.30
1YMB	HIS	119	6.4	-3.97	0.08	-0.14
1YMB	HIS	36	7.8	-1.66	-1.59	-1.00
1YMB	HIS	81	6.6	-0.13	-0.21	-0.14
2CPL	HIS	126	6.34	-0.74	-0.14	0.06
2CPL	HIS	70	5.84	0.52	-0.15	0.32
2LZT	LYS	1	10.6	-0.20	-0.26	-3.71
2LZT	GLU	7	2.73	1.02	1.08	0.90
2LZT	LYS	13	10.3	0.33	0.27	-3.03
2LZT	HIS	15	5.58	-0.42	0.29	0.16
2LZT	ASP	18	2.78	0.89	0.57	0.45
2LZT	LYS	33	10.4	0.40	-0.61	-3.07
2LZT	GLU	35	6.15	0.22	-1.66	-0.97
2LZT	ASP	48	3.4	-0.17	-1.17	-0.67
2LZT	ASP	52	3.67	0.19	0.18	0.02
2LZT	ASP	66	2	0.75	-0.25	-0.06
2LZT	ASP	87	2.84	0.46	0.26	0.14
2LZT	LYS	96	10.7	-0.74	-0.6	-3.94
2LZT	LYS	97	10.1	0.47	0.45	-2.83
2LZT	ASP	101	4.17	0.02	-0.46	-0.65
2LZT	LYS	116	10.2	0.03	-0.19	-3.48

2LZT	ASP	119	2.85	0.38	-0.26	-0.02
2MB5	HIS	113	5.44	0.31	0.94	1.14
2MB5	HIS	116	6.49	-0.21	-0.31	0.00
2MB5	HIS	119	6.13	-1.69	-0.43	0.10
2MB5	HIS	12	6.29	-0.17	0.17	0.33
2MB5	HIS	48	5.25	0.95	0.93	1.17
2MB5	HIS	81	6.68	-0.14	0.07	0.02
2MB5	HIS	97	5.63	0.38	-0.32	-0.22
2RN2	GLU	6	4.1	0.47	-0.4	-0.04
2RN2	ASP	10	5.52	-1.24	-0.56	-1.39
2RN2	GLU	32	3.5	-0.26	0.19	-0.36
2RN2	GLU	48	4.2	0.03	-0.31	-0.22
2RN2	GLU	57	3.67	-0.37	0.05	-0.47
2RN2	GLU	61	4.03	-0.27	-0.57	-0.67
2RN2	HIS	62	7	-0.25	0.2	0.06
2RN2	GLU	64	4.47	-0.19	-0.71	-0.57
2RN2	ASP	70	3.37	0.85	0.45	0.24
2RN2	HIS	83	5.5	0.37	0.43	0.43
2RN2	ASP	94	3.27	-0.02	-0.96	-0.56
2RN2	ASP	102	2	2.14	0.81	1.18
2RN2	ASP	108	3.55	0.46	-0.1	-0.07
2RN2	HIS	114	5	0.25	0.24	1.93
2RN2	GLU	119	4.47	-0.72	-0.4	-0.94
2RN2	HIS	124	7.1	-1.25	-0.49	-0.35
2RN2	HIS	127	7.9	-0.82	-0.32	-0.17
2RN2	GLU	129	3.7	0.53	0.17	0.07
2RN2	GLU	131	4.47	0.00	-0.08	-0.22
2RN2	ASP	134	4.12	-0.67	0.39	-0.68
2RN2	GLU	135	4.5	-0.13	-0.29	-0.34
2RN2	GLU	147	4.23	0.08	0.1	-0.11
2RN2	ASP	148	2	1.25	-0.65	0.26
2RN2	GLU	154	4.35	-0.35	-0.42	-0.70
2TGA	ASP	102	1.4	1.34	3.29	1.23
2TGA	ASP	194	2.3	1.72	2.73	0.94
2TGA	HIS	40	4.6	1.23	1.63	2.45
2TGA	HIS	57	7.3	-1.19	-0.97	-0.20
2TRX	ASP	26	7.5	0.50	0	-1.84
3EBX	HIS	26	5.8	-0.32	-0.09	0.42
3ICB	LYS	1	10.6	0.54	0.09	-3.31
3ICB	LYS	7	11.35	-0.14	-0.84	-3.58
3ICB	LYS	12	11	0.75	-0.15	-3.13
3ICB	LYS	16	10.09	0.92	0.98	-2.64
3ICB	LYS	25	11.69	0.06	-0.65	-3.62
3ICB	LYS	29	10.99	0.29	-0.94	-3.21
3ICB	LYS	41	10.89	-0.25	-0.55	-3.84
3ICB	LYS	55	11.39	0.36	-0.77	-3.71
3ICB	LYS	71	10.72	-0.07	-0.26	-3.54
3ICB	LYS	72	10.97	-0.22	-0.4	-3.70

3RN3	GLU	2	2.81	1.03	0.73	0.58
3RN3	GLU	9	4	0.56	0.12	0.23
3RN3	HIS	12	6.2	-0.37	-0.91	-0.67
3RN3	ASP	14	2	0.62	-0.53	0.33
3RN3	ASP	38	3.1	0.71	0.2	0.11
3RN3	HIS	48	6	0.44	-0.84	-1.61
3RN3	GLU	49	4.7	-0.36	-0.47	-0.66
3RN3	ASP	53	3.9	0.05	0.31	-0.25
3RN3	ASP	83	3.5	-0.39	-0.52	-0.62
3RN3	GLU	86	4.1	-0.34	-0.21	-0.54
3RN3	HIS	105	6.7	0.40	-0.32	-0.07
3RN3	GLU	111	3.5	0.87	0.46	0.58
3RN3	HIS	119	6.09	0.16	0.26	0.25
3RN3	ASP	121	3.1	0.82	-0.1	-0.01
3RNT	HIS	27	7.3	-1.33	-0.2	0.07
3RNT	GLU	28	5.9	-0.96	-1.41	-1.38
3RNT	HIS	40	7.9	-0.86	-1.18	-0.72
3RNT	GLU	58	4.3	0.77	-0.16	0.15
3RNT	HIS	92	7.8	-0.93	-1.43	-0.78
3SRN	HIS	105	6.66	-0.13	-0.3	-0.22
3SRN	HIS	119	6.31	-0.08	0.03	-0.26
3SRN	HIS	12	5.85	-0.62	-0.24	-0.30
3SSI	HIS	106	6	-0.73	-0.52	-0.26
4HHB	HIS	112	7.6	-0.36	-0.8	-0.20
4HHB	HIS	72	6.6	0.24	0.37	0.73
4HHB	HIS	89	7.18	-0.68	-0.8	-0.46
4HHB	HIS	143	6.25	-0.67	-0.13	0.24
4HHB	HIS	77	6.75	-0.40	0.12	-0.01
4PTI	ASP	3	3.57	0.26	-0.04	-0.26
4PTI	GLU	7	3.89	0.17	0.44	-0.11
4PTI	LYS	15	10.43	0.02	-0.2	-3.63
4PTI	LYS	26	10.1	0.32	0.23	-3.27
4PTI	LYS	41	10.6	0.03	-0.11	-3.50
4PTI	LYS	46	9.87	0.50	0.43	-3.00
4PTI	GLU	49	4	0.16	0.17	-0.03
4PTI	ASP	50	3.18	-0.42	-0.12	-0.49
5PNT	HIS	157	7.49	-1.02	-0.83	-0.81
5PNT	HIS	66	7.67	-2.26	-1.04	-0.86
5PNT	HIS	72	9.23	-2.01	-2.17	-2.60
6GST	HIS	167	7.77	-0.74	-0.5	-0.28
6GST	HIS	83	5.18	0.61	1.06	1.72
6GST	HIS	84	7.08	-0.74	-0.74	-0.68

Chapter 6
Discussions and Future Directions

The thesis focuses on protein-peptide interactions and more specifically, protein-PPII interactions. In this thesis, first I have helped to solve two crystal structures as well as applied MD simulation to study the mechanism why CLIP is abundant in DQ2.5. As the abundance of CLIP is connect to the editing process catalyzed by DM in MHCII proteins, in the next chapter, a pure computational MD study was applied on six systems to reveal the dynamic of peptide editing process. These two chapters focus on one example of protein-PPII interactions, in the following chapter, the characteristics of PPII-binding protein was generalized and applied to predict the PPII receptors. Another aspect that was tacked to reveal the protein-peptide interaction is the pK_a prediction of ionizable residues. The pK_a prediction could help to estimate the protonation state of those ionizable residues and hence understand protein functions. The conclusions and future directions of each topic are discussed in separate sections.

6.1 3D structures of DQ2.5-CLIP complexes

6.1.1 Summary

In this section, the structures of DQ2.5 in complex with two CLIP peptides, namely CLIP1 and CLIP2 have been determined by X-ray crystallography. The analysis of crystal structures shows that DQ2.5 has an unusually large P4 pocket and a positively charged peptide binding groove. These two features together promote preferential binding of CLIP2 over CLIP1. In addition, there is a α9-α22-α24-α31-β86-β90 hydrogen bond network which locates at the bottom of the peptide binding groove of DQ2.5. The hydrogen bond network spanning from the P1 to P4 pockets results in the relative immobility of hydrogen bond making residues. This network, as well as the deletion mutation at α53, may lead to the DM insensitivity of DQ2.5. Later, this hypothesis is proven by the MD simulations. The recent biochemical studies by other groups also support our hypothesis. In conclusion, diminished DM sensitivity is a reason for the CLIP-rich phenotype of DQ2.5.

