Microscopic Insights To Relaxation Phenomena In Proteins

Thesis submitted for the degree of

Doctor of Philosophy (Science)

in

Physics (Theoretical)

by

Abhik Ghosh Moulick

Department of Physics

University of Calcutta

2023

Dedicated to

Covid warriors who saved the society during pandemic

"You can grinding four years with no results and the fifth year become the biggest thing on planet. The power of not giving up is real" - Anonymous

"You can't always get what you want, but if you try, sometimes, you might find, you get what you need" - The Rolling Stones

"If every thing is under control you are going too slow" - Mario Andretti

Acknowledgement

After completing 5 years of my Ph.D. life at S.N. Bose National Centre for Basic Sciences, I am thrilled to present my thesis, which is the culmination of those 5 years of hard work. Throughout my journey, I was accompanied by several individuals whose support and assistance were indispensable in achieving this feat. I am deeply grateful for this opportunity to express my gratitude towards them.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Prof. Jaydeb Chakrabarti (JC), who introduced me to the fascinating world of molecular simulations and guided me throughout my Ph.D. His mentoring enabled me to shape the project independently and think critically. Without his invaluable support, it would have been difficult for me to complete my Ph.D. His unwavering support during the challenging times of COVID-19 in 2021 helped me immensely.

I am also grateful to my thesis committee members, Dr. Sakuntala Chatterjee and Dr. Suman Chakrabarty, for their valuable comments and suggestions. I would like to give special thanks to Dr. Suman Chakrabarty for his numerical methods classes during our Ph.D. coursework, which helped me understand the beauty of computational methods. Furthermore, I would like to express my gratitude to the Technical Research Centre for the CRAY supercomputing facility, which provided me with the necessary computing power to carry out my research. Special thanks to DST, Govt. Of India for providing me the INSPIRE fellowship for Ph.D. I would like to thank CSIR for offering me travel grant during my visit at EPFL, Switzerland for the conference organised by CECAM.

I would like to express my gratitude towards Prof. Anjan Dasgupta, who served as the external examiner during my SRF viva. He introduced me to research in biomolecules during my summer project in B.Sc. in his lab at the Department of Biochemistry, University of Calcutta.

It gives me immense pleasure to have such a wonderful family-like group that has been with me through every phase of my Ph.D. journey. I am deeply grateful to my senior, Dr. Sutapa Dutta, for providing guidance and suggestions during the initial days of my Ph.D. career. I also extend my sincere thanks to Dr. Piya Patra and Dr. Sasthi Charan Mandal for their invaluable guidance as group seniors throughout my Ph.D. tenure. My heartfelt gratitude goes out to my dear batchmate, Rahul, who accompanied me on the seven-year journey (M.Sc+Ph.D.) that began during our M.Sc days. Rahul has been a great friend and a constant support system, especially during the tough times of COVID when I was isolated in Basundhara hostel. I would also like to specially acknowledge my beloved junior, Anirban, who patiently listened to my Ph.D.-related

frustrations and provided unwavering support through scientific discussions and fun-filled memories throughout my Ph.D. tenure. Anirban's smiling face and motivation helped me a lot to overcome Ph.D.-related frustrations. Anirban, I really miss your company. Last but not least, I express my gratitude to other group members Edwin, Aayattidi, Suravi, Kanikadi, Avik, Anusree, and Sabuj for their wonderful company. I would like to specially acknowledge Krishnendu for his helpful discussion.

Now, I would like to mention those people without whom I can not imagine these five years of my SNB life. I want to give thanks to Siddhartha, one of my batchmate for his company all over my Ph.D. life. I would like to thank all my friends, seniors, and juniors of SNB Social, SNB Mess, Cultural group, and Muktangan. Many thanks to Dulal da, Utpal da, and all the staff of SNB Students Mess, Bhagirathi Canteen for serving nutritious and delicious food that greatly alleviated my fatigue and stress throughout the day.

I would like to recognize my friends from my past, whose influence has had a significant impact on both my academic and personal growth. Kartik, Souradeep, Koustav, and Biplab are invaluable treasures in my life, and I cherish their friendship dearly.

During my B.Sc. and M.Sc. days, I was fortunate to have several brilliant faculty members who inspired me to pursue research in physics. Dr. Ashim K. Mukherjee (AKM) introduced me to the beauty of statistical physics for the first time. I am grateful for his guidance and inspiration. I would also like to thank Dr. Shamik Gupta, Dr. Abhijit Bandyopadhyay, and Dr. Bobby Ezhuthachan of RKMVERI for their exceptional guidance during my master's program. Their expertise and support helped me to develop a deeper understanding of the subject. Furthermore, I would like to acknowledge my alma maters, Asansol Ramakrishna Mission and Purulia Ramakrishna Mission Vidyapith, for providing me with a strong foundation in education and values. I would like to express my gratitude to Swapanda for his guidance during my school days at Purulia.

I am incredibly grateful to the medical professionals at Heart Clinic, including Dr. Biplab Chandra and Dr. Supratim Nandy, who provided exceptional care and support during my battle with COVID-19 in 2021. Their tireless efforts and dedication helped me overcome the virus and regain my health. I would also like to thank my seniors, friends, and juniors from RKMVP Purulia, who stood beside me during the tough times of COVID-19. Their prayers and support were invaluable and helped me overcome the challenges posed by the disease. Without their support, it would have been impossible for me to recover.

I am blessed to have a supportive family who plays a pivotal role in my life. I am

indebted to Bordi (Dr. Ranjita Ghosh Moulick), my elder sister, and Jaydeep Dada (Dr. Jaydeep Bhattacharya), my brother-in-law, for instilling the motivation to pursue a Ph.D. during my school days. Additionally, I am deeply grateful to Mami and Jethu for their unconditional support throughout my life. I would also like to express my gratitude to Mejdi and Poka for their unwavering love and affection. My heartfelt appreciation, respect, and love go to Maa and Baba. I owe them not only for this thesis but also for everything good in my life. Their constant motivation and unwavering support have been a driving force behind my success. Finally, I would like to thank Indrani, my best friend, philosopher, and guide, and also my life partner, for being the most precious gift during my Ph.D. journey. Thank you for your unending support, encouragement, and love that helped me navigate through the tough times.

In conclusion, I would like to express my appreciation for the open source and free software that facilitated the completion of my Ph.D. work. I am deeply grateful to all those who supported me in any way during my Ph.D. journey, without whom the completion of this thesis would not have been possible. I regret that I cannot personally acknowledge each and every individual who helped me. I extend my sincere thanks to everyone for their contributions.

Abhik Ghosh Moulick
Department of Physics of Complex Systems,
S. N. Bose National Centre for Basic Sciences,
Salt lake, Kolkata-700106, India
2023.

Abstract

Microscopic understanding of relaxation in biomolecular systems like protein is much more complicated than a typical condensed matter system due to involvement a large number of degrees of freedom over a wide range of timescales (femto-seconds to seconds). Structural changes can be induced in a protein by various agents like approach of ligands, altering solvent conditions, and thermodynamic properties. In this thesis we aim to describe microscopic aspects of structural relaxation in protein through change in conformational fluctuations.

Dihedral angles are microscopic degrees of freedom to describe protein conformations. We address static and dynamic aspects of conformational fluctuations in terms of dihedral angle using the molecular dynamics trajectory. The relaxation of dihedral angle is not probed via the existing experimental tools. The backbone dihedral angle have been estimated from NMR data. This leads us to examine if protein dihedral fluctuations can be associated to the NMR cross correlated dipolar fluctuations. We shows that the zero frequency spectral density function of dipole and dihedral fluctuations, using all atom molecular dynamics trajectory, are well correlated. Thus, structural relaxation of protein in terms of dihedral angle fluctuations can be probed using CCR rates of NMR.

We illustrate conformational fluctuations of protein in molten globule (MG) state. The MG state of a protein is dynamic in nature, where conformational states are inter converted on nanosecond time scales. Here we compare the conformational fluctuations of the MG state to those of intrinsic disordered proteins (IDPs). We consider protein, α-lactalbumin (aLA), which shows an MG state at low pH upon removal of the calcium (Ca^{2+}) ion. We find that the MG state of a protein behaves as an intrinsic disorder protein, although the disorder here is induced by external conditions. We further explore functionality of protein at MG state. In MG state, aLA acts as a carrier of fatty acid like oleic acid (OLA) that shows cytotoxic activity against cancer cell line. We characterize binding of aLA with OLA microscopically. We find that the metastable conformational fluctuations shift from ligand binding site to the Ca^{2+} ion binding site as OLA

is removed from the protein.

We also build a coarse-grained model of protein with structural information using the effective free energy profile obtained from all-atom molecular dynamics. We show that using such a model, one can reproduce the protein structure comparable to the crystal structure and all atom simulation.

সারাংশ

প্রোটিনের ন্যায় জৈব অনুর ক্ষেত্রে , তাদের গতি প্রকৃতি বোঝার অন্যতম উপায় হলো প্রোটিনের মধ্যে থাকা বিভিন্ন অ্যামিনো এসিডের শিথিলকরণ | সাধারণত প্রোটিনের গঠনগত পরিবর্তনে মধ্যে দিয়ে অ্যামিনো এসিডের শিথিলকরন ঘটে থাকে | এই থিসিসে আমরা প্রোটিনের গঠনগত পরিবর্তনের বিভিন্ন দিকগুলি বিশেষত সময় সাপেক্ষ দিকগুলির দিকে নজরপাত করেছি | আমরা প্রোটিনের বিভিন্ন অ্যামিনো এসিডের ডিহেড্রাল কোন কে প্রোটিনের প্রধান ডিগ্রী অফ ফ্রিডম হিসেবে ব্যবহার করেছি | শিথিলকরণ এর ওপর ভিত্তি করে মূলত চারটি বিষয়ে আমরা আলোকপাত করেছি : (ক) প্রোটিনের শিথিলকরন কে কোনো ভাবে কোনোরকম পরীক্ষামূলক বিষয়ের সাথে সম্বন্ধ স্থাপন করা যায় কিনা , (খ) আলফা ল্যাক্টালবুমিন প্রোটিনের মোল্টেন গ্লোবুল অবস্থায় প্রোটিনের গঠনমূলক পরিবর্তন, (গ) মোল্টেন গ্লোবুল অবস্থায় আলফা ল্যাক্টালবুমিন প্রোটিনের কোনো ফ্যাটি এসিডের সাথে বন্ধনের প্রবণতা, (ঘ) মডেল গণনার মাধ্যমে প্রোটিনের কোর্স গ্র্যান বর্ণনা |

ডিহেড্রাল কোন এর পরিপ্রেক্ষিতে এখনো প্রোটিনের শিথিলতা কোনো ধরনের পরীক্ষামূলক পরিমাপের মাধ্যমে অনুসন্ধান করা সম্ভব হয়ে ওঠে নি | যদিও, প্রোটিনের প্রধান চেন এর ডিহেড্রাল কোণকে নিউক্লেয়ার ম্যাগনেটিক রেসোনেন্স এর তথ্য থেকে অনুমান করা যায় । আমরা এখানে দেখতে চেয়েছি প্রোটিনের ডিহেড্রাল কোন এর সময় সাপেক্ষ পরিবর্তন কে নিউক্লেয়ার ম্যাগনেটিক রেসোনেন্স এর দ্বারা পরিমাপযোগ্য প্রোটিনের ডাইপোলের ক্রস কোরেলেটেড রিলাক্সেশন হারের সাথে সম্পর্ক যুক্ত করা যায় কিনা | আমরা দেখিয়েছি যে ডাইপোল এবং ডাইহেড্রাল এর সময়সাপেক্ষ পরিবর্তন মলিকুলার ডাইনামিক্স এর গতিপথ ব্যবহার করে পাওয়া শূন্য ফ্রিকোয়েন্সির স্পেক্ট্রাল ফাঙ্কশন মাধ্যমে ভালভাবে সম্পর্কযুক্ত। এইভাবে, ডিহেড্রাল কোণ এর সময় সাপেক্ষ পরিবর্তন এর মাদ্ধমে প্রোটিনের শিথিলতা, নিউক্লেয়ার ম্যাগনেটিক রেসোনেন্স এর থেকে পাওয়া ক্রস কোরেলেটেড রিলাক্সেশন হারের সাহায্যে পরিমাপ করা যেতে পারে।

আমরা প্রোটিনের মোল্টেন গ্লোবুল অবস্থায় প্রোটিনের গঠনগত পরিবর্তন নিয়ে আলোচনা করেছি | প্রোটিন মোল্টেন গ্লোবুল অবস্থায় অত্যন্ত গতিশীল, যেখানে গঠনমূলক অবস্থা ন্যানোসেকেন্ড টাইম স্কেলে আন্তঃরূপান্তরিত হয়। এখানে আমরা প্রোটিনের মোল্টেন গ্লোবুল অবস্থায় তার গঠন মূলক অবস্থার পরিবর্তনকে ইন্ট্রিন্সিক ডিস্ অর্ডার প্রোটিনের সাথে তুলনা করেছি | আমরা প্রোটিন আলফা-ল্যাকটালবুমিনকে বিবেচনা করেছি, যা ক্যালসিয়াম আয়ন অপসারণের ফলে কম পি এইচ -এ মোল্টেন গ্লোবুল অবস্থায় পরিবর্তিত হয়। আমরা দেখতে পেয়েছি যে প্রোটিন তার মোল্টেন গ্লোবুল অবস্থা ইন্ট্রিন্সিক ডিস্ অর্ডার প্রোটিনের ন্যায় আচরণ করে, যদিও এখানে ডিস্ অর্ডার প্রকৃতি বাহ্যিক অবস্থার দ্বারা প্ররোচিত হয়।

আমরা প্রোটিনের মোল্টেন গ্লোবুল অবস্থায় প্রোটিনের কার্যকারিতার ওপর আলোকপাত করেছি | মোল্টেন গ্লোবুল অবস্থায় আলফা ল্যাক্টালবুমিন প্রোটিন, ওলিক এসিডের ন্যায় ফ্যাটি এসিডের বাহক হিসাবে কাজ করে যা ক্যান্সার কোষ লাইনের বিরুদ্ধে সাইটোটক্সিক কার্যকলাপ দেখায়। আমরা ওলিক এসিডের সাথে প্রোটিনের বন্ধন প্রবণতা নিয়ে আলোচনা করেছি | আমরা পেয়েছি যে প্রোটিনের সাথে ওলিক এসিডের বন্ধনের জন্য যে পরিমান শক্তি দরকার তা আগের পরীক্ষামূলক ফলাফলের সাথে তুলনীয়। আমাদের কাজ সম্ভাব্য ড্রাগ অ্যাপ্লিকেশনের জন্য সহায়ক হতে পারে |

আমরা প্রোটিনের মডেল গণনার মাধ্যমে প্রোটিনের কোর্স গ্র্যান বর্ণনা দিতে সক্ষম হয়েছি | আমাদের মডেল গণনার মাদ্ধমে আমরা ডিহেড্রাল কোণ এর মাধ্যমে প্রোটিনের কাঠামোগত বর্ণনা দিতে পেরেছি | মলিকুলার ডাইনামিক্স এর গতিপথ থেকে প্রাপ্ত কার্যকর মুক্ত শক্তি প্রোফাইল ব্যবহার করে মূলত প্রোটিনের মডেলটি তৈরী করেছি | মডেল তৈরির ক্ষেত্রে আমরা মন্টি কার্লো সিমুলেশন পদ্ধতির ব্যবহার করেছি | আমরা দেখিয়েছি যে এই ধরনের একটি মডেল ব্যবহার করে পাওয়া প্রত্যেক অ্যামিনো এসিডের ডিহেড্রাল কোণ, পরীক্ষামূলক ক্রিস্টাল কাঠামো থেকে পাওয়া প্রত্যেক অ্যামিনো এসিডের ডিহেড্রাল কোনের সাথে তুলনীয় | এই মডেলটি কোষের মধ্যে অনেকগুলি প্রোটিন সম্পর্কিত কোনো বায়োফিজিক্যাল বা বায়োকেমিকাল পদ্ধতির মডেল অধ্যয়ন এর ক্ষেত্রে উপযোগী হতে পারে |

List Of Publications

1. **Abhik Ghosh Moulick**, J. Chakrabarti, Conformational fluctuations in molten globule state of α-lactalbumin, Physical Chemistry Chemical Physics, 2022, 24, 21348 - 21357.

2. **Abhik Ghosh Moulick**, J. Chakrabarti, Correlated dipolar and dihedral fluctuations in a protein, Chemical Physics Letters 797 (2022) 139574.

3. **Abhik Ghosh Moulick**, J. Chakrabarti, Correlation between protein bond vector and dihedral fluctuations, AIP Conference Proceedings 2265, 030036 (2020).

4. **Abhik Ghosh Moulick** & J. Chakrabarti, Microscopic understanding of fatty acid binding with α-lactalbumin at molten globule state (To be submitted).

5. **Abhik Ghosh Moulick** , Anirban Paul, J. Chakrabarti Coarse-grained model of protein with structural information (Manuscript under preparation).

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Relaxation phenomena describe time dependent changes of physical quantities following external perturbations. They are good probes for information on the local environment of the system. Microscopic understanding of relaxation phenomena in biomolecular systems like protein is much more complicated than a typical condensed matter system due to involvement of a wide range of dynamical timescales (femto-seconds to seconds) and a large number of degrees of freedom.[1,2] Microscopic understanding of protein function in terms of structural relaxation presents major challenge due to this extended range of timescales as well as various interactions.

Structural relaxation of protein molecules can be induced by various agents like approach of ligands, altering solvent conditions, and thermodynamic properties. The structural relaxation in protein typically includes conformational fluctuations. Microscopic descriptions of conformational fluctuations involve relaxation of structural degree of freedom like the dihedral angles of the protein.