6.1.2 Suggestions and Future Directions

In this study, we use an implication that the DQ2.5–DM interaction is similar to DR–DM interaction, which is shown by the available crystal structure. However, as DQ2.5 structure has a deletion at α53. It could be possible that DQ2.5–peptide–DM structure at the interaction site is different from DR–peptide-DM. DQ2.5 protein has been shown to be correlated with celiac disease and type 1 diabetes, and hence, understanding DQ2.5–DM interaction could give some clues for understanding the disease mechanism as well as proposing specific treatments. We suggest that the DQ2.5–DM structure should be solved in order to have an elaborate and accurate view of the DQ2.5–DM interaction.

6.2 The Molecular Mechanism behind the Peptide-editing Process of MHCII by DM Catalyst: an MD study

6.2.1 Summary

In this study, we investigated the mechanism on six systems: model_5.5, model_6.5, DR–HA–DM, DR–DM, DR–HA and DR. The simulations revealed that model_5.5 and model_6.5 have fluctuation differences at the β2 domain of DR. Both systems also have smaller conformational change with respect to the starting structure. The DR–HA–DM simulations showed the stabilization of DR β2 domain by interacting with the DM β2 domain. All three DR–DM complex in the presence of peptide presented the conformational change at the α69-75 region in DM. Although this DM α region is far from the DR–DM interaction site, it could be possible that the conformational change in that region could result in the long-range effect onto the DR–DM interaction. In this study, we also showed the stable close state of the peptide-free DR. The closing conformation is not observed in the presence of DM. This observation is consistent with the hypothesis that DM stabilizes the peptide-free DR.

6.2.2 Suggestions and future directions

In this study the change from the apo to the holo conformation has not been observed yet. The differences between the apo and the holo forms are at the peptide-binding site and β2 domains of DR/DM proteins. We suggest that the replica-exchange MD simulations, but not the standard MD simulations used in this study, in the interaction site could be used to enhance the sampling at that region. The MHCII proteins are polymorphic, and the MHCII–DM interaction could be fully understood if the simulations of the complexes between DM and all types of human MHCII proteins, namely DR, DQ and DP are extensively studied. Another suggestion is that the experimental studies on β2 domain of DR and α69-75 region of DM should get more attention.

6.3 Prediction of PPII receptors

6.3.1 Summary

Protein-PPII interaction occurs in many signalling network, immune response *etc.* Finding the PPII receptors could elaborate the possible network of PPII peptide or proteins containing PPII peptides. In this study, we have shown the important features of the PPII-binding site. We also used those features to predict the PPII-binding site of a query protein. After applying the prediction protocol, on a non-redundant dataset of 17, 000 proteins, 125 possible PPII-binding proteins have been detected. This PPII prediction program is comparable to the state-of the-art methods in predicting protein-peptide interactions.

6.3.2 Suggestions and future directions

In this study, we suggest 125 proteins that could be plausible PPII receptors. These data remain to be tested by experiments. Our assumption in this study is that the apo and the holo conformations of the PPII receptors are not significantly different. In other words, we simplified the protocol by using only the holo forms as templates. Both the apo and the holo conformations should be used for searching the PPII-binding sites. The apo forms could be generated by MD simulations.

Our PPII prediction used the Trp residues as a requirement for PPII-binding sites. There are other PPII receptors, such as collagen-bind proteins or PDZ domain that do not have Trp. It could be possible to generalize the protocol by using any two donor residues as templates.

In addition, similar protocol could be applied to other protein-bound peptide conformations such as α-helix. Particularly, the features for the α-helix binding proteins should be extracted from the known α-helix binding proteins. The CLICK structural alignment program could be used to compare a query structure with the template. Classification approaches, not limited to SVM could be applied to distinguish the binding/non-binding positions. The α peptide can also be built by Monte Carlo simulations to reduce the clashes between the template peptide and the query protein.

The PPII prediction protocol focused only on the conformation of the protein and the PPII peptide. The sequences of either PPII or receptors were ignored. More attention should be paid to the sequence characteristics of the peptides, as it could be helpful for the design of a potential drug candidate. This could be a future research direction to be explored.

6.4 pK_a prediction of ionizable residues

6.4.1 Summary

From chapter two to chapter four, we have studied protein-peptide interactions. In chapter 3, the protein-peptide interactions are shown to be highly dependent on the protonation states of ionizable residues. In a broader context, the protein function is monitored by the pK_a of all ionizable residues. And hence, in chapter five, we utilized two linear regression models in order to predict the pK_a of ionizable residues. The first model used residue depth in describing the microenvironment. This model gives the RMSD of pK_a prediction values of 1 pH unit in comparison with experimental values. The first model was also shown to complement with another state-of-the-art pK_a prediction program, MMSCP. As a consequence, a meta-algorithm was applied to build the second model that improved the prediction to 0.7 pH unit RMSD. This study has been of benefit in studying the protein interactions with changing protonation states of ionizable residues, such as in the case of triad catalytic Ser-His-Asp motif or MHCII–DM interactions. The web-server of the first model is also freely available (http://mspc.bii.a-star.edu.sg/depth).

6.4.2 Suggestion and future direction

One major possible improvement in pK_a prediction approach is the shift of pK_a value in different pH conditions. Our pK_a prediction implies that the protonation states of neighbourhood residues are typically oversimplified by using the protonation states of isolated residues at pH 7. The calculation should consider the pK_a shift of ionizable residues that are surrounding the interested residue. In addition, we only predicted the pK_a for ASP, GLU, HIS and LYS. It should be expanded to CYS, TYR and ARG. Another possible improvement in the pK_a prediction approach is that in the case of HIS we only predicted whether the HIS is either in the protonated or deprotonated. In the deprotonated state, there are two possible conformations, either deprotonated at $N^{\delta1}$ or at $N^{\varepsilon2}$. However, due to the paucity of experimental data, these predictions are currently not made as they are not testable.

6.5 Conclusion remarks

In this thesis, I have extensively investigated protein-peptide interactions, where the peptides have the PPII conformation. First, I helped to solve the crystal structures of DQ2.5 with two different CLIP peptides. Two specificities were suggested as the reasons why CLIP is retained in DQ2.5, but not DR1, namely, the deletion at α53 position in DQ2.5 and the hydrogen bond network from P1 to P4 pocket. The reason why DQ2.5 prefers to bind CLIP2 was proposed as DQ2.5 has a bigger P4 pocket than other MHCII homolog. And hence, CLIP2 peptide that has Met at P4 position prefers to bind DQ2.5. DQ2.5 is highly correlated with type 1 diabetes and celiac disease and hence, this study could contribute to the knowledge of these diseases and their treatments.

As the peptide editing in MHCII is catalyzed by DM. We next performed MD simulations on six systems to address the following questions, (i) why DR–DM interaction occurs only at pH 5.5 but not at pH 6.5, (ii) what are the important residues for the DR–DM interaction (iii) how these residues change from the apo to the holo conformations, (iv) how peptide releases from DR and (v) how DM stabilizes free-peptide DR. We revealed that the conformational change during DR–DM interaction happens not only at the α1 and β1 domain, but also at the β2 domains. We also showed that without DM, peptide-free DR closed the peptide binding groove. This study deciphers the DR–DM interaction that is important for peptide exchange and hence immune response process. In other words, this study gives a broad overview of how our immune system response when the pathogens entry our body.

These two studies above focus on the example of MHCII-peptide complexes. We expanded the knowledge on protein-PPII interactions by analyzing important features for the PPII-binding proteins. Several features, particularly, the number of hydrogen bonds, residue depth and entropy conservation were also shown to be important for the PPII-binding site. We then applied those requirements to predict the PPII receptors. The prediction of PPII receptors could be applied on studying the network of the PPII-peptide or protein containing the PPII conformation stretch.

Last, chapter 3 shows in that the protein-peptide interactions highly depend on the pH condition. This pH condition monitors the protonation state of all ionizable residues, and hence, we conducted pK_a prediction protocol. Two models were built. One used linear regression of residue depth, ASA, number of hydrogen bond and electrostatic. The second is the meta-algorithm of the first model and its complementary method, MMSCP. The later model showed the RMSD of only 0.7 pH units. Assigning protonation state of ionizable residues is important to understand the protein functions. And hence, this fast and accurate pK_a prediction tool could give biologist some ideas about the charge states of ionizable residues, as well as the total net charge of the protein, and hence, the protein functions.

In summary, using computational approaches the protein-peptide interactions have been tackled broadly. In each study, we have addressed several important problems. We believe that this work makes a positive contribution to the knowledge of protein-peptide interactions.

Publications

Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand binding cavities and predict the pKa of ionizable residues in proteins, K. P. Tan, T. B. Nguyen, S. Patel, R. Varadarajan, M.S. Madhusudhan, Nucl. Acids Res. (2013) 41 (W1): W314-W321.

DEMM: a meta-algorithm to predict the pK_a of ionizable amino acids in proteins. Nguyen T.B., Tan K.P., Madhusudhan M.S. IFMBE Proceedings, (2015) 46, 343-346.

Structural basis for CLIP-rich phenotype of HLA-DQ2.5, Nguyen T. B., Jayaraman P., Bergseng E., Madhusudhan M. S., Sollid L. M., Kim C-Y (under JBC re-revision).

DEMM server for pK_a prediction, Nguyen T. B., Tan K.P., Madhusudhan M.S. (in preparation)

Prediction of PPII receptors, Nguyen T. B., Madhusudhan M.S. (in preparation)

Mechanism of peptide editing from DR by DM catalyst; an MD study (in preparation), Nguyen T. B., Kim C-Y, Verma C.S., Madhusudhan M.S. (in preparation)

Bibliography

1. Petsalaki, E. & Russell, R.B. Peptide-mediated interactions in biological systems: new discoveries and applications. Current Opinion in Biotechnology **19**, 344-50 (2008).

2. Dutta, S., Chen, T.S. & Keating, A.E. Peptide ligands for pro-survival protein Bfl-1 from computationally guided library screening. *ACS Chemical Biology* **8**, 778-88 (2013).

3. Ilari, A. & Savino, C. Protein structure determination by x-ray crystallography. *Methods in Molecular Biology* **452**, 63-87 (2008).

4. Norton, R.S. Nuclear magnetic resonance (NMR) spectroscopy: applications to protein structure and engineering. *Australian Journal of Biotechnology* **4**, 114-20 (1990).

5. Cunningham, B.C. & Wells, J.A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081-5 (1989).

6. Domon, B. & Aebersold, R. Mass Spectrometry and Protein Analysis. *Science* **312**, 212 (2006).

7. Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G. & Gibson, T.J. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Frontiers Bioscience* **13**, 6580-603 (2008).

8. Cresswell, P. Assembly, transport, and function of MHC class II molecules. *Annual Review of Immunology* **12**, 259-93 (1994).

9. Roche, P.A., Marks, M.S. & Cresswell, P. Formation of a nine-subunit complex by HLA class II glycoproteins and the invariant chain. *Nature* **354**, 392-4 (1991).

10. Brodsky, F.M. & Guagliardi, L.E. The cell biology of antigen processing and presentation. *Annual Review of Immunology* **9**, 707-44 (1991).

11. Riberdy, J.M., Newcomb, J.R., Surman, M.J., Barbosa, J.A. & Cresswell, P. HLA-DR molecules from an antigen-processing mutant cell line are associated with invariant chain peptides. *Nature* **360**, 474-7 (1992).

12. Sette, A., Southwood, S., Miller, J. & Appella, E. Binding of major histocompatibility complex class II to the invariant chain-derived peptide, CLIP, is regulated by allelic polymorphism in class II. *Journal of Experimental Medicine* **181**, 677-83 (1995).

13. Sloan, V.S., Cameron, P., Porter, G., Gammon, M., Amaya, M., Mellins, E. & Zaller, D.M. Mediation by HLA-DM of dissociation of peptides from HLA-DR. *Nature* **375**, 802-6 (1995).

14. Germain, R.N. & Rinker, A.G., Jr. Peptide binding inhibits protein aggregation of invariant-chain free class II dimers and promotes surface expression of occupied molecules. *Nature* **363**, 725-8 (1993).

15. Wiesner, M., Stepniak, D., de Ru, A.H., Moustakis, A.K., Drijfhout, J.W., Papadopoulos, G.K., van Veelen, P.A. & Koning, F. Dominance of an alternative CLIP sequence in the celiac disease associated HLA-DQ2 molecule. *Immunogenetics* **60**, 551-5 (2008).

16. Fallang, L.E., Roh, S., Holm, A., Bergseng, E., Yoon, T., Fleckenstein, B., Bandyopadhyay, A., Mellins, E.D. & Sollid, L.M. Complexes of two cohorts of CLIP peptides and HLA-DQ2 of the autoimmune DR3-DQ2 haplotype are poor substrates for HLA-DM. *Journal of Immunology* **181**, 5451-61 (2008).

17. Bergseng, E., Dorum, S., Arntzen, M.O., Nielsen, M., Nygard, S., Buus, S., de Souza, G.A. & Sollid, L.M. Different binding motifs of the celiac disease-associated HLA molecules DQ2.5, DQ2.2, and DQ7.5 revealed by relative quantitative proteomics of

endogenous peptide repertoires. *Immunogenetics* **67**, 73-84 (2015).

18. Karell, K., Louka, A.S., Moodie, S.J., Ascher, H., Clot, F., Greco, L., Ciclitira, P.J., Sollid, L.M. & Partanen, J. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. *Human Immunology* **64**, 469-77 (2003).

19. Sollid, L.M., Markussen, G., Ek, J., Gjerde, H., Vartdal, F. & Thorsby, E. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *Journal of Experimental Medicine* **169**, 345-50 (1989).

20. Todd, J.A., Bell, J.I. & McDevitt, H.O. HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* **329**, 599-604 (1987).

21. Trier, J.S. Celiac Sprue. *New England Journal of Medicine* **325**, 1709-1719 (1991).

22. Kim, C.Y., Quarsten, H., Bergseng, E., Khosla, C. & Sollid, L.M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. *Proceedings of National Academy of Sciences USA* **101**, 4175-9 (2004).

23. Petersen, J., Montserrat, V., Mujico, J.R., Loh, K.L., Beringer, D.X., van Lummel, M., Thompson, A., Mearin, M.L., Schweizer, J., Kooy-Winkelaar, Y., van Bergen, J., Drijfhout, J.W., Kan, W.T., La Gruta, N.L., Anderson, R.P., Reid, H.H., Koning, F. & Rossjohn, J. T-cell receptor recognition of HLA-DQ2-gliadin complexes associated with celiac disease. *Nature Structural & Molecular Biology* **21**, 480-8 (2014).

24. Stepniak, D., Wiesner, M., de Ru, A.H., Moustakas, A.K., Drijfhout, J.W., Papadopoulos, G.K., van Veelen, P.A. & Koning, F. Large-scale characterization of natural ligands explains the unique gluten-binding properties of HLA-DQ2. *Journal of Immunology* **180**, 3268-78 (2008).

25. Busch, R., De Riva, A., Hadjinicolaou, A.V., Jiang, W., Hou, T. & Mellins, E.D. On the perils of poor editing: regulation of peptide loading by HLA-DQ and H2-A molecules associated with celiac disease and type 1 diabetes. Expert Reviews in Molecular Medicine **14**, e15 (2012).

26. Vartdal, F., Johansen, B.H., Friede, T., Thorpe, C.J., Stevanović, S., Eriksen, J.E., Sletten, K., Thorsby, E., Rammensee, H.-G. & Sollid, L.M. The peptide binding motif of the disease associated HLA-DQ (α 1* 0501, β 1* 0201) molecule. *Eur. Journal of Immunology* **26**, 2764-2772 (1996).

27. Chicz, R.M., Urban, R.G., Lane, W.S., Gorga, J.C., Stern, L.J., Vignali, D.A. & Strominger, J.L. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* **358**, 764-8 (1992).

28. Zhou, Z., Reyes-Vargas, E., Escobar, H., Rudd, B., Rockwood, A.L., Delgado, J.C., He, X. & Jensen, P.E. Type 1 diabetes associated HLA-DQ2 and DQ8 molecules are relatively resistant to HLA-DM mediated release of invariant chain-derived CLIP peptides. *Eur. Journal of Immunology* (2015).

29. Quarsten, H., McAdam, S.N., Jensen, T., Arentz-Hansen, H., Molberg, O., Lundin, K.E. & Sollid, L.M. Staining of celiac disease-relevant T cells by peptide-DQ2 multimers. *Journal of Immunology* **167**, 4861-8 (2001).

30. Kozono, H., White, J., Clements, J., Marrack, P. & Kappler, J. Production of soluble MHC class II proteins with covalently bound single peptides. *Nature* **369**, 151-4 (1994).

31. Kalandadze, A., Galleno, M., Foncerrada, L., Strominger, J.L. & Wucherpfennig, K.W. Expression of recombinant HLA-DR2 molecules. Replacement of the hydrophobic transmembrane region by a leucine zipper dimerization motif allows the assembly and secretion of soluble DR alpha beta heterodimers. *Journal of Biological Chemistry* **271**, 20156-62 (1996).

32. Scott, C.A., Garcia, K.C., Carbone, F.R., Wilson, I.A. & Teyton, L. Role of chain

pairing for the production of functional soluble IA major histocompatibility complex class II molecules. *Journal of Experimental Medicine* **183**, 2087-95 (1996).

33. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. Vol. 276 307-326 (Elsevier, 1997).

34. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. & Read, R.J. Phaser crystallographic software. *Journal of Applied Crystallography* **40**, 658-674 (2007).

35. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallographica Section D* **66**, 486-501 (2010).

36. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallographica Section D* **53**, 240-255 (1997).

37. Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C. & Zwart, P.H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography* **66**, 213-21 (2010).

38. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**, 283-291 (1993).

39. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Science.* **Chapter 2**, Unit 2 9 (2007).

40. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **234**, 779-815 (1993).

41. Shen, M.Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Science* **15**, 2507-24 (2006).

42. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. & Ferrin, T.E. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605-12 (2004).

43. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.7. (2014).

44. Li, Y., Yang, Y., He, P. & Yang, Q. QM/MM study of epitope peptides binding to HLA-A*0201: the roles of anchor residues and water. *Chemical Biology & Drug Design* **74**, 611-8 (2009).

45. Petrone, P.M. & Garcia, A.E. MHC-peptide binding is assisted by bound water molecules. *Journal of Molecular Biology* **338**, 419-35 (2004).

46. Ogata, K. & Wodak, S.J. Conserved water molecules in MHC class-I molecules and their putative structural and functional roles. *Protein Engineering* **15**, 697-705 (2002).

47. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712-25 (2006).

48. Case, D.A., Darden, T.A., Cheatham, T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Walker, R.C., Zhang, W., Merz, K.M., Roberts, B., Hayik, S., Roitberg, A., Seabra, G., Swails, J., Goetz, A.W., Kolossváry, I., Wong, K.F., Paesani, F., Vanicek, J., Wolf, R.M., Liu, J., Wu, X., Brozell, S.R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Hsieh, M.J., Cui, G., Roe, D.R., Mathews, D.H., Seetin, M.G., Salomon-Ferrer, R., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A. & Kollman, P.A. AMBER 12. (University of California, San Francisco, 2012).

49. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **98**, 10089-10092 (1993).

50. Voronoi, G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die reine und angewandte Mathematik* **133**, 97 (1907).

51. Ghosh, P., Amaya, M., Mellins, E. & Wiley, D.C. The structure of an intermediate in class II MHC maturation: CLIP bound to HLA-DR3. *Nature* **378**, 457-62 (1995).

52. Gunther, S., Schlundt, A., Sticht, J., Roske, Y., Heinemann, U., Wiesmuller, K.H., Jung, G., Falk, K., Rotzschke, O. & Freund, C. Bidirectional binding of invariant chain peptides to an MHC class II molecule. *Proceedings of the National Academy of Sciences USA* **107**, 22219-24 (2010).

53. Zhu, Y., Rudensky, A.Y., Corper, A.L., Teyton, L. & Wilson, I.A. Crystal structure of MHC class II I-Ab in complex with a human CLIP peptide: prediction of an I-Ab peptide-binding motif. *Journal of Molecular Biology* **326**, 1157-74 (2003).

54. Camacho, C.J., Weng, Z., Vajda, S. & DeLisi, C. Free energy landscapes of encounter complexes in protein-protein association. *Biophysical Journal* **76**, 1166-78 (1999).

55. Drozdov-Tikhomirov, L.N., Linde, D.M., Poroikov, V.V., Alexandrov, A.A. & Skurida, G.I. Molecular mechanisms of protein-protein recognition: whether the surface placed charged residues determine the recognition process? *Journal of Biomolecular Structure and Dynamics* **19**, 279-84 (2001).

56. Chu, X., Wang, Y., Gan, L., Bai, Y., Han, W., Wang, E. & Wang, J. Importance of electrostatic interactions in the association of intrinsically disordered histone chaperone Chz1 and histone H2A.Z-H2B. *PLoS Computational Biology* **8**, e1002608 (2012).

57. Kumar, V., Dixit, N., Zhou, L.L. & Fraunhofer, W. Impact of short range hydrophobic interactions and long range electrostatic forces on the aggregation kinetics of a monoclonal antibody and a dual-variable domain immunoglobulin at low and high concentrations. *International Journal of Pharmaceutics* **421**, 82-93 (2011).

58. Uchikoga, N., Takahashi, S.Y., Ke, R., Sonoyama, M. & Mitaku, S. Electric charge balance mechanism of extended soluble proteins. *Protein Science.* **14**, 74-80 (2005).

59. Wal, Y., Kooy, Y.C., Drijfhout, J., Amons, R. & Koning, F. Peptide binding characteristics of the coeliac disease-associated DQ(α1*0501, β1*0201) molecule. *Immunogenetics* **44**, 246-253 (1996).

60. Pos, W., Sethi, D.K., Call, M.J., Schulze, M.S., Anders, A.K., Pyrdol, J. & Wucherpfennig, K.W. Crystal structure of the HLA-DM-HLA-DR1 complex defines mechanisms for rapid peptide selection. *Cell* **151**, 1557-68 (2012).

61. Doebele, R.C., Busch, R., Scott, H.M., Pashine, A. & Mellins, E.D. Determination of the HLA-DM interaction site on HLA-DR molecules. *Immunity* **13**, 517-27 (2000).

62. Painter, C.A., Negroni, M.P., Kellersberger, K.A., Zavala-Ruiz, Z., Evans, J.E. & Stern, L.J. Conformational lability in the class II MHC 310 helix and adjacent extended strand dictate HLA-DM susceptibility and peptide exchange. *Proceedings of National Academy of Sciences USA* **108**, 19329-34 (2011).

63. Hou, T., Macmillan, H., Chen, Z., Keech, C.L., Jin, X., Sidney, J., Strohman, M., Yoon, T. & Mellins, E.D. An insertion mutant in DQA1*0501 restores susceptibility to HLA-DM: implications for disease associations. *Journal of Immunology* **187**, 2442-52 (2011).

64. Stratikos, E., Wiley, D.C. & Stern, L.J. Enhanced catalytic action of HLA-DM on the exchange of peptides lacking backbone hydrogen bonds between their N-terminal region and the MHC class II alpha-chain. *Journal of Immunology* **172**, 1109-17 (2004).

65. Yin, L., Trenh, P., Guce, A., Wieczorek, M., Lange, S., Sticht, J., Jiang, W., Bylsma, M., Mellins, E.D., Freund, C. & Stern, L.J. Susceptibility to HLA-DM protein is determined by a dynamic conformation of major histocompatibility complex class II molecule bound with peptide. *Journal of Biological Chemistry* **289**, 23449-64 (2014).

66. Schlundt, A., Gunther, S., Sticht, J., Wieczorek, M., Roske, Y., Heinemann, U. & Freund, C. Peptide linkage to the alpha-subunit of MHCII creates a stably inverted antigen presentation complex. *Journal of Molecular Biology* **423**, 294-302 (2012).

67. Bergseng, E., Xia, J., Kim, C.Y., Khosla, C. & Sollid, L.M. Main chain hydrogen bond interactions in the binding of proline-rich gluten peptides to the celiac disease-associated HLA-DQ2 molecule. *Journal of Biological Chemistry* **280**, 21791-6 (2005).

68. Jardetzky, T.S., Gorga, J.C., Busch, R., Rothbard, J., Strominger, J.L. & Wiley, D.C. Peptide binding to HLA-DR1: a peptide with most residues substituted to alanine retains MHC binding. *EMBO Journal* **9**, 1797-803 (1990).

69. Falk, K., Rotzschke, O., Stevanovic, S., Jung, G. & Rammensee, H.G. Pool sequencing of natural HLA-DR, DQ, and DP ligands reveals detailed peptide motifs, constraints of processing, and general rules. *Immunogenetics* **39**, 230-42 (1994).

70. Vartdal, F., Johansen, B.H., Friede, T., Thorpe, C.J., Stevanovic, S., Eriksen, J.E., Sletten, K., Thorsby, E., Rammensee, H.G. & Sollid, L.M. The peptide binding motif of the disease associated HLA-DQ (alpha 1* 0501, beta 1* 0201) molecule. *European Journal of Immunology* **26**, 2764-72 (1996).

71. van de Wal, Y., Kooy, Y.M., Drijfhout, J.W., Amons, R. & Koning, F. Peptide binding characteristics of the coeliac disease-associated DQ(alpha1*0501, beta1*0201) molecule. *Immunogenetics* **44**, 246-53 (1996).

72. Karell, K., Louka, A.S., Moodie, S.J., Ascher, H., Clot, F., Greco, L., Ciclitira, P.J., Sollid, L.M. & Partanen, J. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. *Human Immunology* **64**, 469-77 (2003).

73. Volz, T., Schwarz, G., Fleckenstein, B., Schepp, C.P., Haug, M., Roth, J., Wiesmuller, K.H. & Dannecker, G.E. Determination of the peptide binding motif and high-affinity ligands for HLA-DQ4 using synthetic peptide libraries. *Human Immunology* **65**, 594-601 (2004).

74. Tinto, N., Cola, A., Piscopo, C., Capuano, M., Galatola, M., Greco, L. & Sacchetti, L. High Frequency of Haplotype HLA-DQ7 in Celiac Disease Patients from South Italy: Retrospective Evaluation of 5,535 Subjects at Risk of Celiac Disease. *PLoS One* **10**, e0138324 (2015).

75. Kropshofer, H., Hammerling, G.J. & Vogt, A.B. How HLA-DM edits the MHC class II peptide repertoire: survival of the fittest? *Immunology Today* **18**, 77-82 (1997).

76. Katz, J.F., Stebbins, C., Appella, E. & Sant, A.J. Invariant chain and DM edit self-peptide presentation by major histocompatibility complex (MHC) class II molecules. *Journal of Experimental Medicine* **184**, 1747-53 (1996).