Experimental techniques like X-ray and NMR are able to probe information on protein conformations. But they have their own limitations. X-Ray based information is possible only if a protein can be crystallized. Fig.1.1(a) shows three dimensional structure of protein Ubiquitin (PDB ID:1UBQ) obtained from X-Ray crystallography. Helix and Sheet are marked in red and yellow color respectively. The data obtained from X-Ray is accompanied by B-factor, which gives deviation of atomic coordinate about their mean position due to thermal fluctuations. B-factor may contain model error, invalid restrains, including lattice imperfections. Alternatively, NMR provides information about protein relaxation, but is limited only to the millisecond time range.[3,4] NMR signal is also limited

to smaller protein and NMR spectra are not well-defined, in particular for large proteins and proteins which lack in secondary structure. All atom simulations, supplementary to the experiments, can provide microscopic information on proteins ranging from a pico-second (ps) to a microsecond (μs) time window, with both dynamic and static properties.[5]

Classical all atom molecular dynamics can simulate physical motion of the atom by solving Newton's equation of motion, which is comparable to experimental conditions. Protein dihedral angle can be used as conformational degrees of freedom to understand protein functionality efficiently.[6–8] Dihedral angle is the angle between two planes. Protein backbone chain consists of three dihedral (ϕ, ψ, ω) angle and side chain consists of five dihedral ($\chi_i, i = 1, .., 5$). Fig.1.1(b) shows a schematic representation of protein backbone dihedral angle ϕ and ψ. ϕ_i is defined as angle between $C_{i-1} - N_i - C_{\alpha,i}$ and $N_i - C_{\alpha,i} - C_i$ planes and ψ_i is defined as angle between $N_i - C_{\alpha,i} - C_i$ and $C_{\alpha,i} - C_i - N_{i+1}$, where i is residue index. χ_1, side chain dihedral is defined as the angle between $N_i - C_{\alpha,i} - C_{\beta,i}$ and $C_{\alpha,i} - C_{\beta,i} - C_{\gamma,i}$.

In this thesis, we address static and dynamic aspects of conformational fluctuations in terms of dihedral angle to understand structural relaxation microscopically.[9] First, we relate the dihedral fluctuations to experimentally measurable quantities. The structural relaxation of protein in terms of dihedral angle is not probed via any kind of experimental measurement. Although, the backbone dihedral angle have been estimated from NMR data. Here, we consider if protein dihedral fluctuations can be associated to the NMR cross correlated dipolar fluctuations. We show that the zero frequency spectral density function of dipole and dihedral fluctuations, using all atom molecular dynamics trajectory, are well correlated. Thus, structural relaxation of protein in terms of dihedral angle fluctuations can be probed using CCR rates of NMR. We illustrate conformational fluctuations of protein at molten globule (MG) state induced by lowering the solvent pH. The MG state of a protein is dynamic in nature, where conformational states are inter converted on nanosecond time scales. Here we compare the conformational fluctuations of the MG state to those of intrinsic disordered proteins (IDPs). We consider protein, α-lactalbumin (aLA), which shows an MG state at low pH upon removal of the calcium (Ca^{2+}) ion. It is observed that the long live fluctuations are localised to the Ca^{2+} binding site. We find that the MG state of a protein behaves as an intrinsic disorder protein, although the disorder here is induced by external conditions. We also explore functionality of protein at MG state. In MG state, aLA acts as a carrier of fatty acid like oleic acid (OLA)

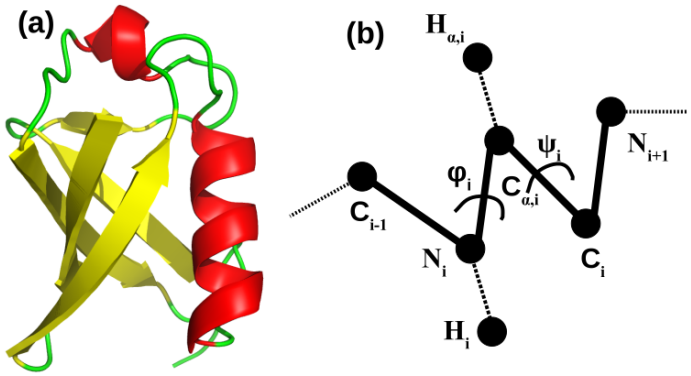

Figure 1.1: (a) Three dimensional structure of protein along with its secondary structure component. Helix is marked in red color and sheet is in yellow color, (b) Schematic representation of protein dihedral angle ϕ and ψ along with NMR sensitive dipole $H_i - N_i/C_{\alpha,i} - H_{\alpha,i}$

that shows cytotoxic activity against cancer cell line. We characterize the dihedral fluctuations as the ligand is liberated from the active site of protein. The long live fluctuations occur near the ligand binding site which eventually transferred towards Ca^{2+} binding site as ligand is taken away from protein. We also build a coarse-grained model of protein with structural information using the effective free energy profile obtained from all-atom molecular dynamics. We show that using such a model, one can reproduce the protein structure comparable to the crystal structure and all atom simulation. This study may be useful to study phenomena involving many protein molecules, like protein aggregations.

The rest of the chapter is organised as follows: In section 1.1, we describe how to relate structural relaxation of protein with some experimental measurable quantities. In section 1.2, we look at structural relaxation of protein out of its native structure where we discuss conformational fluctuations of protein at MG state. Section 1.3 illustrates our investigations on ligand binding ability of protein at MG state. Section 1.4 depicts our finding on coarse-grained representation of protein including structural information.

1.1 Correlated dihedral and dipolar fluctuations in protein

Coordination between functional parts of protein controls ligand binding, essential for various biophysical and biochemical phenomena inside the cell. Sometimes proteins become functional following ligand binding at a far site, known

as allostery. Recent study[10,11] suggest that dihedral angles are convenient microscopic variables to describe the equilibrium aspects, including thermodynamics of conformational changes upon ligand binding[7]and allostery in protein.[6,12] Although dihedral fluctuations play a key role in protein function, they are not directly amenable to experiment due to the limitations of the experimental probe.

The NMR experiments probe dipolar fluctuations in terms of cross-correlated relaxation (CCR) rates, given by the zero frequency spectral density function of the fluctuations.[13,14] Fig.1.1(b) shows that backbone dihedrals ϕ_i and ψ_i involve the NMR sensitive dipole pairs like $H_i - N_i/C_{\alpha,i} - H_{\alpha,i}$ where i is residue index. One would, therefore, expect the backbone dihedral to be sensitive to the fluctuations of mutual orientation of these dipoles.

The major problem in relating the dihedral fluctuations to those of dipoles is the issue of timescale. While the dihedral fluctuations take place in nanoseconds, the CCR rates provide integrated information over time on dipole orientation correlation function. If the dipolar fluctuations at timescales of nanoseconds have an important contribution to CCR, CCR may act as markers for backbone dihedral fluctuations as well.

We use molecular dynamics (MD) computer simulations to compute different time dependent correlation functions (TDCF) of the mutual dipolar orientations for both intra-residue and sequential dipole vectors and the backbone dihedral angles over the equilibrated portion of molecular dynamics trajectory for proteins, GB3 and Ubiquitin(Ub). The TDCFs of fluctuations, for dipolar orientation TDCFs, are given by the second order Legendre polynomial of angle between two dipole vectors. We calculate time dependent autocorrelation functions for dihedral fluctuations as well.

We observe that the dipolar and dihedral fluctuations decay typically within few nanosecond (ns) time scale. Zero frequency spectral function integrated over nanoseconds of dipolar fluctuations can capture the experimental CCR suggesting that the short time (ns) decay of the dipolar fluctuations well describe the experimental data. Next, we check to what extent the intra-residual dipolar fluctuations are correlated to the dihedral fluctuations. We find that the zero-frequency intra-residual dipolar fluctuations are well correlated to those of the dihedral fluctuations. We also examine correlation plot for sequential dipolar fluctuations and the dihedral fluctuations. We find that ψ auto-correlations in particular are better correlated to the sequential orientation fluctuations.

Thus, we conclude that the fluctuations at the timescale of a few tens of nanoseconds can capture the experimental CCR of dipolar fluctuations. Within

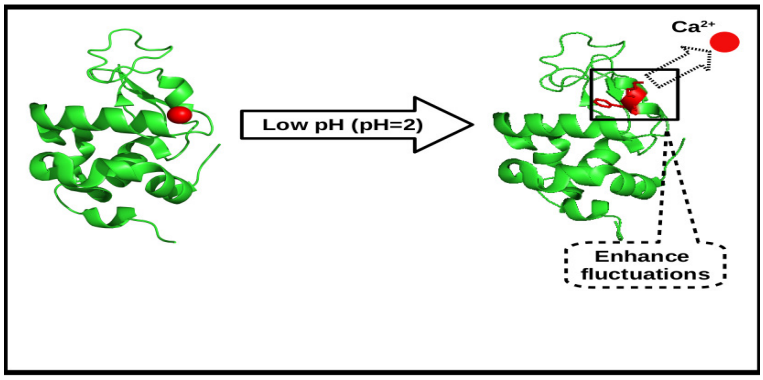

Figure 1.2: At low pH, enhance fluctuations occur in Ca^{2+} binding region of α-lactalbumin and Ca^{2+} ion comes out. The protein goes into a molten globule state.

this timescale, the zero frequency mode of dihedral fluctuations and dipolar fluctuations are correlated well. Hence, CCR obtained through NMR can be a marker of dihedral fluctuations.

1.2 Conformational fluctuations in molten globule state

Many proteins show structural fluctuations in a near denaturing-condition, while retaining their overall tertiary structures.[15] Such states are called Molten Globule (MG) state of the protein. The MG state is induced by various denaturing conditions like high temperature, pH, high pressure and due to the presence of various denaturing chemicals like urea.[16] The structural and functional characterization of the MG is largely lacking, since the MG states are not directly amenable to crystallization. Earlier experimental and theoretical studies show that the MG state of the protein is dynamic in nature, where conformational states are inter converted on nanosecond time scales.[17] These observations lead us to study and compare conformational fluctuations of the MG state induced by the external agents to those of intrinsically disordered proteins (IDP).[18]

Here, we consider a milk protein α-lactalbumin (aLA). Fig.1.2 shows protein structure at its holo state (with Ca^{2+} ion). The protein converts to MG state at low pH upon removal of calcium (Ca^{2+}) ion.[19] We compare the microscopic fluctuations in the state of aLA to those of IDP. Since maintaining low pH is essential for the formation of the MG state of aLA, we perform biased constant pH molecular dynamics (CpHMD) via a hybrid scheme where explicit water molecules are taken into account in the MD simulations. We use the dihedral

principal component analysis, the density based clustering method, and the machine learning technique to identify the conformational fluctuations[8,20–25] in the MG state.

We find the presence of metastable states spanned by the dihedral angles in the MG state of aLA . We find that residues belonging to a stable secondary structure in the crystal structure are responsible for the fluctuations, suggesting enhanced conformational fluctuations.

Thus, our results suggest that protein in MG state is similar to IDP. However, it is important to keep in mind that the MG state is induced by external effects like lowering of the solution pH and, hence, can be viewed as an induced disordered protein. Our results suggest the following scenario of the MG state. At low pH, there are enhanced fluctuations near the Ca^{2+} binding region and eventually Ca^{2+} ion will come out from protein. As a result, protein converts into molten globule state(Fig.1.2). Our study will be helpful to understand the functionality of a protein in partly denatured conditions, as in the MG state.

1.3 Fatty acid binding with α-lactalbumin in molten globule (MG) state

In the previous section we find fluctuations in the MG state of aLA (MG-aLA). MG-aLA is known to bind with ligand[26] like oleic acid (OLA). The complex formed between MG state of aLA (MG-aLA) and OLA, known as XAMLET (aLA made lethal where X stands for the name of the mammal) draws considerable attention due to its potential application of cytotoxic activity against cancer cells.[19] It is envisaged that OLA has cytotoxic activities where MG-aLA acts like a carrier of OLA. Earlier experimental observations suggest that OLA binds with aLA at MG state near the cleft region of protein, with binding energy approximately -9.45 kcal/mol.[27] The microscopic understanding of binding is not yet established due to lack of crystal structure. Here, we explore OLA binding to MG state of aLA through microscopic simulations.

Since the binding modes are not known experimentally, so first we find the active residues of MG-aLA for OLA binding. Recent experiments and theoretical works suggest that thermodynamics based on the dihedral fluctuations of a protein in different conformations play a vital role in ligand binding.[7] The thermodynamic free energy (ΔG) and entropy ($T\Delta S$) costs are estimated using histogram base method(HBM) from the distributions of dihedral angle in the

respective conformations. Residues having, $\Delta G, T\Delta S > 0.0$ are identified as destabilized and disordered residues. These residues are active in ligand binding.[28] This yields the binding mode between the ligand and the protein. We apply similar approach to identify active residues in MG-aLA for OLA binding. We calculate the changes in conformational free energy(ΔG) and entropy($T\Delta S$) of dihedral angle at MG state and neutral state. Fig.1.3 show the active residues in green color over the energy minimized structure of MG-aLA-OLA complex.

We perform MD simulations on the complex and observe that the ligand remain in the vicinity of the active sites. We estimate the binding free energy using the umbrella sampling method.[29] Here, we generate a series of configurations by keeping the centre of mass of the protein and the ligand at fixed distances. We estimate the binding free energy $\sim$ 8.3 kcal/mol which compares well with the experimental results.

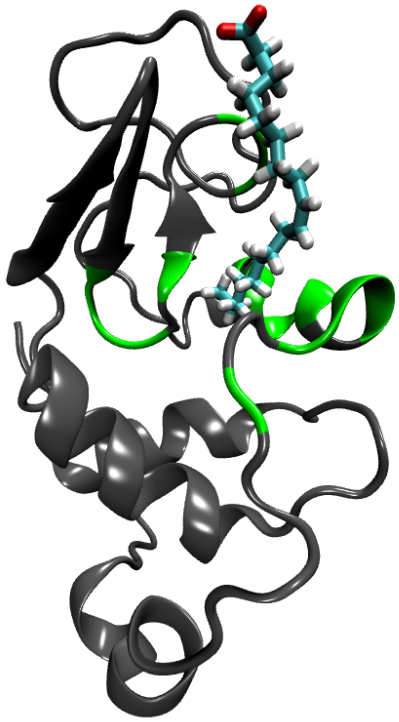

Figure 1.3: Cartoon representation of aLA at MG state along with the OLA. Hydrophobic tail of OLA binds in the cleft region. Binding residues are marked in green color over the energy minimized structure of the complex.

We find that upon complex formation metastability decrease. We perform clustering and machine learning base analysis to identify essential coordinates (EC) of the system at complex state. We find that the binding site residues play role as EC in the conformation fluctuations. We further perform the machine learning based analysis on the steered MD trajectories where the protein and ligand are hold at a given distance. As the ligand goes further away, the ECs are shifted towards the residues of Ca^{2+} binding region, suggesting that the Ca^{2+} binding residues have allosteric control on the ligand binding, confirming the suggestions in earlier work.[30,31]

We also check the relaxation of water molecules near the protein surface in

terms of various dynamical quantities like solvent residence time, translational and rotational diffusion. We observe that water molecules near the protein and OLA surface exhibit sub diffusive behaviour. We find that the water molecules close to both protein and ligand have slower timescale and lower diffusion constant compared to the bulk. This suggests heterogeneous behaviour of water at the protein ligand interface in agreement to other systems reported in the literature.[32]

Thus, we have related the conformational change of protein aLA at MG state due to binding with fatty acid like OLA. Our analysis reveals that protein MG state become stabilized upon OLA binding and the binding energy is comparable with earlier experimental observations. The fluctuations at the binding site may be helpful to release the ligand.

1.4 Coarse grained model of protein

Many cellular phenomena involve a large number of bio-molecules ranging from water, small and medium-size oligomers and co-polymers (peptides, proteins, RNA, etc) to huge co-polymers, such as DNA. For instance, protein aggregation is a complex process for which computational study is still difficult due to the involvement of broad range of lengths and time scale.[33] Classical all-atom molecular modelling is still limited by its algorithmic efficiency and the available computing power.[34] Lowering the representation from all atom to coarse grained (CG) model shows possibility to study systems involving large numbers of bio-molecules where typically a group of chemical moieties are represented by a single entity.[35,36] It is known that protein functionality largely depends on its structure.[37] However, structural information has not been properly addressed in the CG models of proteins.[35,36]

We build a CG model of protein incorporating structural information. We represent an amino acid as a bead with spring given by the radius of gyration over the all-atom trajectory. The model interactions between beads are governed by bonding and stretching as bonded interactions and Van Der Waals type interaction as non-bonded interactions. Each bead is assigned with a couple of degrees of freedom representing the backbone dihedral angles. The energy cost of these degree of freedom is obtained from the negative logarithmic of the joint distribution of the fluctuations of the backbone dihedrals in fully atomic simulations. We consider solvent explicitly in the model where solvents are interacted with polymer beads via repulsive part of Lennard-Jones interaction.

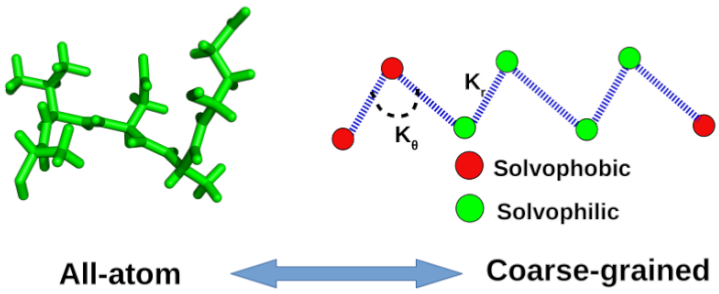

Figure 1.4: All atom to coarse-grained representation of protein amino acid residues. Red corresponds to hydrophobic and green corresponds to hydrophilic amino acid. Beads are connected via harmonic spring

We use the Monte Carlo (MC) method to generate conformations considering model solvent interactions. The initial values of the dihedral angle of each amino acid residues are taken from protein crystal structure data. When we construct MC move, we compute the energy cost due to different interactions like bonded, non-bonded and solvent contributions. We also change the dihedral degree of freedom and construct the energy cost by using potential generated from all-atom simulation.

We show that using such model one can reproduce well the protein structure comparable to the crystal structure and all atom data. We also used this dihedral coupling information for other proteins to check transferability of the model. We find that one can get well structural information for other protein like lambda N protein and ubiquitin protein based on dihedral coupling information of protein GB3.

Thus, we propose a simple coarse-grained description of protein models where dihedral information is included. This can pave the way for CG model with structural information.

The organization of the rest of the thesis is as follows: In chapter 2 we discuss how structural relaxation of protein in terms of dihedral angle could be related to some experimental measurable quantity using time dependent correlation function analysis. Chapter 3 consider conformational fluctuations of protein like aLA at molten globule state (MG). We extend this work in Chapter 4 to study the OLA binding of aLA in MG state. In Chapter 5, we discuss on CG model of protein along with structural information in terms of dihedral angle.

Chapter 2 Correlated dihedral and dipolar fluctuations in a protein

2.1 Introduction

Ligand binding sites in a protein are often separated by distances much larger than atomic sizes.[12,38] However, their dihedral angles fluctuate in a correlated manner, which is revealed in a number of studies by the Pearson correlation coefficient.[11,39] The Pearson correlation coefficient, given by covariance of two variables, is a purely static quantity. In the case of functional residues, the correlations have intrinsic timescales that control the rate of binding events. It has been shown that one can generalize the covariance between two variables with a time lag to construct a time-dependent correlation function (TDCF)[40] of fluctuations of dihedral angles.[41] The time-dependent correlations of dihedral fluctuations persist till nanosecond (ns) even for spatially distant functional residues.[41] Such time scales are suitable for ligand binding. Although the dihedral fluctuations can easily be constructed theoretically, the nanosecond timescales of the correlated fluctuations down to the spatial resolution of the residues are not yet directly amenable to experimental probes. It is interesting to ask if the dihedral fluctuations can be related to an experimentally measurable quantity. Nuclear magnetic resonance (NMR) techniques[42] are used to probe atomic motions through dipolar fluctuations in a protein.[13,14,43–59] The collective

Based on publications: 1. Abhik Ghosh Moulick, J. Chakrabarti, Correlated dipolar and dihedral fluctuations in a protein, Chemical Physics Letters 797 (2022) 139574.; 2.Abhik Ghosh Moulick, J. Chakrabarti, Correlation between protein bond vector and dihedral fluctuations, AIP Conference Proceedings 2265, 030036 (2020)

dynamics of a group of atoms[60–62] is given in terms of cross correlated relaxation (CCR)[63] rates, usually expressed as the zero frequency spectral density function J(0) of the fluctuations of the mutual orientation of two spatially separated dipole vectors in a protein.[64,65] Experiments often consider fluctuations in mutual orientation between $H^N - N$ and $C_\alpha - H_\alpha$ dipole pairs. Such data have been reported for GB3.[66] The effects of dynamics of protein on CCR rates have been probed in a number of studies.[67–70]

The angle between $C'_{i-1} - N_i - C^\alpha_i$ and $N_i - C^\alpha_i - C'_i$ planes, defines the backbone dihedral ϕ. Similarly, the angle between $N_i - C^\alpha_i - C'_i$ and $C^\alpha_i - C'_i - N_{i+1}$ planes defines the backbone dihedral ψ. Thus, the backbone dihedrals involve the dipole pairs $H^N_i - N_i/C^\alpha_i - H^\alpha_i$ (Figure2.1). Earlier work[46] shows that NMR cross correlation coefficients can be utilized to determine simultaneously the backbone ϕ and ψ angles. The method depends on measuring cross correlated dipolar relaxation, which in turn depends on the angle γ between dipole vectors $H^N - N$ and $C_\alpha - H_\alpha$. Another way of probing dihedral angle experimentally is J-coupling, where the relationship between 3 bond J-coupling (^3J) and intervening dihedral is used. Those experimental studies depend on either average angle obtained from a multi-state structure or summing up of CCR rates of each individual states. Such an approach, however, loses information on the correlated motion. It is not, therefore, obvious if the dihedral fluctuations can be captured from dipolar fluctuations. One would, however, expect the backbone dihedral to be sensitive to the fluctuations of mutual orientation of these dipoles, which is supported by preliminary data.[71] The major problem in relating the dihedral fluctuations to those of dipoles is the issue of timescale. While the dihedral fluctuations take place in nanoseconds, the CCR rates provide integrated information over time on dipole orientation correlation function. If the dipolar fluctuations at timescales of nanoseconds have an important contribution to CCR, CCR may act as markers for backbone dihedral fluctuations as well.

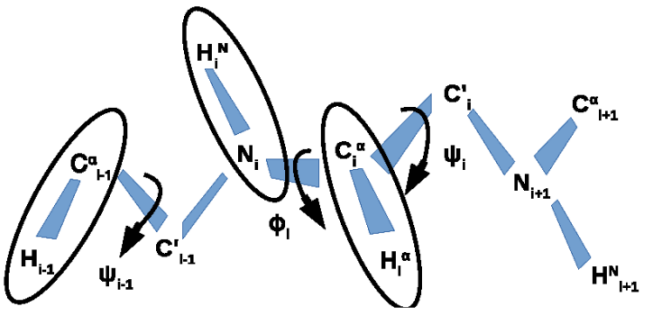

Figure 2.1: Schematic diagram of dipoles (within ellipses) used in the present study. The backbone dihedral angles ϕ and ψ are shown by arrows.

Motivated by this, we examine the zero frequency spectral functions of intra-residual backbone dihedral angles and $H^N - N/C_\alpha - H_\alpha$ dipolar orientation fluctuations in two proteins, GB3 and Ubiquitin (Ub). We use molecular dynamics (MD) computer simulation to compute different TDCFs of the mutual dipolar orientations for both intra-residue and sequential dipole vectors and the backbone dihedral angles over equilibrated portion of the trajectory. We extract the zero frequency spectral function for all the correlation functions by numerically integrating the correlation functions up to a time so that the entire decay of the initial correlations is captured.

We observe that the correlations of dipolar fluctuations of both the proteins persist up to a few nanoseconds, in agreement with earlier works.[72] We compare the zero frequency spectral functions for the both intra-residue and sequential dipole orientation fluctuations to the experimentally available CCR data for GB3. We find that the theoretical data for dipolar fluctuations capture well the experimental CCR data for both intra-residue and sequential dipole vectors. Thus, the experimental CCR is largely accounted for by short time decay of the correlation functions. Similarly, time dependent correlation fluctuations of the dihedral angle (ϕ,ψ) possess correlation up to tens of nanoseconds, which agrees to earlier study.[41] We extract the zero frequency spectral functions for these fluctuations also, considering the short time decay. We observe that the residues belonging to the helix region of both proteins show faster decay of dihedral auto-correlation functions compared to those belonging to the beta-sheet and loop regions. The dihedral ϕ of helix residues show moderate correlation with intra-residue dipolar fluctuation. On the other hand, the sequential dipolar fluctuations are strongly correlated with dihedral ψ.

2.2 Methods

2.2.1 System preparation & simulation details

We consider GB3 (PDB id: 2OED)[73] with 56 amino acids and Ubiquitin (PDB id : 1UBQ)[74] with 76 amino acids in our studies. Amber99sb force field (ff)[75] is used in the GROMACS software package[76] for simulation. The details of force field and simulation techniques are given in Appendix A1 and A2 respectively. The TIP3P[77] water model is considered for solvent molecules, and counter-ions are added for electroneutrality. Particle Mesh Ewald method is used to assess long ranged electrostatic energy. 10 Angstrom (Å) is considered as truncation

limit for both Lennard-Jones and short range interactions. Both proteins are solvated in cubic box. After energy minimization, the systems are equilibrated through NVT and NPT simulations using position restraints on heavy atoms at 300K Temperature and 1 Bar pressure, respectively. The production NPT runs are executed for 1.05 μs with 2 fs time step integration employing periodic boundary conditions in all directions.

2.2.2 Analysis

Time-dependent correlation function (TDCF) between two dynamic variables A and B for time interval Δt where $\Delta t = |t_2 - t_1|$ is defined as:

$$C_{A;B}(\Delta t) = \langle (A[t_2] - \overline{A}\rangle)(B[t_1] - \overline{B})\rangle \tag{2.1}$$

Here angular $\overline{A}$ and $\overline{B}$ denote ensemble average of the corresponding quantity over the entire simulation trajectory. The angular brackets denote the average over the choices of the initial time. One can choose without loss of generality, $t_1 = 0$ and $t_2 = \Delta t$. The details of computation are given in Ref.[41]

The time-dependent correlation function (TDCF) of fluctuations for dipolar orientation TDCFs, are given by the second order Legendre polynomial of angle between two dipole vectors. Let the orientations of two dipole vectors at two different times are given by $\theta(t), \phi(t)$ and $\theta(0), \phi(0)$ in a given reference frame. The addition theorem of the spherical harmonics yields $P_2(cos(\gamma)) = \sum_{m=-2}^{2} Y_{2,m}(\theta(t), \phi(t)) Y_{2,m}(\theta(0), \phi(0))$ where γ is the angle between the two dipole vectors. Using this, we find for intra-residue dipolar mutual orientation fluctuations, $C^i_{dipole}(\Delta t) = \langle P_2(cos(\gamma))\rangle$. Here $\langle ...\rangle$ denotes the average over initial time (0) with respect to which the time difference Δt is measured. Both the dipoles belong to the same (i-th) residue. Similarly, TDCFs $C^{i,i-1}_{dipole}(\Delta t)$ for sequential dipolar fluctuations are constructed from the second order Legendre polynomial of angle between dipole vectors belonging to the i-th residue and the $(i-1)$-th residue.

We calculate the dihedral angles for backbone ϕ and ψ from the atomic positions. Time-dependent auto correlation functions for dihedral fluctuation is defined using Eq. 2.1 where variables are replaced by sine of dihedral angles (ϕ and ψ of a given residue R).

The spectral density function $J(\omega)$ is defined as the Fourier transform of the correlation function $C(t)$.[78,79]

$$J_{A,B}(\omega) = 2\int_0^\infty C_{A,B}(t).cos(\omega t)dt \tag{2.2}$$

which corresponds to an integral over time for zero frequency. The integrals are computed numerically from the auto-correlation functions using the 10 point Gaussian quadrature. The correlation between variables are expressed in terms of Pearson correlation (r) coefficient. r-value is defined as the covariance of the two variables divided by the product of their standard deviations.[80]

2.3 Results & discussions

The equilibration of all-atom MD simulations of the system is ensured from the saturation of root-mean-square deviation (RMSD) of the backbone atoms, as shown in Figure2.2(a) for protein GB3 and (b) for Ub. We perform the analysis over the equilibrated part of the trajectory (beyond 250 ns). To validate the simulations, we compare the crystal structure of the proteins with the average structures over the equilibrated trajectory. The aligned conformations for both structures are in Figure2.2(c) for GB3 and Figure2.2(d) for Ub. For GB3, all heavy-atom RMSD is 1.21 Å and for Ub, RMSD is 1.80 Å. We also compare $\psi - \phi$ Ramachandran plot for both the crystal and average structures for both proteins Figure2.2(e)-(f). The Ramachandran plots show good structural similarities of both the proteins with the experimental crystal structures.

We divide the simulation trajectory after RMSD saturation into 8 equal windows, each consisting of configurations stored for 100ns. This time interval is sufficient to guarantee the decay of the TDCFs so that each window can be treated as independent as far as the dipolar fluctuations are concerned. We consider the TDCF of fluctuations of dipolar orientation using Eq.2.1 in each separate window. Finally, average over all windows is considered for final analysis.

2.3.1 Zero frequency spectral functions for dipolar orientation fluctuations

We consider the intra-residue dipole pairs $H^N - N$ and $C_\alpha - H_\alpha$ of residue i. Experimental CCR data cannot distinguish between the orientation fluctuations between the direct and the cross pairs, $H^N - N/C_\alpha - H_\alpha$ and, $H^N - C_\alpha/N - H_\alpha$ respectively. Hence, we consider both the pairs in our calculations. The TDCF is normalized by $\Delta t = 0$ value. Let us illustrate the case of Isoleucine (I7) of GB3 which belongs to the beta-sheet of the protein crystal structure. Fig.2.3(a) and (b) show the direct and cross dipole TDCFs for I7, $C^{I7}_{dipole}(\Delta t)$. The correlation functions decay within a few ns, along with oscillations of small amplitudes about zero at larger times. The insets show semi-log plots of the correlation function,

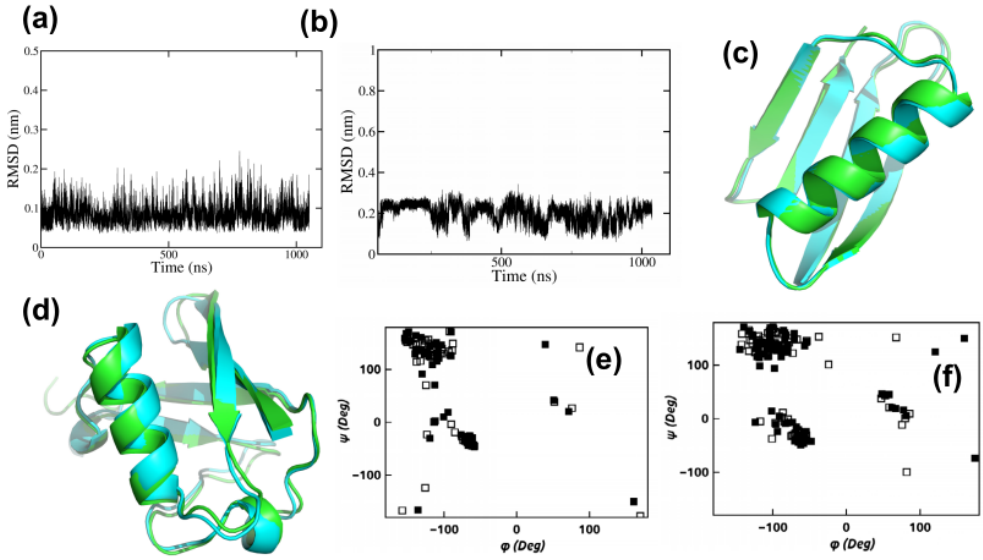

Figure 2.2: RMSD plot of (a) GB3, (b) Ub over 1.05 µs using Amber force field.Overlapped image of initial and average structure of (c) GB3, (d) Ub. Green color represent initial structure and cyan represents average structure. The Ramachandran plot of residues in (e) GB3 (f) Ub using Amber force field; Filled rectangle represents the crystal structure and hollow rectangle shows simulated average structure obtained from the equilibrated trajectory.

where the vertical axis is the logarithm of the correlation function. The solid line shows the fitted line, and the symbols show the simulated data. The plots suggest that at short time, the decay is exponential. The temporal correlation functions for the other cases show similar behaviour. We also show typical cases of dipolar fluctuations of Ub. Figure2.3(c) and (d) show the direct and cross dipolar TDCF $C^{K6}_{dipole}(\Delta t)$ for Lysine, K6. This residue belongs to the beta sheet region in the crystal structure. Here again, the TDCFs decay within a few ns to zero value. The fitted semi-log plots are shown in the in corresponding insets. The short time data in the insets confirm an exponential decay. The dipolar orientation fluctuations of the other residues show similar behaviour.

As representative cases for sequential dipolar TDCF, we consider $H^N - N$ dipole of i-th residue and $C_\alpha - H_\alpha$ dipole of $(i-1)$-th residue. Fig.2.4(a) and (b) show sequential dipolar TDCFs for residues Isoleucine (I7) and Valine (V6) of GB3 protein. The correlation function $C^{I7,V6}_{dipole}(\Delta t)$ (Fig.2.4(a)) decays to zero value within 10 ns for GB3. TDCFs for cross pair shows similar behaviour (Fig.2.4(b)). Insets show the semi-log plot of the correlation functions up to which it decays to zero, where the solid lines represent the fitted lines. We also illustrate the sequential dipolar fluctuations for Ub Figure. 2.4(c) and (d). TDCFs follows a

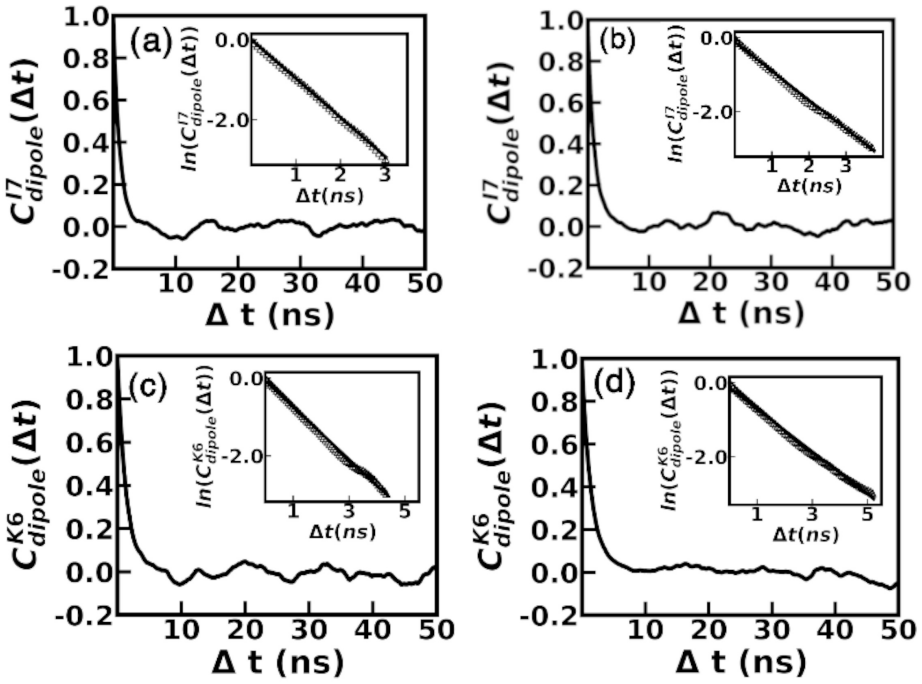

Figure 2.3: TDCFs of dipolar orientational fluctuations for Isoleucine, I7 of GB3 ($C^{I7}_{dipole}(\Delta t)$) (a) direct, and (b) cross pair of dipoles. Similar TDCFs for Lysine, K6 ($C^{K6}_{dipole}(\Delta t)$) (c) direct, and (d) cross pair, of Ub. Inset:Short time behaviour of TDCFs in terms of semi log plot of the correlation functions. The solid line represents the best linear fit.

similar trend as GB3 protein. Correlation decays to zero value within 10ns, as well as oscillations about zero value. For cross dipole pairs, long time oscillations are larger for both proteins. The short time behaviours in both cases are shown in terms of semi log plot in the corresponding insets. Here also, these plots confirm exponential decay of the TDCF in time.