77. Weber, D.A., Evavold, B.D. & Jensen, P.E. Enhanced dissociation of HLA-DR-bound peptides in the presence of HLA-DM. *Science* **274**, 618-20 (1996).

78. Rabinowitz, J.D., Vrljic, M., Kasson, P.M., Liang, M.N., Busch, R., Boniface, J.J., Davis, M.M. & McConnell, H.M. Formation of a highly peptide-receptive state of class II MHC. *Immunity* **9**, 699-709 (1998).

79. Yaneva, R., Springer, S. & Zacharias, M. Flexibility of the MHC class II peptide binding cleft in the bound, partially filled, and empty states: a molecular dynamics simulation study. *Biopolymers* **91**, 14-27 (2009).

80. E. Fermi , J.P., S. Ulam. *Los Alamos Scientific Laboratory report LA-1940* (1955).

81. B. J. Alder, T.E.W. Phase Transition for a Hard Sphere System *The Journal of Chemical Physics* **27**, 1208 (1957).

82. Alder, B.J., Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics* **31**, 459 (1959).

83. Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Physical Review* **136**, A405 (1964).

84. Stillinger, F.H., Rahman, A. Improved Simulation of Liquid Water by Molecular Dynamics. *Journal of Chemical Physics* **60**, 1545 (1974).

85. McCammon, J.A., Gelin, B.R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585-90 (1977).

86. Berendsen, H.J.C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91**, 43-56 (1995).

87. Berendsen H. J. C., P.J.P.M., van Gunsteren W. F., Hermans J. (ed.) *Interaction Models for Water in Relation to Protein Hydration*, (Springer Netherlands, 1981).

88. Tan, K.P., Nguyen, T.B., Patel, S., Varadarajan, R. & Madhusudhan, M.S. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Research* **41**, W314-21 (2013).

89. Nguyen, T.B., Tan, K.P. & Madhusudhan, M.S. DEMM: A Meta-Algorithm to Predict the pKa of Ionizable Amino Acids in Proteins. in *5th International Conference on Biomedical Engineering in Vietnam* (eds. Toi, V.V. & Lien Phuong, H.T.) 343-346 (Springer International Publishing, Cham, 2015).

90. Tom Darden, D.Y., and Lee Pederse. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems *The Journal of Chemical Physics* **98**(1993).

91. Hess Berk, B.H., Berendsen Herman J. C., Fraaije Johannes G. E. M. . LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**(1997).

92. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. & Caves, L.S. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**, 2695-6 (2006).

93. Humphrey, W., Dalke, A. and Schulten, K. VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 6 (1996).

94. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637 (1983).

95. Anders, A.K., Call, M.J., Schulze, M.S., Fowler, K.D., Schubert, D.A., Seth, N.P., Sundberg, E.J. & Wucherpfennig, K.W. HLA-DM captures partially empty HLA-DR molecules for catalyzed removal of peptide. *Nature Immunology* **12**, 54-61 (2011).

96. Painter, C.A., Negroni, M.P., Kellersberger, K.A., Zavala-Ruiz, Z., Evans, J.E. & Stern, L.J. Conformational lability in the class II MHC 310 helix and adjacent extended strand dictate HLA-DM susceptibility and peptide exchange. *Proceedings of the National Academy of Sciences USA* **108**, 19329-34 (2011).

97. Carven, G.J. & Stern, L.J. Probing the ligand-induced conformational change in HLA-DR1 by selective chemical modification and mass spectrometric mapping. *Biochemistry* **44**, 13625-37 (2005).

98. Walensky, L.D. & Bird, G.H. Hydrocarbon-stapled peptides: principles, practice, and progress. *Journal of Medical Chemistry* **57**, 6275-88 (2014).

99. Adzhubei, A.A. & Sternberg, M.J. Left-handed polyproline II helices commonly occur in globular proteins. *Journal of Molecular Biology* **229**, 472-93 (1993).

100. Creamer, T.P. Left-handed polyproline II helix formation is (very) locally driven. *Proteins* **33**, 218-26 (1998).

101. Stapley, B.J. & Creamer, T.P. A survey of left-handed polyproline II helices. *Protein Science* **8**, 587-95 (1999).

102. Creamer, T.P. & Campbell, M.N. Determinants of the polyproline II helix from modeling studies. *Advances in Protein Chemistry* **62**, 263-82 (2002).

103. Chellgren, B.W. & Creamer, T.P. Short sequences of non-proline residues can adopt the polyproline II helical conformation. *Biochemistry* **43**, 5864-9 (2004).

104. Chellgren, B.W., Miller, A.F. & Creamer, T.P. Evidence for polyproline II helical structure in short polyglutamine tracts. *Journal of Molecular Biology* **361**, 362-71 (2006).

105. Ramachandran, G.N. & Sasisekharan, V. Conformation of polypeptides and proteins. *Advances in Protein Chemistry* **23**, 283-438 (1968).

106. Traub, W. & Shmueli, U. Structure of Poly-L-Proline I. *Nature* **198**, 1165-1166 (1963).

107. Sasisekharan, V. Structures of poly-L-proline II. *Acta Crystallographica* **12**(1959).

108. Adzhubei, A.A., Sternberg, M.J. & Makarov, A.A. Polyproline-II helix in proteins: structure and function. *Journal of Molecular Biology* **425**, 2100-32 (2013).

109. Pauling, L. & Corey, R.B. The structure of fibrous proteins of the collagen-gelatin group. *Proceedings of the National Academy of Sciences USA* **37**, 272-81 (1951).

110. Cowan, P.M., McGavin, S. & North, A.C. The polypeptide chain configuration of collagen. *Nature* **176**, 1062-4 (1955).

111. Ananthanarayanan, V.S., Soman, K.V. & Ramakrishnan, C. A novel supersecondary structure in globular proteins comprising the collagen-like helix and beta-turn. *Journal of Molecular Biology* **198**, 705-9 (1987).

112. Berisio, R. & Vitagliano, L. Polyproline and triple helix motifs in host-pathogen recognition. *Current Protein & Peptide Science* **13**, 855-65 (2012).

113. Polverini, E., Rangaraj, G., Libich, D.S., Boggs, J.M. & Harauz, G. Binding of the proline-rich segment of myelin basic protein to SH3 domains: spectroscopic, microarray, and modeling studies of ligand conformation and effects of posttranslational modifications. *Biochemistry* **47**, 267-82 (2008).

114. Peterson, F.C. & Volkman, B.F. Diversity of polyproline recognition by EVH1 domains. *Frontiers BioscienceFrontiers Bioscience (Landmark Ed)* **14**, 833-46 (2009).

115. Kursula, P., Kursula, I., Massimi, M., Song, Y.H., Downer, J., Stanley, W.A., Witke, W. & Wilmanns, M. High-resolution structural analysis of mammalian profilin 2a complex formation with two physiological ligands: the formin homology 1 domain of mDia1 and the proline-rich domain of VASP. *Journal of Molecular Biology* **375**, 270-90 (2008).

116. Zarrinpar, A., Bhattacharyya, R.P. & Lim, W.A. The structure and function of proline recognition domains. *Science Signaling* **2003**, RE8 (2003).

117. Liu, Y., Chen, W., Gaudet, J., Cheney, M.D., Roudaia, L., Cierpicki, T., Klet, R.C., Hartman, K., Laue, T.M., Speck, N.A. & Bushweller, J.H. Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity. *Cancer Cell* **11**, 483-97 (2007).

118. Cole, D.K., Bulek, A.M., Dolton, G., Schauenberg, A.J., Szomolay, B., Rittase, W., Trimby, A., Jothikumar, P., Fuller, A., Skowera, A., Rossjohn, J., Zhu, C., Miles, J.J., Peakman, M., Wooldridge, L., Rizkallah, P.J. & Sewell, A.K. Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity. *Journal of Clinical Investigation* **126**, 2191-204 (2016).

119. Kundu, K., Mann, M., Costa, F. & Backofen, R. MoDPepInt: an interactive web server for prediction of modular domain-peptide interactions. *Bioinformatics* **30**, 2668-9 (2014).

120. Donnes, P. & Kohlbacher, O. SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Research* **34**, W194-7 (2006).

121. Petsalaki, E., Stark, A., Garcia-Urdiales, E. & Russell, R.B. Accurate prediction of peptide binding sites on protein surfaces. *PLoS Computational Biology* **5**, e1000335

(2009).

122. London, N., Raveh, B., Cohen, E., Fathi, G. & Schueler-Furman, O. Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions. *Nucleic Acids Research* **39**, W249-53 (2011).

123. Lee, H., Heo, L., Lee, M.S. & Seok, C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Research* **43**, W431-5 (2015).

124. Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A. & Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Research* **43**, W419-24 (2015).

125. Nguyen, M.N. & Madhusudhan, M.S. Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Research* **39**, e94 (2011).

126. Nguyen, M.N., Tan, K.P. & Madhusudhan, M.S. CLICK--topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Research* **39**, W24-8 (2011).

127. Bernhard Schölkopf, K.T., Jean-Philippe Vert (ed.) *Kernel Methods in Computational Biology*, (MIT press, 2004).

128. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).

129. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-402 (1997).