We examine timescales of the exponential decay of the TDCFs. Since the long time oscillations have low amplitude, we restrict only to the short time decay shown in the insets of Fig.2.3 and Fig.2.4. We fit the initial decay data to an exponential form, $C^i_{dipole}(\Delta t)/C^i_{dipole}(0) = \exp(-\Delta t/\tau^i_{dipole})$ where τ^i_{dipole} is the timescale of decay of correlation for the ith residue, also known as the correlation timescale. We estimate τ^i_{dipole} from the slopes of linear fitting of semi-log plots in the insets. Fig.2.5(a) shows the histogram $H(\tau_{dipole})$ of the timescale where τ_{dipole} belongs to the set $\{\tau^i_{dipole}\}$ considering dipolar fluctuations, both direct and cross, for all the residues in GB3. We similarly estimate $\tau^{i,i-1}_{dipole}$ from the small-time decay in $C^{i,i-1}_{dipole}(\Delta t)$ data. This histogram $H(\tau_{dipole})$ with τ_{dipole} from the set of $\{\tau^{i,i-1}_{dipole}\}$ is also shown in Fig.2.5(a). The histograms for both intra-residual and

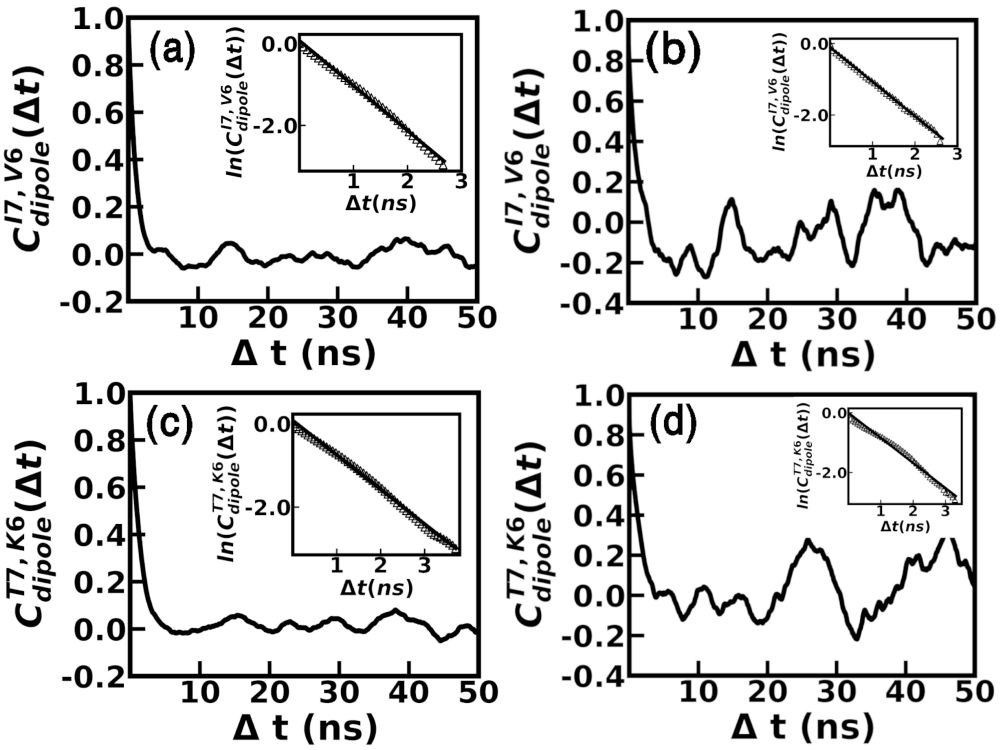

Figure 2.4: TDCFs ($C^{I7,V6}_{dipole}(\Delta t)$) of dipolar orientational fluctuations for two neighbouring residues (i^{th} with $i^{th}-1$):residues I7 and V6 of GB3 (a) for direct pair, and (b) cross pair. Similar TDCFs ($C^{T7,K6}_{dipole}(\Delta t)$) plot for residues T7 and K6 of Ub (c) for direct, and (d) cross pair of dipoles. Inset:Short time behaviour of TDCFs in terms of semi log plot of the correlation functions. The solid line represents the best linear fit.

sequential dipolar correlation timescales are quite similar. Fig.2.5(b) represents similar $H(\tau_{dipole})$ plot for Ub for both $\{\tau^{i}_{dipole}\}$ and $\{\tau^{i,i-1}_{dipole}\}$. The histograms show that the majority of the correlation timescale is just a few ns ($\sim$2 ns).

We check if the fluctuations of the dipolar orientations in the time window of the decay time of the correlated dipolar fluctuations captures the experimental CCR. The time integral of the corresponding TDCF has been computed numerically up to twice of the correlation timescale so that the initial decay of the correlations is entirely captured in the integral. Since the experimental CCRs are given in terms of the algebraic sum of zero frequency spectral function for both $H^N - N/C_\alpha - H_\alpha$ and $H^N - C_\alpha/N - H_\alpha$ dipole pairs, we compute the zero frequency spectral function $J^{(i)}_{dipole}$ as sum for both the direct and cross dipole pair for the intra-residue and similarly $J^{(i,i-1)}_{dipole}$ for the sequential dipolar fluctuations.

We consider in particular GB3, for which experimental data are available.[66,81] The experimental CCR rates are given by average of data set measured using doubly in-phase and anti-phase inter-conversion (DIAI) method and an average

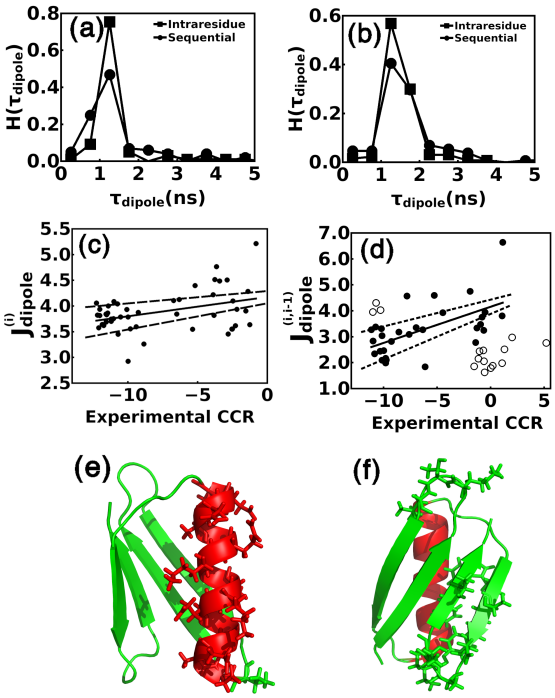

Figure 2.5: Histogram ($H(\tau_{dipole})$) of autocorrelation timescale for (a) GB3 and (b) Ub considering dipole pair $H^N - N/C_\alpha - H_\alpha$ (both direct and cross pairs). Intra-residual and sequential correlation timescales are marked with different symbol. (c) Correlation diagram between experimental CCR and theoretical $J^{(i)}_{dipole}$ for GB3 for intra-residual $H^N - N/C_\alpha - H_\alpha$ dipole, considering both dipole and cross pairs. (d) Correlation diagram between experimental CCR and theoretical J(0) ($J^{(i,i-1)}_{dipole}$) considering $H^N - N$ dipole in i^{th} residue and $C_\alpha - H_\alpha$ dipole in (i-1)th residue of GB3. Hollow scatter points are represented as outlier residues. Dipole pairs for outlier $J^{(i,i-1)}_{dipole}$ values in correlation plot, belonging to the (e) helix and (f) loop and sheet.

of data set measured using both all component evolution (ACE) and mixed multiple quantum (MMQ) method. Fig.2.5(c) shows the correlation plot between the estimated theoretical $J^{(i)}_{dipole}$, and available experimental CCR data. Fig.2.5(c) shows theoretical $J^{(i)}_{dipole}$ of the majority of the residues of GB3 are moderately correlated with the experimental CCR with the Pearson correlation coefficient, r=0.41. Amino acid Glutamine, Q2 and Glycine residues are excluded due to unavailability of experimental data. The terminal residues are not included as well. We show linear regression through the data, along with the upper and lower boundaries determined from the errors in the regression parameters. The upper boundary around the linear regression line is plotted by considering minimum intercept and maximum slope. Similarly, the lower boundary about the regression line is plotted, considering maximum intercept and minimum slope. It turns out that the majority of observations fall within the boundaries, suggesting that the short time decay of the dipolar fluctuations well describe the experimental data. Note that this is by no means obvious because the experimental data is an integrated information over experimentally available large time, while we have considered time only large enough to capture the initial decay in the integration.

Fig.2.5(d) shows correlation plot between theoretical $J^{(i,i-1)}_{dipole}$ which is a sum of the sequential direct and cross dipolar fluctuations and the experimental CCR.[81] Here also the experimental CCR rates are given by average of data set measured

using ACE and DIAI method. Terminal residues and residues next to Glycine are not considered in the calculation due to lack of experimental data. It shows that most of the residues are well correlated with experimental CCR. The measured Pearson correlation value by considering residues represented by filled circle in the Fig.2.5(d), r=0.59. Thus, the initial decay of dipolar fluctuation even for distant residues are able to capture experimental CCR efficiently. Residue pairs for which $J_{dipole}^{(i,i-1)}$ values are considered outlier (hollow circle) in correlation plot shows a specific feature. Most of the residues pair belongs to central alpha-helix region in crystal structure (Fig.2.5(e)), and a few others to loop and sheet region Fig.2.5(f).

One can as well compute angle between two dipoles (γ_{avg}) averaged over a trajectory. An estimate of CCR rate can be given $P_2(\cos\gamma_{avg})$.[64] We find that $P_2(\cos\gamma_{avg})$ show poorer correlation with experimental CCR (Fig.2.6). This establishes the importance of dynamic feature of dipolar fluctuations.

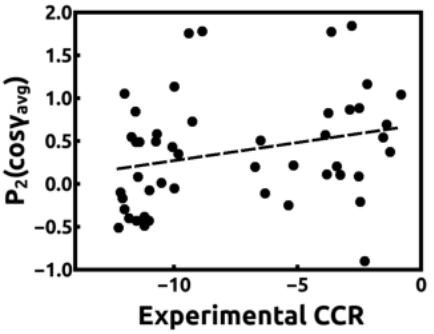

Figure 2.6: Correlation plot between experimental CCR and second order Legendre polynomial of cosine of average angle obtained from simulation. Value of correlation coefficient is 0.26.

2.3.2 TDCFs and zero frequency spectral functions for dihedral fluctuations

We show the TDCF for backbone dihedral fluctuations for representative cases in Fig.2.7 where the insets show the short time data. Fig.2.7(a) and (b) show cases for GB3. $C_{\phi\phi}^{I7}(\Delta t)$ decays to zero in 20 ns (Fig.2.7(a)). $C_{\psi\psi}^{I7}(\Delta t)$ show similar time dependence (Fig.2.7(b)). Both cases reveal that at long times, the TDCFs fluctuate around the zero value. The semi log plots in the insets represent short time behavior, which confirms exponential decay. As representative cases in Ub, we show the dihedral correlation functions for Lysine, K6 in Figs.2.7(c)-(d) and the small-time behaviour in the corresponding insets. $C_{\phi\phi}^{K6}(\Delta t)$ loses initial correlation within 10 ns (Fig.2.7(c) and inset). $C_{\psi\psi}^{K6}(\Delta t)$ fluctuates around zero value following initial decay within 20 ns (Fig.2.7(d)). The overall nature of TDCF of the dihedral fluctuations is in agreement with the previous report.[41]

The short time behaviour in the semi-log plots in the insets reveal that the initial decay has an extremely fast component of the duration of less than a ns.

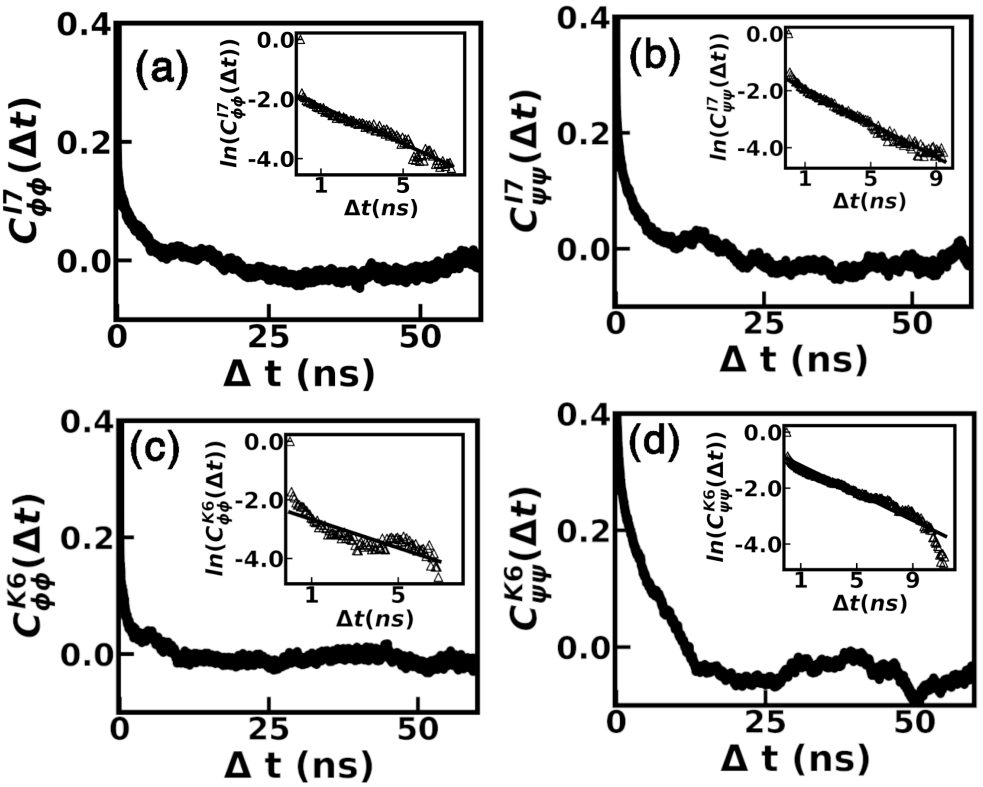

Figure 2.7: TDCFs between various dihedral fluctuations of Isoleucine, I7 of GB3:(a) $C^{I7}_{\phi\phi}(\Delta t)$, (b) $C^{I7}_{\psi\psi}(\Delta t)$; (c) $C^{K6}_{\phi\phi}(\Delta t)$, (d) $C^{K6}_{\phi\phi}(\Delta t)$ of Lysine, K6 of Ub. Insets show short time nature of TDCFs in semilog plot. The solid lines are the best linear fits and the symbols are simulated data.

This is followed by a slower decay of a few ns. This is in contrast to dipolar orientation fluctuations, which typically exhibit a single decay. Inset of Fig.2.8(a) shows the histograms $H(\tau_{dihedral})$ of the slower timescales $\tau_{dihedral}$ belonging to the set $\{\tau_{R,\phi}, \tau_{R,\psi}\}$ of the dihedral auto-correlation functions of all the residues in GB3 where $\tau_{R,\phi}$ and $\tau_{R,\psi}$ are timescales of ϕ and ψ fluctuations respectively. Inset of Fig.2.8(b) shows similar plot of histogram of dihedral auto-correlation timescale ($H(\tau_{dihedral})$) of Ub. The histograms show that the timescales of correlations for dihedral fluctuations are mostly similar to the dipolar fluctuations (Fig.2.5(a)), although in some cases the dihedral fluctuations are slower than the dipolar orientation fluctuations. This observation is in agreement with earlier work that dihedral fluctuations occur in nanosecond timescale.[41]

We consider twice $\tau_{dihedral}$ as the upper limit of integration over time for computing the zero frequency spectral function of the dihedral fluctuations to include the initial decay. We denote the zero frequency spectral function for the dihedral fluctuations by $J_R(\phi\phi, 0)$ and $J_R(\psi\psi, 0)$ for the backbone dihedral

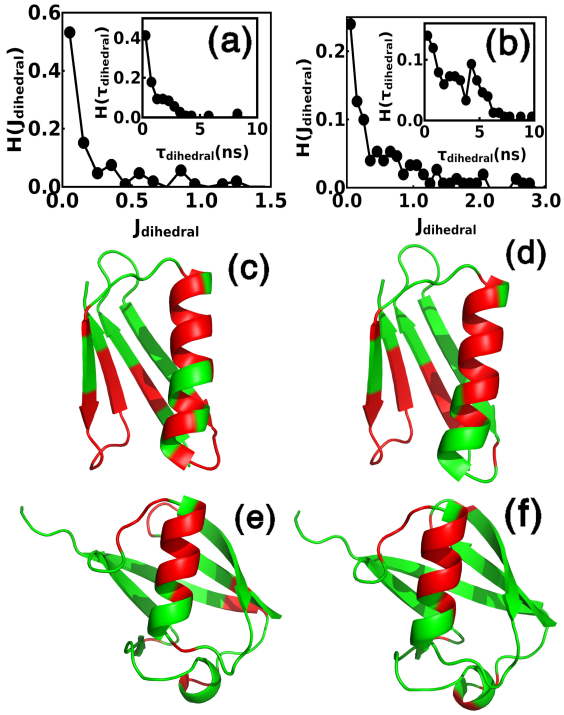

Figure 2.8: Histogram $H(J_{dihedral})$ of $J_{dihedral}$ values for (a) GB3 protein considering both ϕ and ψ dihedral angle. Inset: Histogram ($H(\tau_{dihedral})$ of correlation timescales for dihedral angle fluctuations $\tau_{dihedral}$. (b) Similar plot for Ub. Crystal structure of GB3(2OED.pdb) where red color represents residues having $J_{dihedral}$ value less than 0.1 for (c) ϕ and (d) ψ. Crystal structure of Ub(1UBQ.pdb) where red color represents residues having $J_{dihedral}$ value less than 0.1 for (e) ϕ and (f) ψ.

angles ϕ and ψ respectively for a given residue R. We show the histogram of zero frequency spectral function ($J_{dihedral}$) belonging to the set $\{J_R(\phi\phi, 0), J_R(\psi\psi, 0)\}$ considering those for both the dihedral angles of all the residues of GB3 in Fig.2.8(a) and Ub in Fig.2.8(b). We observe that there is a sharp peak for low $J_{dihedral}$ along with a tail extending to larger values. This behaviour is similar to the histograms of $\tau_{dihedral}$ (Inset). This similarity is expected since the integration over the entire time range yields the correlation timescale if the TDCF has a purely exponential time dependence. This suggests that the effect of long time tail in the zero frequency spectral functions of the dihedral fluctuations is not strong. Next, we denote the $J_{dihedral}$ value corresponding to half of the peak value of $H(J_{dihedral})$ by J_c. We mark the residues in red in Fig.2.8(c) residues with $J_R(\phi\phi, 0) < J_c$ and in Fig.2.8(d) those with $J_R(\psi\psi, 0) < J_c$. These residues primarily belong to the helix structure. Thus, the perturbation in the backbone dihedral angles of the residues belonging to the helix decay faster than the residues in the other secondary structure, which could be assigned to more stability of the helix residues.[82] Fig.2.8(e) and (f) which show that the residues having $J_R(\phi\phi, 0) < J_c$ and $J_R(\psi\psi, 0) < J_c$ respectively for Ub also belong to the helix as in the case of GB3.

In order to distinguish the dihedral fluctuations based on the secondary structural element, we calculate dihedral angle distributions for residues belonging

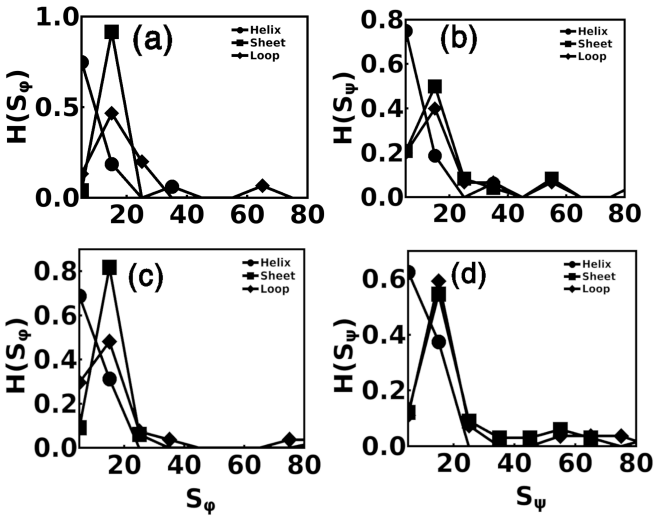

Figure 2.9: (a) Histogram $H(S_\phi)$ of standard deviation S_ϕ for distribution of dihedral ϕ and (b) histogram $H(S_\psi)$ of ψ for GB3. Different symbols are used for different secondary structure, i.e. helix, sheet and loop. (c) and (d) show similar plot for protein Ub.

to different secondary structures, namely, helix, sheet and loop. The width of the distribution, given by the standard deviation, is a measure of fluctuation of the corresponding dihedral angle. We show in Fig.2.9(a) histogram ($H(S_\phi)$) of standard deviation of distribution of dihedral angle ϕ, S_ϕ for residues based on the secondary structures of GB3. The figure shows that standard deviation of fluctuation for residues of helix region is less than compared to residues of sheet and loop region. Similar feature is present in $H(S_\psi)$ plot in Fig.2.9(b) where S_ψ stands for standard deviation of distribution of ψ. The histograms $H(S_\phi)$ (Fig.2.9(c)) and $H(S_\psi)$ (Fig.2.9(d)) of Ub show similar trend.

2.3.3 Correlation between dipolar and dihedral fluctuations

Next, we check to what extent the intra-residual dipolar fluctuations are correlated with the dihedral fluctuations. The correlation data based on the secondary structural elements are shown together for GB3 and Ub in Fig.2.10, using different symbols. We take both the proteins together, for the dihedral fluctuations show similar feature in both cases. Moreover, this allows us to have better statistics. Fig.2.10(a)and (b) show the correlation plot for dihedral ϕ and ψ respectively for the residues in helix region of the proteins. Linear regression lines are shown in both cases, considering both proteins. Pearson correlation coefficient (r) for dihedral ϕ is 0.49 and 0.45 for dihedral ψ. Fig.2.10 (c) and (d) show correlation plot for ϕ and ψ of residues in sheet region. Here r=0.35 for ϕ (Fig.2.10(c)) and r=0.46 for ψ (Fig.2.10(d)). Residues of loop region also show similar characteristics in Fig.2.10(e)-(f) where r=0.41 and r=0.34 for dihedral ϕ and ψ respectively. Linear regression line are shown in all cases. Thus, overall, the dihedral and

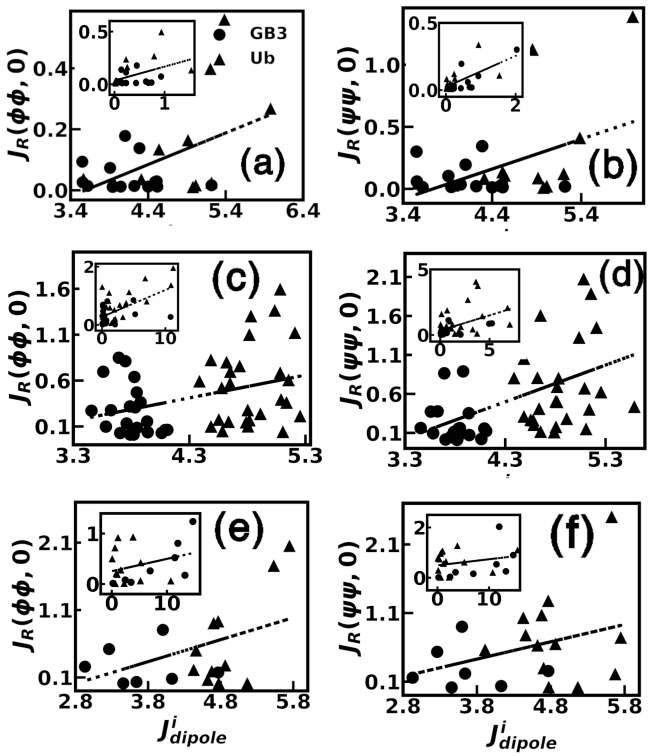

Figure 2.10: Correlation plot between zero frequency spectral functions of dihedral angle and intra-residual dipolar fluctuations of ith residue based on secondary structure.: (a) $J_R(\phi\phi,0)$ vs $J^{(i)}_{dipole}$; (b) $J_R(\psi\psi,0)$ vs $J^{(i)}_{dipole}$ for residues in the helix. (c) $J_R(\phi\phi,0)$ vs $J^{(i)}_{dipole}$, (d) $J_R(\psi\psi,0)$ vs $J^{(i)}_{dipole}$ for residues in sheet structure. (e) Similar plot for dihedral ϕ and (f) ψ respectively for loop residues. Different proteins are represented using different symbol. Inset shows similar correlation plot by considering only internal dynamics.

dipolar fluctuations show moderate correlation for these proteins, the Pearson correlations being in the range 0.35-0.45.

Next, we examine correlation plot for sequential dipolar fluctuations and the dihedral fluctuations. Residues $J_R(\phi\phi,0) < J_c$ and $J_R(\psi\psi,0) < J_c$ do not show correlation with dipolar fluctuations. Hence, we consider cases for $J_R(\phi\phi,0) > J_c$ and $J_R(\psi\psi,0) > J_c$. Since the residue pairs may belong to different secondary structure, we do not classify the residues in terms of structural elements. Fig.2.11(a) and (b) show correlation plots for $J^{(i,i-1)}_{dipole}$ with dihedral $J_R(\phi\phi,0)$ and $J_R(\psi\psi,0)$ respectively for GB3. The value of r=0.79 for ϕ and 0.55 for ψ. However, for Ub, $J^{(i,i-1)}_{dipole}$ values are poorly correlated with dihedral fluctuations $J_R(\phi\phi,0)$ (Fig.2.11(c)), while the correlation is strong between $J^{(i,i-1)}_{dipole}$ and $J_R(\psi\psi,0)$ (Fig.2.11(d))with the r=0.70. These observations suggest that ψ auto-correlations in particular are better correlated with the sequential orientation fluctuations.

Despite moderate values of the Pearson coefficients, our data suggest that the intra-residue dipolar orientation fluctuations show marginally stronger correlation with ϕ fluctuations, particularly in the helix region of the proteins. This may be related to the earlier result[46] that the intra-residue dipole vector involves dihedral ϕ. On the other hand, ψ fluctuations are well correlated to sequential

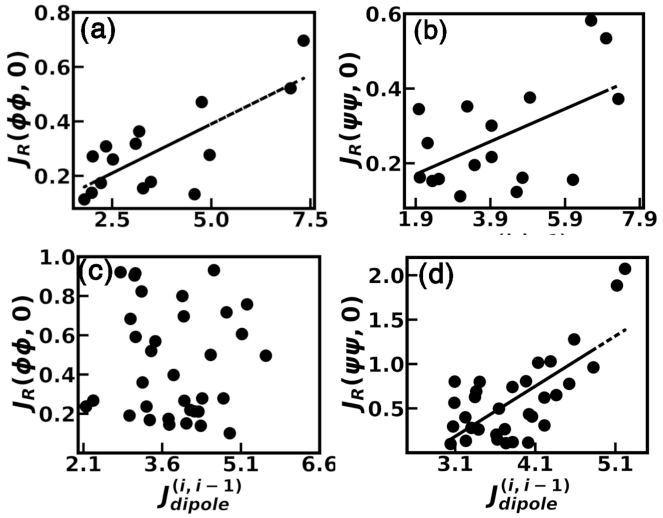

Figure 2.11: Correlation plot between zero frequency spectral functions of intra-residue dihedral angle and sequential dipole: (a) $J_R(\phi\phi, 0)$, and (b) $J_R(\psi\psi, 0)$, with $J^{(i,i-1)}_{dipole}$ for GB3 Similar plot for Ub: (c) $J_R(\phi\phi, 0)$, (d) $J_R(\psi\psi, 0)$, with $J^{(i,i-1)}_{dipole}$.

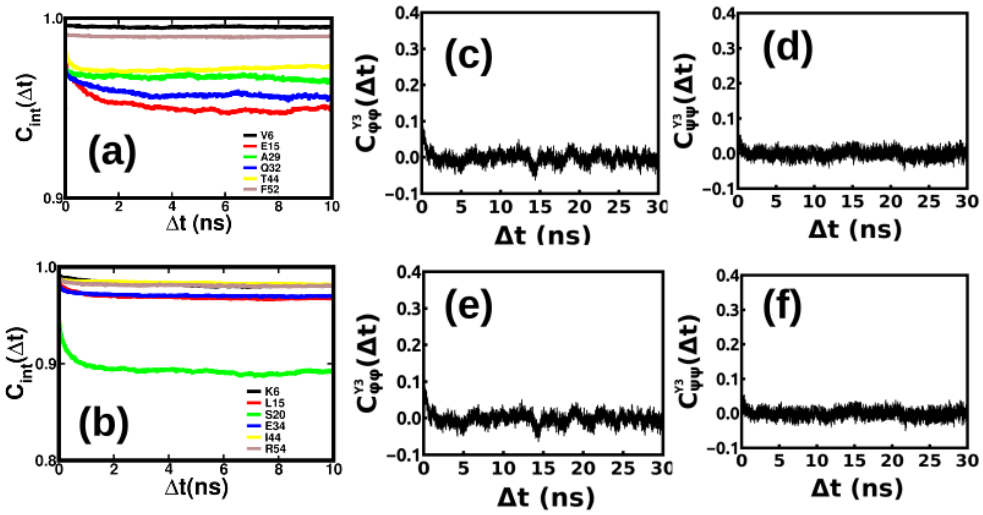

Figure 2.12: TDCFs of $H^N - N/C_\alpha - H_\alpha$ dipole pair due to the internal motion of the protein. (a) For GB3, Valine (V6), Glutamic acid (E15), Alanine(A29), Glutamine (Q32), Threonine(T44) and Phenylalanine (F62) are considered. (b) For Ub, Lysine (K16), Leucine (L15), Serine (S20), Glutamic acid (E34), Isoleucine (I44), Arginine (R54) are considered. Unit of Δt is nano second. Comparison of intrnal and total TDCF for residue Tyrosine(Y3) of GB3. Internal correlation for dihedral (c)ϕ and dihedral (d)ψ. Total correlation for dihedral (e)ϕ and dihedral (f)ψ. Nature of TDCFs are same in both cases.

dipolar fluctuations, which is qualitatively in agreement that the sequential dipolar angles involves ψ.[46] So far, we have not factored out the overall protein motion (rotation and translation) in computing the correlation functions. The contributions of the internal motion only $C^{dipole}_{int}(\Delta t)$ can be obtained by removing rotational and translational motion of protein by aligning the simulated structures to a reference one. The internal correlation function can be fitted to the dipolar

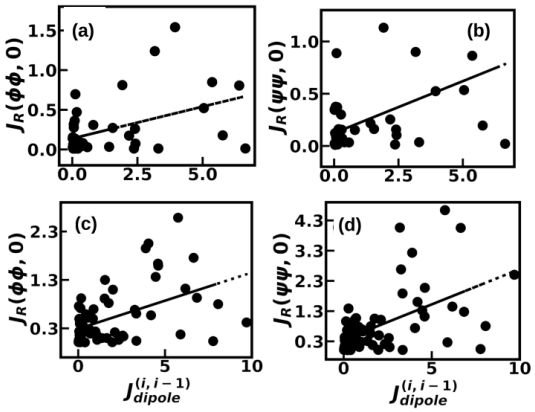

Figure 2.13: Correlation plot considering internal dynamics for GB3: (a) $J_R(\phi\phi, 0)$, (b)$J_R(\psi\psi, 0)$ with $J^{(i,i-1)}_{dipole}$. Similar plot for Ub:(c) $J_R(\phi\phi, 0)$, (d) $J_R(\psi\psi, 0)$, with $J^{(i,i-1)}_{dipole}$.

	Intra-residual	
Secondary structure	r_ϕ	r_ψ
Helix	0.44	0.66
Sheet	0.60	0.47
Loop	0.32	0.25
	Sequential	
Protein	r_ϕ	r_ψ
GB3	0.46	0.50
Ub	0.48	0.454

Table 2.1: Pearson correlation coefficients for intra-residual and sequential.

correlation decay curve using Lipari-Szabo model free[78] approach:

$$C^{dipole}_{int}(\Delta t) = S^2 + (1 - S^2)e^{-(\frac{\Delta t}{\tau_e})} \tag{2.3}$$

Here S^2 and τ_e define the order parameter and effective internal correlation time.[78] Some representative data are shown in Fig.2.12(a) for GB3 and Fig.2.12(b) for Ub. The dihedral correlations remain unaffected(Fig.2.12(c)-(f)). Fig.2.12(c)-(d) shows dihedral TDCFs considering internal motion only for dihedral ϕ and ψ respectively for protein GB3. Total correlation plot (Fig.2.12(e)-(f)) for same dihedral angles suggest that dihedral angle TDCFs nature remain unchanged. Insets of Fig.2.10 and Fig.2.13 show correlation plots between zero frequency spectral functions of the dihedral and dipolar fluctuations. The Pearson correlation coefficients shown in Table. 2.1 for different cases suggest well correlated dipolar and dihedral fluctuations even for internal dynamics.