130. Fabian Pedregosa, G.V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**(2011).

131. Wang, G. & Dunbrack, R.L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-91 (2003).

132. Wang, G. & Dunbrack, R.L., Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research* **33**, W94-8 (2005).

133. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* **247**, 536-40 (1995).

134. Sonnhammer, E.L., Eddy, S.R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-20 (1997).

135. Kwan, J.J. & Donaldson, L.W. A lack of peptide binding and decreased thermostability suggests that the CASKIN2 scaffolding protein SH3 domain may be vestigial. *BMC Structural Biology* **16**, 14 (2016).

136. Chong, P.A., Lin, H., Wrana, J.L. & Forman-Kay, J.D. An expanded WW domain recognition motif revealed by the interaction between Smad7 and the E3 ubiquitin ligase Smurf2. *Journal of Biological Chemistry* **281**, 17069-75 (2006).

137. Haywood, J., Qi, J., Chen, C.C., Lu, G., Liu, Y., Yan, J., Shi, Y. & Gao, G.F. Structural basis of collagen recognition by human osteoclast-associated receptor and design of osteoclastogenesis inhibitors. *Proceedings of the National Academy of Sciences USA* **113**, 1038-43 (2016).

138. Prior, S.H., Byrne, T.S., Tokmina-Roszyk, D., Fields, G.B. & Van Doren, S.R. Path to Collagenolysis: collagen V triple-helix model bound productively and in encounters by matrix metalloproteinase-12. *Journal of Biological Chemistry* **291**, 7888-901 (2016).

139. Hotta, K., Ranganathan, S., Liu, R., Wu, F., Machiyama, H., Gao, R., Hirata, H.,

Soni, N., Ohe, T., Hogue, C.W., Madhusudhan, M.S. & Sawada, Y. Biophysical properties of intrinsically disordered p130Cas substrate domain--implication in mechanosensing. *PLoS Computational Biology* **10**, e1003532 (2014).

140. Dunker, M.S.a.A.K. Proline Rich Motifs as Drug Targets in Immune Mediated Disorders. *International Journal of Peptides* **2012**(2012).

141. Vitali, A. Proline-rich peptides: multifunctional bioactive molecules as new potential therapeutic drugs. *Current Protein & Peptide Science* **16**, 147-62 (2015).

142. Bartlett, G.J., Porter, C.T., Borkakoti, N. & Thornton, J.M. Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology* **324**, 105-21 (2002).

143. Garcia-Moreno, E.B. & Fitch, C.A. Structural interpretation of pH and salt-dependent processes in proteins with computational methods. *Methods Enzymology* **380**, 20-51 (2004).

144. Hendsch, Z.S., Jonsson, T., Sauer, R.T. & Tidor, B. Protein stabilization by removal of unsatisfied polar groups: computational approaches and experimental tests. *Biochemistry* **35**, 7621-5 (1996).

145. Schaefer, M., Sommer, M. & Karplus, M. pH-Dependence of Protein Stability: Absolute Electrostatic Free Energy Differences between Conformations†. *The Journal of Physical Chemistry B* **101**, 1663-1683 (1997).

146. Tjong, H. & Zhou, H.X. Prediction of protein solubility from calculation of transfer free energy. *Biophysical Journal* **95**, 2601-9 (2008).

147. Maciej, D. & Jan, M.A. The impact of protonation equilibria on protein structure. *Journal of Physics: Condensed Matter* **17**, S1607 (2005).

148. Warshel, A. Calculations of enzymatic reactions: calculations of pKa, proton transfer reactions, and general acid catalysis reactions in enzymes. *Biochemistry* **20**, 3167-77 (1981).

149. Castaneda, C.A., Fitch, C.A., Majumdar, A., Khangulov, V., Schlessman, J.L. & Garcia-Moreno, B.E. Molecular determinants of the pKa values of Asp and Glu residues in staphylococcal nuclease. *Proteins* **77**, 570-88 (2009).

150. Bartik, K., Redfield, C. & Dobson, C.M. Measurement of the individual pKa values of acidic residues of hen and turkey lysozymes by two-dimensional 1H NMR. *Biophysical Journal* **66**, 1180-4 (1994).

151. Oda, Y., Yamazaki, T., Nagayama, K., Kanaya, S., Kuroda, Y. & Nakamura, H. Individual ionization constants of all the carboxyl groups in ribonuclease HI from Escherichia coli determined by NMR. *Biochemistry* **33**, 5275-84 (1994).

152. Oda, Y., Yoshida, M. & Kanaya, S. Role of histidine 124 in the catalytic function of ribonuclease HI from Escherichia coli. *Journal of Biological Chemistry* **268**, 88-92 (1993).

153. Li, L., Li, C., Zhang, Z. & Alexov, E. On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *Journal of Chemical Theory and Computation* **9**, 2126-2136 (2013).

154. Markley, J.L. Observation of histidine residues in proteins by nuclear magnetic resonance spectroscopy. *Accounts of Chemical Research* **8**, 70-80 (1975).

155. Matthew, J.B., Gurd, F.R., Garcia-Moreno, B., Flanagan, M.A., March, K.L. & Shire, S.J. pH-dependent processes in proteins. *CRC Critical Review in Biochemistry* **18**, 91-197 (1985).

156. Forsyth, W.R., Antosiewicz, J.M. & Robertson, A.D. Empirical relationships between protein structure and carboxyl pKa values in proteins. *Proteins* **48**, 388-403 (2002).

157. Giletto, A. & Pace, C.N. Buried, charged, non-ion-paired aspartic acid 76 contributes favorably to the conformational stability of ribonuclease T1. *Biochemistry* **38**, 13379-84 (1999).

158. Baran, K.L., Chimenti, M.S., Schlessman, J.L., Fitch, C.A., Herbst, K.J. & Garcia-Moreno, B.E. Electrostatic effects in a network of polar and ionizable groups in staphylococcal nuclease. *Journal of Molecular Biology* **379**, 1045-62 (2008).

159. Dwyer, J.J., Gittis, A.G., Karp, D.A., Lattman, E.E., Spencer, D.S., Stites, W.E. & Garcia-Moreno, E.B. High apparent dielectric constants in the interior of a protein reflect water penetration. *Biophysical Journal* **79**, 1610-20 (2000).

160. Fitch, C.A., Karp, D.A., Lee, K.K., Stites, W.E., Lattman, E.E. & Garcia-Moreno, E.B. Experimental pK(a) values of buried residues: analysis with continuum methods and role of water penetration. *Biophysical Journal* **82**, 3289-304 (2002).

161. Whitten, S.T. & Garcia-Moreno, E.B. pH dependence of stability of staphylococcal nuclease: evidence of substantial electrostatic interactions in the denatured state. *Biochemistry* **39**, 14292-304 (2000).

162. Khandogin, J. & Brooks, C.L., 3rd. Toward the accurate first-principles prediction of ionization equilibria in proteins. *Biochemistry* **45**, 9363-73 (2006).

163. Baker, N., Holst, M. & Wang, F. Adaptive multilevel finite element solution of the Poisson–Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems. *Journal of Computational Chemistry* **21**, 1343-1352 (2000).

164. Baker, N.A., Sept, D., Joseph, S., Holst, M.J. & McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences USA* **98**, 10037-41 (2001).

165. Lu, B., Cheng, X., Huang, J. & McCammon, J.A. An Adaptive Fast Multipole Boundary Element Method for Poisson-Boltzmann Electrostatics. *Journal of Chemical Theory and Computation* **5**, 1692-1699 (2009).

166. Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M. & Karplus, M. CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry* **30**, 1545-614 (2009).

167. Warwicker, J. & Watson, H.C. Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *Journal of Molecular Biology* **157**, 671-9 (1982).

168. Alexov, E.G. & Gunner, M.R. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophysical Journal* **72**, 2075-93 (1997).

169. Still, W.C., Tempczyk, A., Hawley, R.C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **112**, 6127-6129 (1990).

170. Im, W., Lee, M.S. & Brooks, C.L., 3rd. Generalized born model with a simple smoothing function. *Journal of Computational Chemistry* **24**, 1691-702 (2003).

171. Sigalov, G., Fenley, A. & Onufriev, A. Analytical electrostatics for biomolecules: beyond the generalized Born approximation. *Journal of Chemical Physics* **124**, 124902 (2006).

172. Laurents, D.V., Huyghues-Despointes, B.M., Bruix, M., Thurlkill, R.L., Schell, D., Newsom, S., Grimsley, G.R., Shaw, K.L., Trevino, S., Rico, M., Briggs, J.M., Antosiewicz, J.M., Scholtz, J.M. & Pace, C.N. Charge-charge interactions are key determinants of the pK values of ionizable groups in ribonuclease Sa (pI=3.5) and a basic variant (pI=10.2). *Journal of Molecular Biology* **325**, 1077-92 (2003).

173. Forsyth, W.R., Gilson, M.K., Antosiewicz, J., Jaren, O.R. & Robertson, A.D. Theoretical and experimental analysis of ionization equilibria in ovomucoid third domain. *Biochemistry* **37**, 8643-52 (1998).