One can generate additional correlation functions using the simulated trajectory. For instance, we construct the correlation function from the time series of the values of $P_2(\cos\gamma)$. We calculate γ from the trajectory used in the calculation of $C^{dipole}_{int}(\Delta t)$. We have calculated $J^{(i)}_2$, by integrating, the corresponding TDCF with timescale capturing the initial decay of the correlation. Fig.2.14(a) shows correlation plot of $J^{(i)}_2$ with $J_R(\phi\phi, 0)$ and Fig.2.14(b) that with $J_R(\psi\psi, 0)$ for GB3. r=0.52 for dihedral ϕ and 0.86 for dihedral ψ respectively. Both figures and high correlation coefficient suggest that dihedral fluctuation is well correlated with

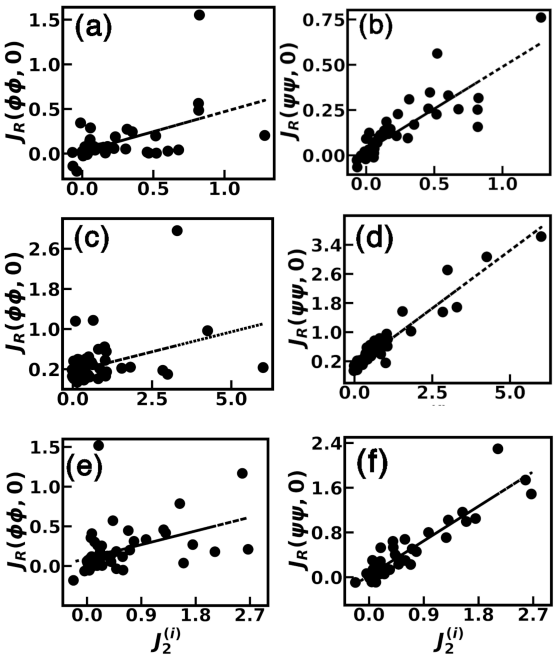

Figure 2.14: Correlation plot between zero frequency spectral functions of dihedral angle and dipole: (a) $J_R(\phi\phi, 0)$, and (b) $J_R(\psi\psi, 0)$ with $J_2^{(i)}$ for GB3; (c) $J_R(\phi\phi, 0)$ (d) $J_R(\psi\psi, 0)$ with $J_2^{(i)}$ for Ub. The Amberff are used in both cases. Correlation plot for dihedral (e) and (f) with $J_2^{(i)}$ for Ub, using the CHARMM force field.

$J_2^{(i)}$. The nature of correlation plot remains the same for Ub as GB3 (Fig.2.14(c) for ϕ and (d) for ψ) with r=0.4 for dihedral ϕ, and r=0.96 for ψ. We carry out the analysis for protein Ub using different force field like the CHARMM force field[83] and get similar results, shown in Fig.2.14(e) and (f). The values of r for ϕ and ψ are 0.44 and 0.93 respectively. Thus, the correlation pattern between $J_2^{(i)}$ and $J_R(\phi\phi, 0)$ and $J_R(\psi\psi, 0)$ appear to be independent of the protein.

2.4 Conclusions

To summarize, the fluctuations at the timescale of a few tens of nanoseconds can capture the experimental CCR of dipolar fluctuations. Within this timescale, the zero frequency spectral functions of the intra-residue dipolar fluctuations and backbone dihedral ϕ show moderate correlations, depending on the secondary structural elements. Similarly, backbone dihedral ψ show good correlation with sequential dipolar fluctuations. Our calculations predict the existence of a universal relation among the zero frequency spectral function of the dihedral fluctuations and that of the time dependent correlation function of the second order Legendre polynomial of angle between two dipoles, which needs experimental verification.Thus, the experimental CCR may act as a good marker for dihedral fluctuations as well. It may be noted that relaxation time scales may depend on the solvent model.[84] However, for constructing the correlation diagram between the two sets of data, correction due to differences in time scales is unlikely

to be important. However, it may be worthwhile to verify the correlation of fluctuations using other solvent models as well.

Appendix

A1. Molecular dynamics(MD) simulation algorithm

MD simulation[85] is necessary tools which can be used to obtain time dependent trajectory of particle numerically. Consider a system of N particles in three dimension where $\vec{r_i} = \vec{r_1}, \vec{r_2}, ..., \vec{r_N}$ and $\vec{p_i} = \vec{p_1}, \vec{p_2}, ..., \vec{p_N}$ denote the position and momentum of $i = 1, 2, ..., N$ particles. The mass of each particle is $m_i = m_1, m_2, ..., m_N$ and the total force acting within the particle is $\vec{F_i}$. Particles are assumed to be interact via conservative pair potential $V(r_{ij})$, where $r_{ij} = \vec{r_i} - \vec{r_j}$ i.e depends only on the pair separation. Hence, the force $\vec{F_i}$ acting on i^{th} particle due to all other j^{th} particles can be obtained using the gradient of $V(r_{ij})$ i.e. $\vec{F_i} = \sum_{j=1}^{N} -\nabla V(r_{ij})$. Thus, time dependent particle trajectory can be obtained using Newton's second law of motion, $\vec{F_i} = m_i.\vec{a_i}$, where a_i is acceleration of i^{th} particle. In general, Verlet algorithm[85] is widely used algorithm which is based on central difference algorithm. In this algorithm, position of particle, $\vec{r}(t + \Delta t)$ at later time $t + \Delta t$ can be obtained from the position and acceleration of particle at time t and positions from the previous step, $\vec{r}(t - \Delta t)$.

Using Taylor series expansion,

$$\vec{r}(t + \Delta t) = \vec{r}(t) + \Delta t.\vec{v}(t) + \frac{1}{2}.\Delta t^2.\vec{a}(t) + ... \tag{2.4}$$

$$\vec{r}(t - \Delta t) = \vec{r}(t) - \Delta t.\vec{v}(t) + \frac{1}{2}.\Delta t^2.\vec{a}(t) + ... \tag{2.5}$$

Finally, combining these equations gives,

$$\vec{r}(t + \Delta t) = 2\vec{r}(t) - \vec{r}(t - \Delta t) + \Delta t^2.\vec{a}(t) \tag{2.6}$$

The velocity can be obtained using the formula,

$$\vec{v}(t) = \frac{\vec{r}(t + \Delta t) - \vec{r}(t - \Delta t)}{2\Delta t} \tag{2.7}$$

The trajectories from MD simulation is useful to calculate both static and dynamic quantities of a system.

A2. Force-field used in bio-molecular simulation

In our study, we used GROMACS and AMBER simulation package which used parallel computation for biomolecular simulation. Here we performed all atom MD simulation for protein. In this package, the interaction between biomolecules and interaction of biomolecules with solvent or any other biomolecules is governed by some well established force field like CHARMM, AMBER and GROMOS. The parameters of force field are derived based on semi empirical quantum mechanical calculations or by fitting experimental data like X-Ray, neutron and electron diffraction, NMR etc. The result of the simulation can be improved based on accurate choice of the force fields. The force field contains both bonded and non-bonded interactions. The bonded interactions is expressed bond stretching, bond rotations, and torsional dihedrals via simple harmonic oscillations. Non bonded interaction includes Lennard-Jones(LJ) and Coulomb interactions.

The form of the potential energy is:

$$V = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi[1 + \cos(n\phi - \delta)] + \sum_{improper} k_w(w - w_0)^2 + \sum_{LJ} 4\epsilon[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6] + \sum_{elec} \frac{q_i q_j}{4\pi\epsilon_r\epsilon_0 r_{ij}} \tag{2.8}$$

Here, the first term is related to energy cost due to bond stretching where k_b is bond force constant and $(r - r_0)$ is deviation of bond length from equilibrium position r_0. Similarly, second term accounts for change in bond angles from equilibrium value θ_0 with force constant k_θ. In the third term, k_ϕ is dihedral force constant, n represents the multiplicity number and δ is phase shift. Fourth term is for the improper i.e. deviation of outer plane bending from equilibration (w_0). The last two term represents non-bonded interaction between pair of atoms where ϵ corresponds to the depth of the potential and σ_{ij} accounts for the distance at which intermolecular potential between two particles is zero. The last term is responsible for the Coulombic interactions, where q_i and q_j corresponds to charge of the i^{th} and j^{th} particles. $r_{ij} = |r_i - r_j|$ represents the magnitude of the distance between particles and ϵ_0, ϵ_r corresponds to permittivity of the vacuum and relativity permittivity respectively.

A3. Periodic boundary condition and minimum image convention

In the all atom MD simulation, system is initially kept in a central box. Now, in order to mimic truly infinite bulk system, periodic boundary condition (PBC)

is implemented in all direction. Now finite length box size, L can influence surface effect, so that particles near the surface exposed to different force than the bulk. Now this effect can be minimized by considering that the central box is surrounded by infinite replica. Hence, if any atom leaves the simulation box during simulation, its image will enter the box via opposite face. In the course of simulation, PBC is accompanied with minimum image convention, where one particle will interact with the closest image of another atom among all the boxes. A cut of ($r_c \sim L/2$) is introduced for truncating the long-ranged interaction to avoid the interaction between an atom in centre box with its mirror image in other replica boxes. This approach makes simulation less expensive.

A4. Particle Mesh Ewald (PME) methods

Long range interactions goes off as r^{-n}, with n$\leq$4. We use particle mesh Ewald (PME)[86] method to take care long ranged contributions of electrostatic interactions. In this method, the interaction potential is splitted into two parts; the long range interaction is estimated in Fourier space and short range part is estimated in real space. Discrete Fast-Fourier transform (FFT) is used to approximate reciprocal-space term of the standard Ewald summation using a discrete convolution on an interpolating grid.

Chapter 3

Conformational fluctuations in the molten globule state

3.1 Introduction

Some proteins show structural fluctuations in certain parts in a near denaturing condition, while retaining their overall tertiary structures. Such states are called Molten Globule (MG) state of the protein.[15] The MG state is induced by various denaturing conditions like high temperature, pH, high pressure and due to the presence of various denaturing chemicals like urea.[16,87,88] MG states, despite having structural fluctuations, show binding with ligands. Many of such complexes have functional relevance. For instance, the complexes of fatty acids and MG of α-lactalbumin (aLA) protein, ubiquitously present in milk,[19,89,90] in an acidic solvent show cytotoxic activities against cancer cells. Such potential applications make the MG states interesting. However, the structural and functional characterization of the MG is largely lacking, since the MG states are not directly amenable to crystallization.

Intrinsic disordered proteins (IDP)[18] are also known to lack well-defined conformation and exist as highly dynamic conformational ensemble either at secondary or tertiary level. The IDPs inherently posses meta-stable conformations[91] instead of well-defined minima in energy landscape, unlike that observed in ordered protein. The intrinsic dynamic nature helps IDP to participate in functional ligand binding.[92] Likewise, the structures of a protein, like aLA in the

Based on publications: Conformational fluctuations in the molten globule state of α-lactalbumin, Abhik Ghosh Moulick, J. Chakrabarti, Phys. Chem. Chem. Phys., 2022, 24, 21348.

MG state, are highly dynamic and flexible.[17] Experimental studies using time resolved fluorescence suggest that conformation fluctuations in the MG state occur on nanosecond (ns) time scales.[93] However, it is not a priori evident that the conformation fluctuations in MG state of such protein have resemblances to those in IDP, since the MG states are induced by external conditions quite unlike the IDP.

With this backdrop, we study the microscopic nature of conformation fluctuations in the MG state of aLA. Fig.3.1 shows the crystal structure of Ca^{2+} loaded (holo-) aLA. The data show that the protein has a α-helical domains (A1-A4) and a beta sheet domain (B1-B3) separated by a cleft. Under physiological conditions, the Ca^{2+} ion is coordinated to this protein by the carbonyl oxygen of Lysine (LYS)79 and Aspartate (ASP)84, side chain carboxylates of ASP82, ASP87, ASP88 and the crystal waters. At acidic pH, protons compete with Ca^{2+} for the carboxylate oxygen and disturb the Ca^{2+} ions coordination. Fluorescence and Circular dichroism (CD) data reveal that binding of Ca^{2+} to aLA show pronounced change in structure and function of the protein.[94–96] Calorimetric study shows that Ca^{2+} ion reduce molecular flexibility and increases thermal stability of the protein.[97] The removal of the ion reduces the overall stability of the protein[98] resulting MG states of the protein. The MG state of aLA show binding with fatty acids[99–103] like oleic acid (OLA) having cytotoxic activities. Thus, aLA in the MG state acts like a carrier of cytotoxic factors. These complexes are called XAMLET, namely, aLA made lethal where X stands for the name of the mammal.[104–106]

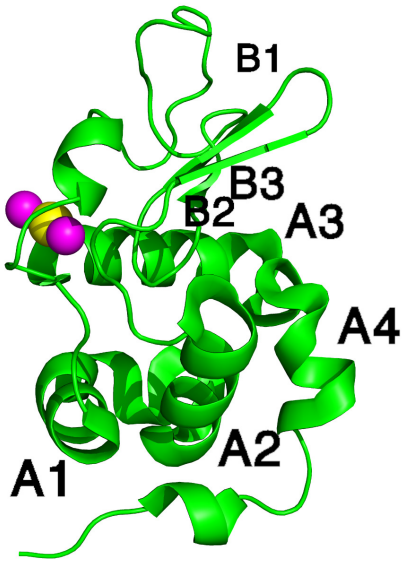

Figure 3.1: Initial crystal structure of holo α-lactalbumin protein. Ca^{2+} ion is shown in yellow, and crystal water participating in the coordination of the ion is shown in magenta. The secondary structure element of the alpha-helical (A1-A4) and the beta-sheet (B1-B3) domains are marked.

Various biophysical techniques are extensively used to characterize MG state of aLA. Secondary structure and tertiary packing interactions in MG state are

probed using the CD method.[107] Refolding kinetics of aLA is addressed following the time-dependent change in the CD spectra. The decay time of aLA is affected by its holo- and apo-forms. Amide hydrogen exchange measurements have been used to investigate the most persistent structure in MG state.[108]

Molecular dynamics (MD) simulations are essential tools to study MG state of protein[17,109,110] due to lack of crystallization. However, the traditional MD simulations have severe limitations in studying the MG state of aLA. The solution pH plays a vital role in the molten globule formation of aLA by changing the protonation states of titrable residues.[17] pK_a value determines the protonation states of a given residue in a protein, which in turns depend on both intra- and inter-molecular electrostatic interactions with neighboring residues. Traditional molecular dynamics simulations use fixed protonation state of titrable residues. As a result, pH must be explicitly included as an external parameter that allows the protonation state of titrable residues to change in response to changes in the chemical environment. Recent development in this direction uses a discrete method where the protonation states of the titrable residues are updated using the Monte Carlo moves as per the generalized Born energy cost due to changes in the charged states of the residues in a continuum.[111] Such scheme, also known as constant pH molecular dynamics method(CpHMD),[111] has been utilized using the continuum solvent model to get MG states of aLA.[17] The MG state has been shown in canine milk lysozyme,[110] using normal molecular dynamics simulation at different temperatures and the replica exchange umbrella sampling method at implicit solvent conditions. These works, however, do not address the conformation fluctuations in the MG state.

Since maintaining low pH is essential for the formation of the MG state of aLA, we perform CpHMD via a hybrid scheme where explicit water molecules are taken into account in the MD simulations, but the charged state is updated using the implicit solvent model. The water molecules are explicitly taken to ensure coordination of the ion in neutral pH conditions. We find that the structure and ion coordination from the hybrid simulation compare well to those from normal all-atom calculations at neutral pH. Next, we study aLA at low pH(=2) to take the system in MG state. We characterize the MG state in terms of radius of gyration (R_g) and contact map. We also examine internal motion of the protein at neutral and pH=2.0 condition in terms of Lipari-Szabo order parameter (S^2)[78] and internal correlation time (τ_e).[72,112] We employ the dihedral angle based principal component analysis,[23] the density based clustering and the machine learning techniques[8,20–25]to identify meta-stability in MG state. In MG state, we

find that R_g increases on average while native contact decreases. We observe a decrease in S^2 along with an increase in τ_e. Our analysis reveals meta-stability via a number of helix residues in the crystal structure. In the MG state, these residues lack well-defined secondary structure, have lower structural persistence (S_P), a longer dihedral autocorrelation timescale, and dynamic correlations with fatty acid binding residues.

3.2 Methods

3.2.1 System preparation

The protein used in this study is bovine α-lactalbumin in both holo (with Ca^{2+} ion, RCSB PDB ID: 1F6R) and apo (without Ca^{2+} ion, RCSB PDB ID: 1F6S)[113] form. Initial structure is shows in Fig.3.1. Both initial structures have six identical chains, in which we choose only the first chain for the simulation.

3.2.2 Simulation Details

We use the protocol suggested by Swails et al.[114] to simulate CpHMD in an explicit solvent, keeping the number of particles (N), volume (V) and temperature (T) fixed. At first, MD (See chapter 2, Appendix A1) is carried out with a constant set of protonation states. Then the MD is stopped, and the solvent is stripped. Here the potential is changed to the Generalized Born (GB) potential and protonation states are changed for each titrable residue randomly. The electrostatic energy due to this transition is used to make a Monte Carlo decision regarding the acceptance or rejection of a new protonation state.[111] If an attempt is accepted, the solute is frozen and MD is performed on the solvent to relax the solvent distribution around the new protonation states.

The AMBER[115] package gives definitions for titrating side chains of aspartate (ASP), glutamate (GLU), histidine (HIS), lysine (LYS), tyrosine (TYR), and cysteine (CYS). We use ASP and GLU for titration at pH=2. Due to high acidic pH (=2), HIS is excluded. At neutral condition, pH=7, only HIS is considered for titration as pK_a (HIS) = 6.5. Since, CpHMD simulation in the AMBER package does not support C or N terminal titrable residues, they are excluded for titration during simulation. Systems are parametrized using the AMBER FF10 force field (ff).[75] This ff is equivalent to the AMBER FF99SB for protein. TIP3P[77] water is used as solvent in truncated octahedron box around protein. Total 7349 water

molecules are added in the box. The system is neutralized by placing 8 Na^+ atoms randomly. The total number of atoms inside the box is 24077. The volume of the box is 272559.837 $Å^3$.

After making the topology, the system is minimized using 5000 steps, following both the steepest descent and the conjugate gradient with 10kcal/mol.$Å^2$ positional restraint applied to the backbone. The system is then heated at constant volume, varying the temperature slowly from 10K to 300K over 1000ps. The salt concentration of the system is set at the default value of 0.15M. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[116] The heated structures are equilibrated for 5 ns, maintaining a constant temperature using Langevin dynamics and a constant pressure using the Berendsen barostat. Consecutive MC trials are separated by 2 femtosecond(fs) time steps. The production run is done using a similar protocol for a further 195 ns at constant volume and temperature using Langevin dynamics. We also perform normal molecular dynamics of the holo form of α-lactalbumin protein without CpHMD using the AMBER FF99SB force field for 200 ns.

We performed a total of 4 sets of simulations: (1) the CpHMD and (2) normal MD on holo-aLA in neutral medium (PH=7.0) in explicit water for validation of the CpHMD method we adopted. (3) Then we perform a simulation of apo-aLA at acidic pH(=2.0) using the CpHMD. (4) We further carry out CpHMD simulation on apo-aLA in neutral medium.

3.2.3 Analysis

Structural characterization

We calculate the following quantities to characterize the structures:

1. The root mean square fluctuations (RMSF) for the backbone C_α atoms.

2. The radius of gyration (R_g) is calculated as the average distance of C_α an atom from their centre of mass ($\vec{R}_{CM}$). The square of R_g is defined as,

$$R_g^2 = \frac{\sum_i m_i(\vec{r_i} - \vec{R}_{CM})^2}{\sum_i m_i.} \tag{3.1}$$

   Here m_i and $\vec{r_i}$ is the mass and position vector of i-th C_α atom.

3. Native contacts are taken to exist between two residues if two specified atoms of the two residues are closer than a specific cutoff value.

All analysis are done using CPPTRAJ tools of AMBER.[117]

Structural persistence

The structural persistence (S_p) parameter is defined as:[118]

$$S_p = \frac{1}{N_{res}} \sum_{i=1}^{N_{res}} e^{-(\Delta\phi_i/\Delta\phi_{max})}.e^{-(\Delta\psi_i/\Delta\psi_{max})} \tag{3.2}$$

Here, N_{res} denotes the total number of residues. $\Delta\phi_i$ and $\Delta\psi_i$ represents change in dihedral ϕ and ψ of i-th residue with respect to initial crystal structure. $\Delta\phi_{max}$ and $\Delta\psi_{max}$ denotes the maximum permissible change in dihedral angle in Ramachandran space without considering direction. S_p = 1 denotes no change in protein conformation, whereas low S_p represents a higher deviation from the reference structure. Structural persistence for a single residue is calculated in a similar manner. However, summation and averaging over time is done for a single residue instead of the full protein.

Correlation functions and order parameter

In general, dipolar interaction between two nuclei can be measured using NMR. The correlation function for such cases is defined as:

$$C_{tot}(t) = \langle P_2(\hat{\mu}_i(t).\hat{\mu}_i(0))\rangle \tag{3.3}$$

Where $\hat{\mu}_i(t)$ is the unit dipole moment vector pointing along the given dipole, and $P_2(x) = (1/2)(3x^2 - 1)$ is a second order Legendre polynomial. $\langle..\rangle$ represents that averaging is performed over all given dipoles at different time origins. For macromolecules like protein, the timescale associated with internal motion is faster than overall motion, and hence, these two motions could be considered independently. Thus, $C_{tot}(t)$ can be factorized into correlation functions for overall motions ($C_O(t)$) and internal motions ($C_I(t)$) of the protein. $C_I(t)$ can be calculated after removing the center of mass of rotational and translational motion of protein with reference to initial structure.

The relaxation behavior of $C_I(t)$ can be quantified by fitting the $C_I(t)$ decay curve with the Lipari-Szabo model free approach[78]

$$C_I(t) = S^2 + (1 - S^2)e^{-(\frac{t}{\tau_e})} \tag{3.4}$$

Where S^2 and τ_e are defined as Lipari-Szabo order parameter and effective internal correlation time of a given dipole. It is important to note that S^2 and $\langle\tau_e\rangle$ can be calculated from NMR relaxation data. Parameter S^2 denotes the level of spatial rigidity of a given dipole. S^2=1 denotes totally rigid state, whereas 0 value represents entirely free motion. Here we considered two dipoles of the protein backbone, i.e. N-H and C-N dipoles. Here, internal correlation functions ($C_I(t)$) averaged over all C-N and N-H dipoles respectively at both neutral and pH=2.

Essential coordinate identification using clustering and machine learning technique

We use a systematic approach to identify low dimensional essential coordinate (EC) of the system using modern clustering and supervised machine learning technique as suggested by Brandt et al.[8] We start with 3N dimensional coordinate of the system obtained from MD simulation. Therefore, we calculate protein backbone dihedral angle ϕ and ψ, referred as MD coordinate using inhouse code. Initially, principal component analysis technique based on dihedral angle by Sittel et al.[23] is used to obtain low dimensional description of the dynamics of the system. The detail of the method is in Appendix A1. Next, density based and dynamical based clustering techniques is used to identify metastable conformational state. Details of the clustering is in Appendix A2-A3. In the final step, a supervised machine learning technique has been used to obtain EC of the system. Details of the method is in the Appendix A4.

Dynamical cross correlation

Correlated motion between various protein segment can be defined in terms of simple cross correlation $C(i,j)$ functions[119,120] of various residues. If Δr_i and Δr_j are the displacement vectors of i-th and j-th C_α atom of the protein, then corresponding $C(i,j)$ is defined as,

$$C(i,j) = \frac{\langle \Delta r_i . \Delta r_j \rangle}{\langle \Delta r_i^2 \rangle^{1/2} \langle \Delta r_j^2 \rangle^{1/2}} \tag{3.5}$$

Here, the angular bracket signifies the ensemble average. $C(i,j)$ varies between -1 (anti correlated motion) and +1 (correlated motion). The movement of correlated residues is in the same direction, whereas anti-correlated residues are in the opposite direction. It is noted that we have used the trajectory at $t = 0$ as the reference structure, and all other structures are aligned with respect to

that reference structure. Calculations are done using the Bio3D suite of R programming packages.[121]

3.3 Results

Residue Name	Residue	Predicted pKa Value	Offset Value	Fraction of protonation
7	GLU	3.675	1.675	0.956
11	GLU	3.334	1.334	0.987
14	GLU	3.869	1.869	0.794
25	GLU	2.586	0.586	0.998
37	ASP	3.239	1.239	0.945
46	ASP	3.652	1.652	0.978
49	GLU	3.933	1.933	0.988
63	ASP	2.472	0.472	0.748
64	ASP	3.270	1.270	0.949
78	ASP	2.714	0.714	0.838
82	ASP	2.594	0.594	0.797
83	ASP	3.077	1.077	0.923
84	ASP	2.784	0.784	0.859
87	ASP	2.052	0.052	0.530
88	ASP	3.499	1.499	0.969
97	ASP	3.392	1.392	0.961
113	GLU	4.661	2.661	0.998
116	ASP	3.282	1.282	0.950
121	GLU	4.091	2.091	0.992

Table 3.1: Predicted pKa values of titrable residues during CpHMD simulations at pH=2. The offset value is defined as the difference between predicted pKa and system pH. Fraction of time titrable residues remain protonated during simulations.

At first, we validate structures obtained from CpHMD and normal MD simulation of the holo-structure at neutral pH. Fig.3.2 (a) shows the overlapped structures. The root-mean-square deviation between the atoms is found to be only 0.479 Å with similar overall secondary structures. This shows the validity of the hybrid simulations. Table 3.1 shows the predicted pK_a value of titrable residues obtained from the CpHMD simulations at pH=2. This prediction is considered accurate if the offset value, i.e., the absolute value of the difference between the predicted pK_a and system pH, is less than 2.0.[111] Table S1 shows that most of the titrable residues have an offset value of less than 2.0.

After equilibration, the whole trajectory is divided into 4 windows. Window

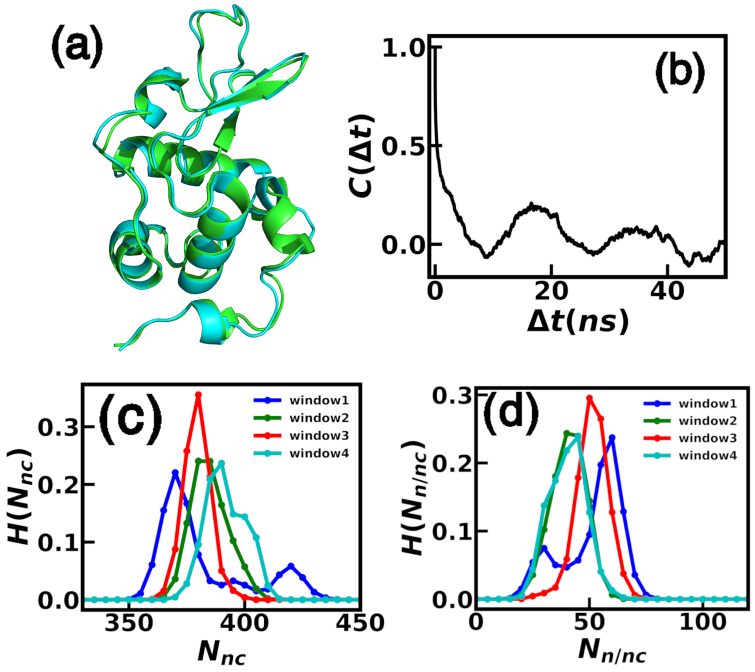

Figure 3.2: (a) Overlapped average structure obtained from normal MD simulation (green) and constant pH MD simulations (cyan) at neutral pH. Root mean square distance between two structure is 0.479 Å, (b)Autocorrelation plot of radius of gyration (R_g). Histogram of (c) native (N_{nc}) and (d) non-native($N_{n/nc}$) contact for different window. Different windows are represented as different colours.

size is chosen based on autocorrelation of radius of gyration (R_g) where R_g is calculated using Eq.3.1. Fig.3.2(b) shows autocorrelation plot of R_g. The decay of correlation takes place within 40ns. We consider each window of ∼45ns to ensure that they can be treated as independent. We check convergence of the simulation based on native and non-native contact between residues based on C_α atom in each window. Fig.3.2(c)-(d) shows histogram of native (N_{nc}) and non-native ($N_{n/nc}$) contact for each window. The overlaps of the histograms in native and non-native contact suggest the convergence of the simulated structures of apo-aLA using the CpHMD method. We check the convergence of all other quantities in a similar way. We use the histograms of the quantities of interest over the windows and compute the window mean. The mean values of the window means are taken as the mean of the given quantity. The error of the mean is given by $\sigma/\sqrt{n}$ (n=4), where σ is the standard deviation of the mean values over the windows and n represents the number of window. The average value of native contact is 382.54±2.67 and non-native contact is 43.93±3.08. In the following, "neutral" refers to the holo-aLA in neutral solvent condition with normal MD production runs.

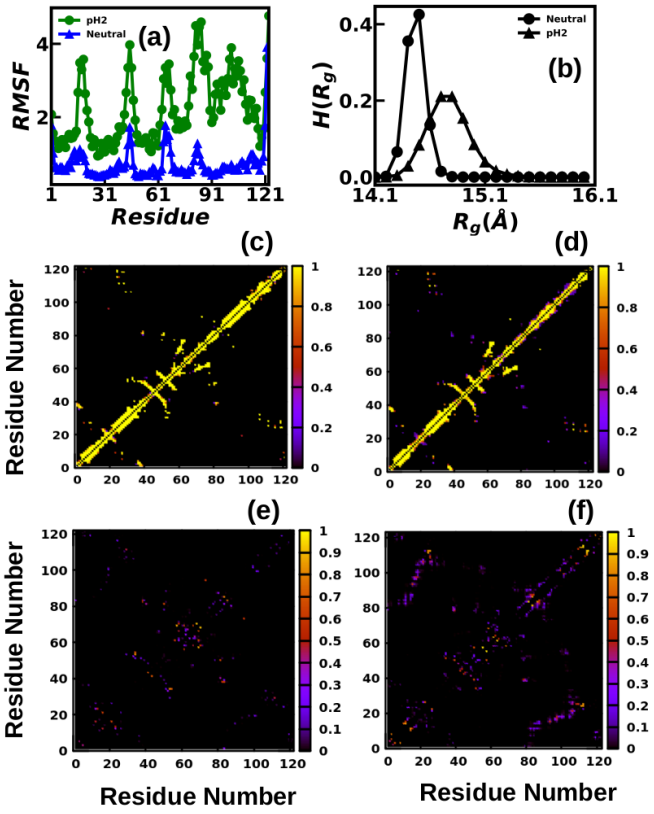

Figure 3.3: a) RMSF per residue for both apo-aLA at pH2 and holo-aLA at neutral. Histogram of (b) radius of gyration ($H(R_g)$), Contact map of protein-native contact at (c) neutral, (d) pH2 and non-native contact at (e) neutral and (f) pH2. Native contact decreases and nonnative contact increases at pH2 as compared to neutral.

3.3.1 Conformations in the MG state

Fig. 3.3(a) shows RMSF per residue for holo-aLA at neutral solvent using normal MD and apo-aLA at acidic solvent using the CpHMD simulations. The data suggest that RMSF increases at acidic pH compared to the holo-aLA in neutral condition. Radius of gyration (R_g) as given in Eq.3.1 measures compactness of any protein. R_g is computed for each conformation. Fig. 3.3(b) shows histograms $(H(R_g))$ of R_g for both holo-aLA in neutral and apo-aLA in acidic pH using over a representative window. The average value of R_g obtained from simulation using window averaging at MG state is 14.75$\pm$0.03 Å is slightly larger than holo-aLA in the neutral case, with an average $R_g = 14.40\pm0.01$Å. Qualitatively, this is in agreement to earlier experimental observations and simulations.[122–125]

The native contacts in a denatured protein decrease, as revealed by the nuclear magnetic resonance (NMR) chemical shift dispersion spectra.[17,126,127] Fig. 3.3(c)-(d) shows native contact map of holo-aLA at neutral medium and apo-aLA at pH=2 with 7 Å cut off between C_α atom as reference atom of the residues. The native contacts decrease at acidic pH. Non-native contact map is shown in Fig. 3.3(e) for neutral and (f) for pH=2 case. The maps show that non-native contacts, on the other hand, increase at acidic pH than in the neutral case, in

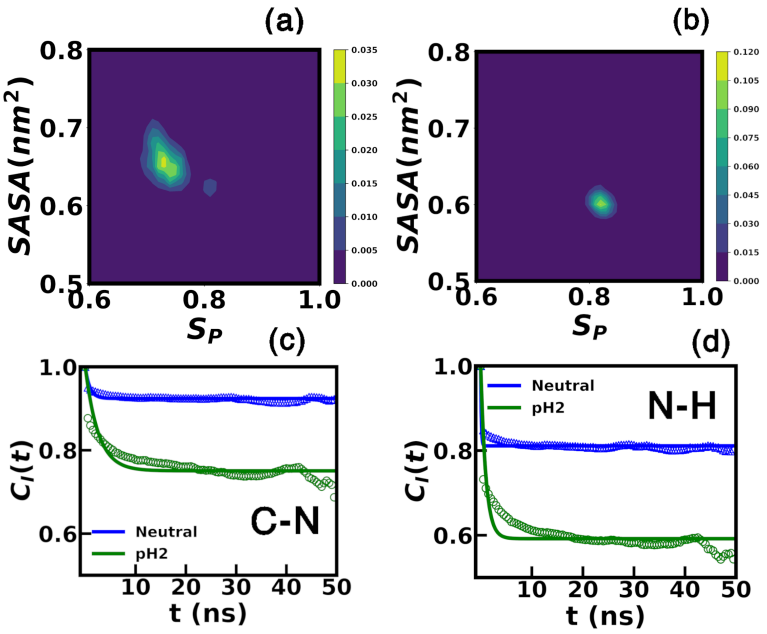

Figure 3.4: Joint probability distribution of SASA and S_P at (a) pH2 and (b) neutral. Internal correlation functions ($C_I(t)$) for the (c) C-N bond dipole and (d) N-H bond dipole. The symbols show the original curve, while the fitted line is shown in solid.

agreement to earlier simulation studies.[17]

Earlier works show changes in the solvent accessible surface area (SASA) in the molten globule state.[17] Fluorescence studies suggest that solvent accessibility of Tryptophan residue in the MG state increases as compared to the neutral state.[128,129] Earlier implicit solvent based analysis on MG state also shows similar result.[17] Accordingly, we calculate the SASA. Table 3.2 shows the SASA values of Tryptophan(W) residue at neutral and molten state. We observe increase of SASA of Tryptophan(W) at MG state compared to the neutral case. The structural persistence (S_P) (see Eq.3.2) is sensitive to changes in structural preferences. We show the joint probability distributions of SASA and S_P at both acidic and neutral pH in Fig.3.4(a) and (b) respectively. Compared to neutral pH, we observe that the lower structural persistence is correlated with a higher SASA value at acidic pH. Thus, the apo-aLA in acidic medium is in MG state.

Condition	W60	W104	W118
Neutral	4Å^2	16Å^2	19Å^2
pH2	19Å^2	96Å^2	34Å^2

Table 3.2: SASA value of Tryptophan(W) residues at neutral and acidic pH.

We also consider the dynamics of protein backbone N-H and C-N dipoles. Fig.3.4(c) and (d) shows internal correlation functions ($C_I(t)$) (see Eq.3.3) averaged over all C-N and N-H dipoles respectively for aLA at acidic pH and holo-aLA at neutral condition. $C_I(t)$ decay slowly at acidic pH compared to the

Dipole	Condition	S^2	τ_e(ns)
N-H	Neutral	0.82(0.003)	0.08(0.02)
N-H	pH 2	0.61(0.02)	0.82(0.17)
C-N	Neutral	0.93(0.003)	0.8(0.06)
C-N	pH 2	0.77(0.02)	2.40(0.17)

Table 3.3: Order parameter (S^2), internal correlation time (τ_e) for the backbone N-H dipole and C-N dipole of α-lactalbumin protein at both neutral and pH2. Error obtained using window analysis are shown in parentheses.