174. Forsyth, W.R. & Robertson, A.D. Insensitivity of perturbed carboxyl pK(a) values in

the ovomucoid third domain to charge replacement at a neighboring residue. *Biochemistry* **39**, 8067-72 (2000).

175. Jensen, J.H., Li, H., Robertson, A.D. & Molina, P.A. Prediction and rationalization of protein pKa values using QM and QM/MM methods. *Journal of Physical Chemistry AJournal of Physical Chemistry A* **109**, 6634-43 (2005).

176. Li, H., Robertson, A.D. & Jensen, J.H. The determinants of carboxyl pKa values in turkey ovomucoid third domain. *Proteins* **55**, 689-704 (2004).

177. Schaefer, P., Riccardi, D. & Cui, Q. Reliable treatment of electrostatics in combined QM/MM simulation of macromolecules. *Journal of Chemical Physics* **123**, 014905 (2005).

178. Khandogin, J. & Brooks, C.L., 3rd. Constant pH molecular dynamics with proton tautomerism. *Biophysical Journal* **89**, 141-57 (2005).

179. de Oliveira, C.A., Hamelberg, D. & McCammon, J.A. Coupling Accelerated Molecular Dynamics Methods with Thermodynamic Integration Simulations. *Journal of Chemical Theory and Computation* **4**, 1516-1525 (2008).

180. Williams, S.L., de Oliveira, C.A. & McCammon, J.A. Coupling Constant pH Molecular Dynamics with Accelerated Molecular Dynamics. *Journal of Chemical Theory and Computation* **6**, 560-568 (2010).

181. Mehler, E.L. The Lorentz-Debye-Sack theory and dielectric screening of electrostatic effects in proteins and nucleic acids. in *Theoretical and Computational Chemistry,* Vol. Volume 3 (eds. Jane, S.M. & Kalidas, S.) 371-405 (Elsevier, 1996).

182. Mehler, E.L. & Guarnieri, F. A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophysical Journal* **77**, 3-22 (1999).

183. Mehler, E.L. Self-Consistent, Free Energy Based Approximation To Calculate pH Dependent Electrostatic Effects in Proteins. *The Journal of Physical Chemistry* **100**, 16006-16018 (1996).

184. Shan, J. & Mehler, E.L. Calculation of pK(a) in proteins with the microenvironment modulated-screened coulomb potential. *Proteins* **79**, 3346-55 (2011).

185. Wisz, M.S. & Hellinga, H.W. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins* **51**, 360-77 (2003).

186. Li, H., Robertson, A.D. & Jensen, J.H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **61**, 704-21 (2005).

187. Bas, D.C., Rogers, D.M. & Jensen, J.H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **73**, 765-83 (2008).

188. Chakravarty, S. & Varadarajan, R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **7**, 723-32 (1999).

189. Tan, K.P., Varadarajan, R. & Madhusudhan, M.S. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Research* (2011).

190. Pedersen, T.G., Sigurskjold, B.W., Andersen, K.V., Kjaer, M., Poulsen, F.M., Dobson, C.M. & Redfield, C. A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution. *Journal of Molecular Biology* **218**, 413-26 (1991).

191. Pintar, A., Carugo, O. & Pongor, S. Atom depth as a descriptor of the protein interior. *Biophysical Journal* **84**, 2553-61 (2003).

192. Pintar, A., Carugo, O. & Pongor, S. Atom depth in protein structure and function. *Trends in Biochemical Sciences* **28**, 593-7 (2003).

193. Adkar, B.V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P.,

Swarnkar, M.K., Gokhale, R.S. & Varadarajan, R. Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* **20**, 371-81 (2012).

194. Word, J.M., Lovell, S.C., Richardson, J.S. & Richardson, D.C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* **285**, 1735-47 (1999).

195. van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P. & Tironi, I.G. *Biomolecular Simulation: The {GROMOS96} manual and userguide*, (Hochschuleverlag AG an der ETH Zürich, 1996).

196. Baker, E.N. & Hubbard, R.E. Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology* **44**, 97-179 (1984).

197. Tan, K.P. Characterization of Physicochemical Environments of Proteins (Doctoral dissertation). (2016).

198. Shrake, A. & Rupley, J.A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology* **79**, 351-71 (1973).

199. Lund-Katz, S., Wehrli, S., Zaiou, M., Newhouse, Y., Weisgraber, K.H. & Phillips, M.C. Effects of polymorphism on the microenvironment of the LDL receptor-binding region of human apoE. *Journal of Lipid Research* **42**, 894-901 (2001).

200. Lund-Katz, S., Zaiou, M., Wehrli, S., Dhanasekaran, P., Baldwin, F., Weisgraber, K.H. & Phillips, M.C. Effects of lipid interaction on the lysine microenvironments in apolipoprotein E. *Journal of Biological Chemistry* **275**, 34459-64 (2000).

201. Assadi-Porter, F.M. & Fillingame, R.H. Proton-translocating carboxyl of subunit c of F1Fo H(+)-ATP synthase: the unique environment suggested by the pKa determined by 1H NMR. *Biochemistry* **34**, 16186-93 (1995).

202. Garcia-Moreno, B., Dwyer, J.J., Gittis, A.G., Lattman, E.E., Spencer, D.S. & Stites, W.E. Experimental measurement of the effective dielectric in the hydrophobic core of a protein. *Biophysical Chemistry* **64**, 211-24 (1997).

203. Harris, T.K., Wu, G., Massiah, M.A. & Mildvan, A.S. Mutational, kinetic, and NMR studies of the roles of conserved glutamate residues and of lysine-39 in the mechanism of the MutT pyrophosphohydrolase. *Biochemistry* **39**, 1655-74 (2000).

204. Zhang, G., Mazurkie, A.S., Dunaway-Mariano, D. & Allen, K.N. Kinetic evidence for a substrate-induced fit in phosphonoacetaldehyde hydrolase catalysis. *Biochemistry* **41**, 13370-7 (2002).

205. Oliveberg, M., Arcus, V.L. & Fersht, A.R. pKA values of carboxyl groups in the native and denatured states of barnase: the pKA values of the denatured state are on average 0.4 units lower than those of model compounds. *Biochemistry* **34**, 9424-33 (1995).

206. Loewenthal, R., Sancho, J. & Fersht, A.R. Histidine-aromatic interactions in barnase. Elevation of histidine pKa and contribution to protein stability. *Journal of Molecular Biology* **224**, 759-70 (1992).

207. Khare, D., Alexander, P., Antosiewicz, J., Bryan, P., Gilson, M. & Orban, J. pKa measurements from nuclear magnetic resonance for the B1 and B2 immunoglobulin G-binding domains of protein G: comparison with calculated values for nuclear magnetic resonance and X-ray structures. *Biochemistry* **36**, 3580-9 (1997).

208. Ebina, S. & Wuthrich, K. Amide proton titration shifts in bull seminal inhibitor IIA by two-dimensional correlated 1H nuclear magnetic resonance (COSY). Manifestation of conformational equilibria involving carboxylate groups. *Journal of Molecular Biology* **179**, 283-8 (1984).

209. Fujii, S., Akasaka, K. & Hatano, H. Acid denaturation steps of Streptomyces subtilisin inhibitor. A proton magnetic resonance study of individual histidine environment. *Biochemistry* **88**, 789-96 (1980).

210. Kesvatera, T., Jonsson, B., Thulin, E. & Linse, S. Measurement and modelling of sequence-specific pKa values of lysine residues in calbindin D9k. *Journal of Molecular Biology* **259**, 828-39 (1996).

211. Chiang, C.M., Chang, S.L., Lin, H.J. & Wu, W.G. The role of acidic amino acid residues in the structural stability of snake cardiotoxins. *Biochemistry* **35**, 9177-86 (1996).

212. Chen, H.A., Pfuhl, M., McAlister, M.S. & Driscoll, P.C. Determination of pK(a) values of carboxyl groups in the N-terminal domain of rat CD2: anomalous pK(a) of a glutamate on the ligand-binding surface. *Biochemistry* **39**, 6814-24 (2000).

213. Tan, Y.J., Oliveberg, M., Davis, B. & Fersht, A.R. Perturbed pKA-values in the denatured states of proteins. *Journal of Molecular Biology* **254**, 980-92 (1995).

214. Yu, L. & Fesik, S.W. pH titration of the histidine residues of cyclophilin and FK506 binding protein in the absence and presence of immunosuppressant ligands. *Biochimica et Biophysica Acta* **1209**, 24-32 (1994).

215. Kohda, D., Sawada, T. & Inagaki, F. Characterization of pH titration shifts for all the nonlabile proton resonances a protein by two-dimensional NMR: the case of mouse epidermal growth factor. *Biochemistry* **30**, 4896-900 (1991).

216. Gooley, P.R., Keniry, M.A., Dimitrov, R.A., Marsh, D.E., Keizer, D.W., Gayler, K.R. & Grant, B.R. The NMR solution structure and characterization of pH dependent chemical shifts of the beta-elicitin, cryptogein. *Journal of Biomolecular NMR* **12**, 523-34 (1998).

217. Perez-Canadillas, J.M., Campos-Olivas, R., Lacadena, J., Martinez del Pozo, A., Gavilanes, J.G., Santoro, J., Rico, M. & Bruix, M. Characterization of pKa values and titration shifts in the cytotoxic ribonuclease alpha-sarcin by NMR. Relationship between electrostatic interactions, structure, and catalytic function. *Biochemistry* **37**, 15865-76 (1998).

218. Kuramitsu, S. & Hamaguchi, K. Analysis of the acid-base titration curve of hen lysozyme. *Biochemistry* **87**, 1215-9 (1980).

219. Szyperski, T., Antuch, W., Schick, M., Betz, A., Stone, S.R. & Wuthrich, K. Transient hydrogen bonds identified on the surface of the NMR solution structure of Hirudin. *Biochemistry* **33**, 9303-10 (1994).

220. Wang, Y.X., Freedberg, D.I., Yamazaki, T., Wingfield, P.T., Stahl, S.J., Kaufman, J.D., Kiso, Y. & Torchia, D.A. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochemistry* **35**, 9945-50 (1996).

221. Gao, G., DeRose, E.F., Kirby, T.W. & London, R.E. NMR determination of lysine pKa values in the Pol lambda lyase domain: mechanistic implications. *Biochemistry* **45**, 1785-94 (2006).

222. Jorgensen, A.M., Kristensen, S.M., Led, J.J. & Balschmidt, P. Three-dimensional solution structure of an insulin dimer. A study of the B9(Asp) mutant of human insulin using nuclear magnetic resonance, distance geometry and restrained molecular dynamics. *Journal of Molecular Biology* **227**, 1146-63 (1992).

223. Forman-Kay, J.D., Clore, G.M. & Gronenborn, A.M. Relationship between electrostatics and redox function in human thioredoxin: characterization of pH titration shifts using two-dimensional homo- and heteronuclear NMR. *Biochemistry* **31**, 3442-52 (1992).

224. Kao, Y.H., Fitch, C.A., Bhattacharya, S., Sarkisian, C.J., Lecomte, J.T. & Garcia-Moreno, E.B. Salt effects on ionization equilibria of histidines in myoglobin. *Biophysical Journal* **79**, 1637-54 (2000).

225. Bashford, D., Case, D.A., Dalvit, C., Tennant, L. & Wright, P.E. Electrostatic calculations of side-chain pK(a) values in myoglobin and comparison with NMR data for histidines. *Biochemistry* **32**, 8045-56 (1993).

226. March, K.L., Maskalick, D.G., England, R.D., Friend, S.H. & Gurd, F.R. Analysis of electrostatic interactions and their relationship to conformation and stability of bovine pancreatic trypsin inhibitor. *Biochemistry* **21**, 5241-51 (1982).

227. Anderson, D.E., Becktel, W.J. & Dahlquist, F.W. pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry* **29**, 2403-8 (1990).

228. Dao-pin, S., Anderson, D.E., Baase, W.A., Dahlquist, F.W. & Matthews, B.W. Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. *Biochemistry* **30**, 11521-9 (1991).

229. Heinz, D.W., Ryan, M., Smith, M.P., Weaver, L.H., Keana, J.F. & Griffith, O.H. Crystal structure of phosphatidylinositol-specific phospholipase C from Bacillus cereus in complex with glucosaminyl(alpha 1-->6)-D-myo-inositol, an essential fragment of GPI anchors. *Biochemistry* **35**, 9496-504 (1996).

230. Baker, W.R. & Kintanar, A. Characterization of the pH titration shifts of ribonuclease A by one- and two-dimensional nuclear magnetic resonance spectroscopy. *Archives Biochemistry Biophysics* **327**, 189-99 (1996).

231. Antosiewicz, J., McCammon, J.A. & Gilson, M.K. Prediction of pH-dependent properties of proteins. *Journal of Molecular Biology* **238**, 415-36 (1994).

232. Spitzner, N., Lohr, F., Pfeiffer, S., Koumanov, A., Karshikoff, A. & Ruterjans, H. Ionization properties of titratable groups in ribonuclease T1. I. pKa values in the native state determined by two-dimensional heteronuclear NMR spectroscopy. *Eur Biophysical Journal* **30**, 186-97 (2001).

233. Inagaki, F., Kawano, Y., Shimada, I., Takahashi, K. & Miyazawa, T. Nuclear magnetic resonance study on the microenvironments of histidine residues of ribonuclease T1 and carboxymethylated ribonuclease T1. *Biochemistry* **89**, 1185-95 (1981).

234. Norton, R.S., Cross, K., Braach-Maksvytis, V. & Wachter, E. 1H-n.m.r. study of the solution properties and secondary structure of neurotoxin III from the sea anemone Anemonia sulcata. *Biochemistry* **293 (Pt 2)**, 545-51 (1993).

235. Isom, D.G., Castañeda, C.A., Cannon, B.R., Velu, P.D. & García-Moreno E., B. Charges in the hydrophobic interior of proteins. *Proceedings of the National Academy of Sciences* **107**, 16096-16100 (2010).

236. Isom, D.G., Castañeda, C.A., Cannon, B.R. & García-Moreno E., B. Large shifts in pKa values of lysine residues buried inside a protein. *Proceedings of the National Academy of Sciences* **108**, 5260-5265 (2011).

237. Meyer, T., Kieseritzky, G. & Knapp, E.W. Electrostatic pKa computations in proteins: role of internal cavities. *Proteins* **79**, 3320-32 (2011).

238. Harms, M.J., Castaneda, C.A., Schlessman, J.L., Sue, G.R., Isom, D.G., Cannon, B.R. & Garcia-Moreno, E.B. The pK(a) values of acidic and basic residues buried at the same internal location in a protein are governed by different factors. *Journal of Molecular Biology* **389**, 34-47 (2009).

239. Harms, M.J., Schlessman, J.L., Chimenti, M.S., Sue, G.R., Damjanovic, A. & Garcia-Moreno, B. A buried lysine that titrates with a normal pKa: role of conformational flexibility at the protein-water interface as a determinant of pKa values. *Protein Science* **17**, 833-45 (2008).

240. Chimenti, M.S., Castaneda, C.A., Majumdar, A. & Garcia-Moreno, E.B. Structural origins of high apparent dielectric constants experienced by ionizable groups in the hydrophobic core of a protein. *Journal of Molecular Biology* **405**, 361-77 (2011).

241. Nguyen, D.M., Leila Reynald, R., Gittis, A.G. & Lattman, E.E. X-ray and thermodynamic studies of staphylococcal nuclease variants I92E and I92K: insights into polarity of the protein interior. *Journal of Molecular Biology* **341**, 565-74 (2004).

242. Lee, K.K., Fitch, C.A. & Garcia-Moreno, E.B. Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein. *Protein Science* **11**, 1004-16 (2002).

243. Inagaki, F., Miyazawa, T., Hori, H. & Tamiya, N. Conformation of erabutoxins a and b in aqueous solution as studied by nuclear magnetic resonance and circular dichroism. *European Journal of Biochemistry* **89**, 433-42 (1978).

244. Zhou, M.M., Davis, J.P. & Van Etten, R.L. Identification and pKa determination of the histidine residues of human low-molecular-weight phosphotyrosyl protein phosphatases: a convenient approach using an MLEV-17 spectral editing scheme. *Biochemistry* **32**, 8479-86 (1993).

245. Schaller, W. & Robertson, A.D. pH, ionic strength, and temperature dependences of ionization equilibria for the carboxyl groups in turkey ovomucoid third domain. *Biochemistry* **34**, 4714-23 (1995).

246. Swint-Kruse, L. & Robertson, A.D. Hydrogen bonds and the pH dependence of ovomucoid third domain stability. *Biochemistry* **34**, 4724-32 (1995).

247. Ogino, T., Croll, D.H., Kato, I. & Markley, J.L. Properties of conserved amino acid residues in tandem homologous protein domains. Hydrogen-1 nuclear magnetic resonance studies of the histidines of chicken ovomucoid. *Biochemistry* **21**, 3452-60 (1982).

248. Betz, M., Lohr, F., Wienk, H. & Ruterjans, H. Long-range nature of the interactions between titratable groups in Bacillus agaradhaerens family 11 xylanase: pH titration of B. agaradhaerens xylanase. *Biochemistry* **43**, 5820-31 (2004).

249. Joshi, M.D., Hedberg, A. & McIntosh, L.P. Complete measurement of the pKa values of the carboxyl and imidazole groups in Bacillus circulans xylanase. *Protein Science* **6**, 2667-70 (1997).

250. Stanton, C.L. & Houk, K.N. Benchmarking pKa Prediction Methods for Residues in Proteins. *Journal of Chemical Theory and Computation* **4**, 951-66 (2008).

251. Simonson, T., Carlsson, J. & Case, D.A. Proton binding to proteins: pK(a) calculations with explicit and implicit solvent models. *Journal of American Chemical Society***126**, 4167-80 (2004).

252. Pokala, N. & Handel, T.M. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. *Protein Science* **13**, 925-36 (2004).

253. Georgescu, R.E., Alexov, E.G. & Gunner, M.R. Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. *Biophysical Journal* **83**, 1731-48 (2002).

254. Song, Y., Mao, J. & Gunner, M.R. MCCE2: improving protein pKa calculations with extensive side chain rotamer sampling. *Journal of Computational Chemistry* **30**, 2231-47 (2009).