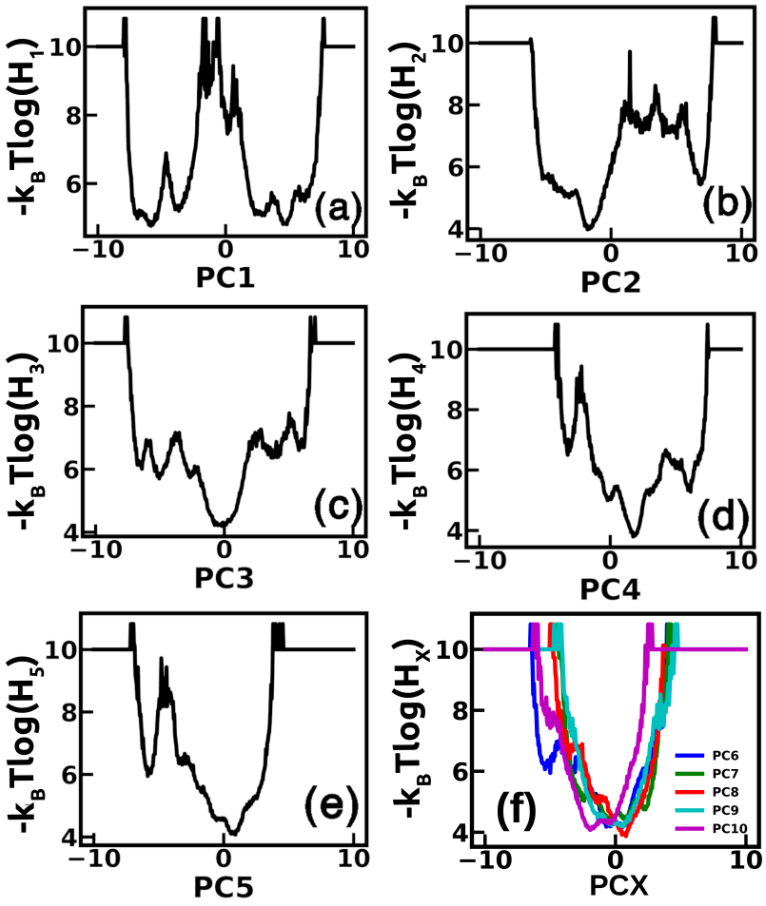

Figure 3.5: Free energy landscape obtained from dPCA+ for (a) PC1, (b)PC2, (c)PC3, (d)PC4, and (e)PC5. (f) PC6-10. Y axis represents negative log of population of PCs.

neutral case. We quantify the correlation relaxation further by fitting $C_I(t)$ using Eq.3.4. The fitted parameters of Eq.3.4 are shown with error bars in Table 3.3. The data suggest that lowering the pH reduce the order parameter (S^2) values for both dipole, while the correlation time (τ_e) increases in comparison to the neutral holo-aLA. The reduction in S^2 is consistent with the formation of the MG state.

3.3.2 Meta-stability in the MG state

Next, we consider the fluctuations in the MG state in the dihedral angles using the dPCA+ method[23] over the simulated trajectory (see in the Method section). One dimensional free energy landscapes for PC1 to PC5 are shown in Fig.3.5(a)-(e) all of which show the presence of meta-stable states. Representative cases of PC 6-10, where meta-stability is absent, are shown in Fig.3.5(f). Metastability gradually decrease for higher PCs. Fig.3.6(a) shows auto-correlation plot($C_X^{PC}(\Delta t)$) using where $X \in (1, 5)$. We fit the data to an exponential form, $C_X^{PC}(\Delta t) = A\exp(-\Delta t/\tau_{PC})$, where τ_{PC} is the auto correlation timescale. τ_{PC} is maximum for PC1, represents the slowest mode in the system. We chose PC 1-5 for further analysis.

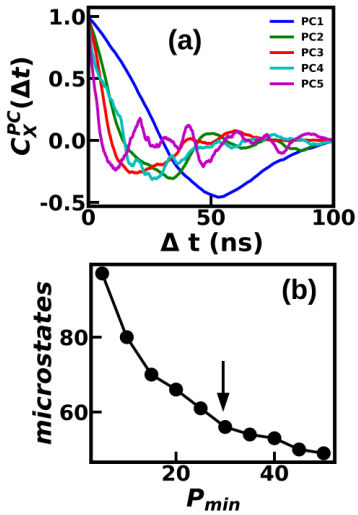

Figure 3.6: (a) Autocorrelation function of principal component 1-5, (b) Number of microstate, plotted as function of the minimal population P_{min}. P_{min}=30 (shown by arrow) value used in the analysis to avoid initial drop .

Next, we use the density based geometrical cluster analysis[22] over the hyperspace spanned by the dihedral PC1-5 (see Methods for details) to identify high density clusters. Fig.3.6(b) shows how the number of microstates changes as P_{min} increases. It is required to choose P_{min} large enough to avoid an initial drop. Here, we chose, $P_{min} = 30$ which provides 56 microstates based on clustering analysis. Next, we use the most probable path (MPP) algorithm (see Methods for details) to reduce microstates into a set of meta-stable states. We obtain 27 metastable conformational states using a lag time of τ=40 picoseconds and a minimum meta-stability of Q_{min}=0.92, eliminating spurious transitions in the vicinity of the energy barrier by coring as discussed in Nagel et al.[25] The first 6 highly populated states out of 27 metastable states carry about 50% of the total population. Fig.3.7(a)-(e) shows the representative secondary structure of first five populated state. Each conformation from metastable states supports conformational fluctuations in MG state.

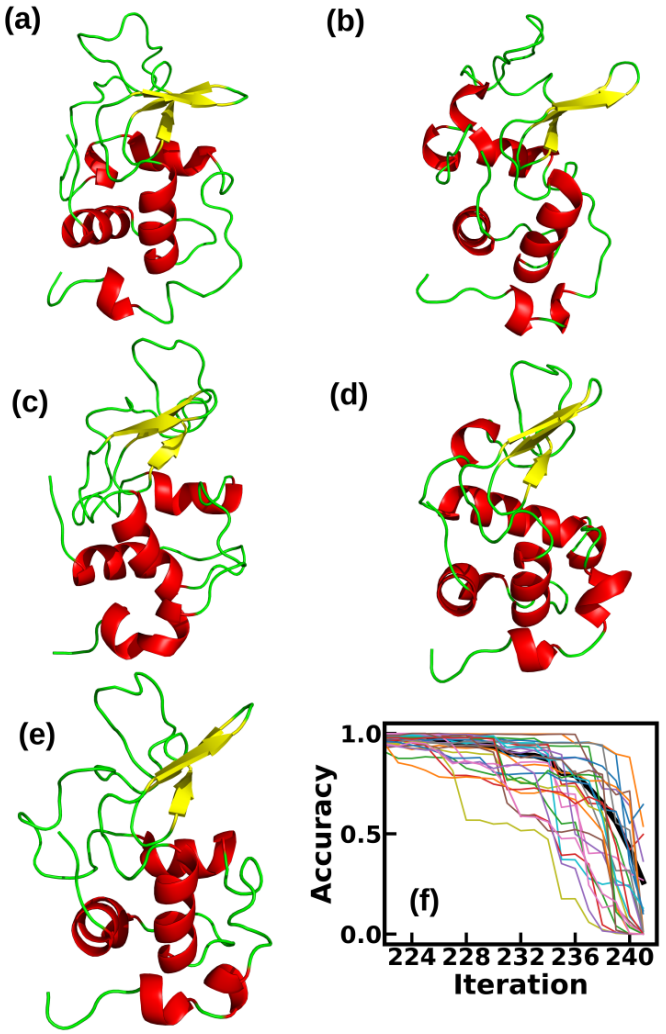

Figure 3.7: Representative secondary structures of the first six highly populated metastable states of aLA in the MG state. Structures are arranged as per decreasing population. Conformation (a) corresponds to the metastable states having the highest populations. Conformations (b–e) correspond to the other 4 populated metastable states, with population decreasing gradually from (b) to (e). (f) Accuracy loss plot of the XGBoost classifier. The figure is shown as a function of the number of discarded coordinates. The accuracy of all metastable states drops drastically upon removing the last 10 coordinates.

3.3.3 Location of the ECs

Given the meta-stable states, we use the XGBoost model (see in the Method section) on MD data to identify the relative importance of the dihedral angles in the MG state. Fig.3.7(f) shows an accuracy plot of the XGBoost algorithm. The thick black line represents state averaged accuracy. The accuracy remains constant upon discarding up to 232 coordinates. Accuracy decreases sharply for most of the states by discarding the last 10 coordinates. These coordinates, shown in Table 4 as per decreasing importance along with the secondary structure element in the crystal structure, are the essential coordinates (EC). The most essential coordinate is $\psi 80$ as it is obtained at the final iteration of the XGBoost, removing all other coordinates. The ECs are shown in Table 3.4. We also perform XGBoost analysis over a randomly selected window to check how much EC changes. We keep P_{min}=30 as earlier and obtain 31 microstates after density based clustering.

Finally, we acquire 15 metastable conformational states using a lag time of τ=40 picoseconds and a minimum meta-stability of Q_{min}=0.92. Obtained ECs are tabulated in Table 3.4 as per decreasing importance. We find $\psi 80$, $\phi 79$ and $\phi 17$ as EC like in the earlier case. This also confirms the convergence of the metastable state over our simulated trajectory. Our analysis suggests that the region near the calcium binding loop plays an essential role in molten globule formation.

We highlight the residues having essential coordinates over the crystal structure (Fig.3.8(a)) in red color. We find that only $\phi 122$ and $\phi 114$ belongs to loop residues, while all others belong to helix residues. The essential coordinate $\psi 80$ belongs to amino acid residue PHE80, which belongs to the helix near the calcium binding loop. Similarly, the residue LYS79 having the essential coordinate $\phi 79$ directly coordinates to calcium ion. We find in Fig.3.8(b) that most residues having the essential coordinates prefer unstructured or bend conformation, suggesting that these residues lack well-defined secondary structural elements in the MG state.

We compare the structural and dynamic features of the ECs to those of the non-ECs. We show the $\phi - \psi$ correlation plot for residue PHE80 in Fig.3.8(c). Comparing the Ramachandran plot, we find that the conformations in PHE80 lie within the helix, sheet, and unstructured region whereas secondary element in crystal structure was in helix region. This indicates that the residue undergoes fluctuations in the structural element. This is consistent with the data in Fig.3.8(b). On the other hand, the $\phi - \psi$ correlation plot for VAL8 which is assigned as non-essential coordinate, suggests conformations only confined within the helix region (Fig.3.8(d)) as it was in crystal structure. This indicates there is no change in the structural element throughout the simulation.

We characterize essential and non-essential coordinates in terms of S_P. We consider two data sets, one of the residues having the ECs and the other consisting of the remaining residues. The histograms of the structural persistence ($H(S_P)$) for these sets are shown in Fig.3.8(e). We find that residues containing EC coordinate posses low S_P, whereas non-EC coordinates have a higher value of S_P.

We further compute the time dependent correlation functions($C^{\theta}_{dihedral}(\Delta t)$) (TDCF) of dihedral fluctuations, using the method discussed in Ref.[130] Fig.3.8(f) shows dihedral auto-correlation functions of the first 5 essential coordinates. The TDCF is normalized by $\Delta t = 0$ value. We fit the initial decay data with an exponential form $C^{\theta}_{dihedral}(\Delta t) = A \exp(-\Delta t/\tau^{\theta}_{dihedral})$. $\tau^{\theta}_{dihedral}$ is the decay timescale. The histogram of the correlation decay timescales ($H(\tau^{\theta}_{dihedral})$) for the EC and non-EC dihedrals are shown in the inset of Fig.3.8(f). We observe that

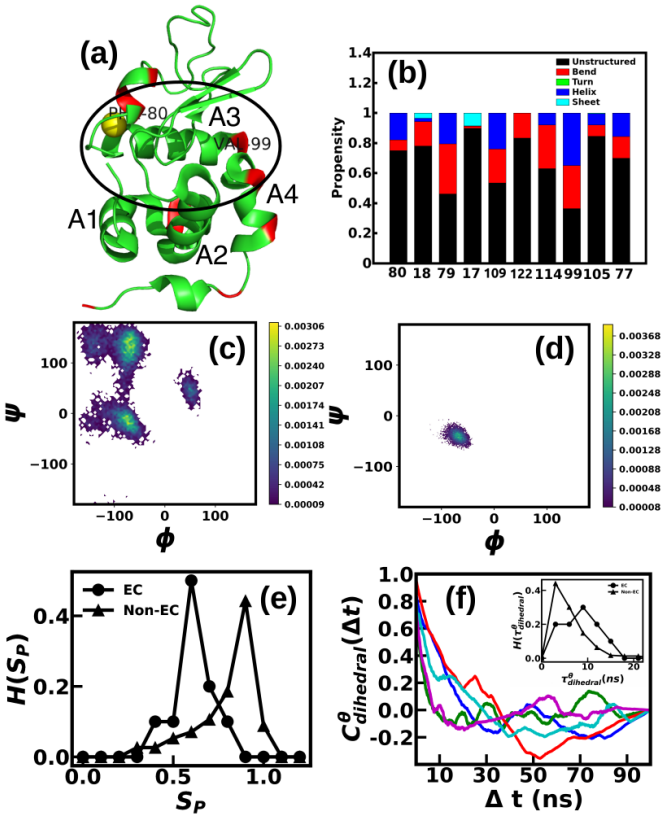

Figure 3.8: (a) Crystal structure of α-Lactalbumin showing the essential in red. Putative binding sites are present within the black circle. (b) Conformational preference of residues having essential coordinates. in ϕ-ψ correlation plots (c) Phenylalanine80 and (d)Valine8. (e) Histogram of structural persistence (S_P) for residues containing essential coordinates and non-essential coordinates. The lines are guides to the eyes. (f) Dihedral auto correlation function of first 5 essential coordinates. The inset shows the histogram of correlation timescales of essential and non-essential coordinates.

ECs have higher value of correlation timescales compared to non-EC. Average correlation timescales obtained using window averaging, $\langle \tau^{\theta}_{dihedral} \rangle$ for ECs is 3.13$\pm$0.76 ns and for non-ECs 2.45$\pm$0.21 ns, suggesting that the fluctuations in the EC are longer lived than those in non-EC degrees of freedom.

Window1			Window2		
EC	**Residue Name**	**Structure**	**EC**	**Residue Name**	**Structure**
ψ80	Phenylalanine(PHE)	Helix	ψ20	Glycine(GLY)	Helix
ψ18	Tyrosine(TYR)	Helix	ϕ100	Glycine(GLY)	Helix
ϕ79	Lysine(LYS)	Helix	ψ80	Phenylalanine(PHE)	Helix
ϕ17	Glycine(GLY)	Helix	ϕ20	Glycine(GLY)	Helix
ψ109	Alanine(ALA)	Helix	ϕ79	Lysine(LYS)	Helix
ϕ122	Lysine(LYS)	Loop	ϕ17	Glycine(GLY)	Helix
ϕ114	Lysine(LYS)	Loop	ϕ87	Aspartate(ASP)	Helix
ψ99	Valine(VAL)	Helix	ψ82	Aspartate(ASP)	Helix
ψ105	Leucine(LEU)	Helix	ψ19	Glycine(GLY)	Helix
ϕ77	Cysteine(CYS)	Helix	ϕ91	Cysteine(CYS)	Helix

Table 3.4: Residues which possess ECs pH=2. Secondary structure for each residue in initial crystal structure is mentioned.

Apo, CpHMD, Neutral			Holo, Normal MD, Neutral		
EC	**Residue Name**	**Structure**	**EC**	**Residue Name**	**Structure**
ϕ18	Tyrosine(TYR)	Helix	ϕ68	Histidine(HIS)	Loop
ψ68	Histidine(HIS)	Loop	ψ121	Glutamic Acid(GLU)	Loop
ψ122	Lysine(LYS)	Loop	ψ120	Cysteine(CYS)	Loop
ϕ19	Glycine(GLY)	Helix	ψ16	Lysine(LYS)	Helix
ϕ45	Aspargine(ASN)	Loop	ϕ63	Aspartate(ASP)	Loop
ϕ121	Glutamic Acid(GLU)	Loop	ψ67	Proline(PRO)	Loop
ϕ17	Glycine(GLY)	Helix	ψ66	Aspargine(ASN)	Loop
ϕ20	Glycine(GLY)	Loop	ϕ121	Glutamic Acid(GLU)	Loop
ψ18	Tyrosine(TYR)	Helix	ψ68	Histidine(HIS)	Loop
ψ16	Lysine(LYS)	Helix	ψ104	Tryptophan(TRP)	Helix

Table 3.5: EC obtained using XGBoost method at neutral condition using both CpHMD and Normal MD.

3.3.4 Comparison to IDP

The IDPs are known to sample conformational space by stochastically switching to different states separated by large energy barriers. This is reflected in terms of meta-stable states observed in the aggregation prone IDP like hIAPP.[91] Likewise, we find metastable states, separated by an energy barrier, in the MG state of aLA (Fig.3.5). We find that ECs are initially belong to a stable secondary structure, suggesting enhanced conformational fluctuations. This nature of the protein at MG state is similar to IDP. However, it is important to keep in mind that the MG state is induced by external effects like lowering of the solution pH and, hence, can be viewed as an induced disordered protein.

Experimental data suggest that the MG in aLA is realized in the absence of the Ca^{2+} ion.[98] This is supported by our data as well. We perform a CpHMD simulation with apo-aLA in neutral solvent conditions. Fig.3.9(a) shows dihedral PCs obtained in this case. High meta-stability is present where many of the ECs belong to stable structures in the initial crystal structure (Table3.5) as in the MG state. We contrast the conformation fluctuations in holo-aLA in the neutral solvent conditions. The free energy profiles obtained from the dPCA+ analysis (Fig.3.9(e)) show the presence of meta-stable states. However, in contrast to apo-aLA, the essential dihedral angles in the neutral case for the holo-protein mostly lie on the loop region in the crystal structure (Table3.5) and prefer unstructured conformation over the trajectory (Fig.3.9(h)). Thus, the loss of Ca^{2+} ion is the primary factor to realize the MG state of aLA.

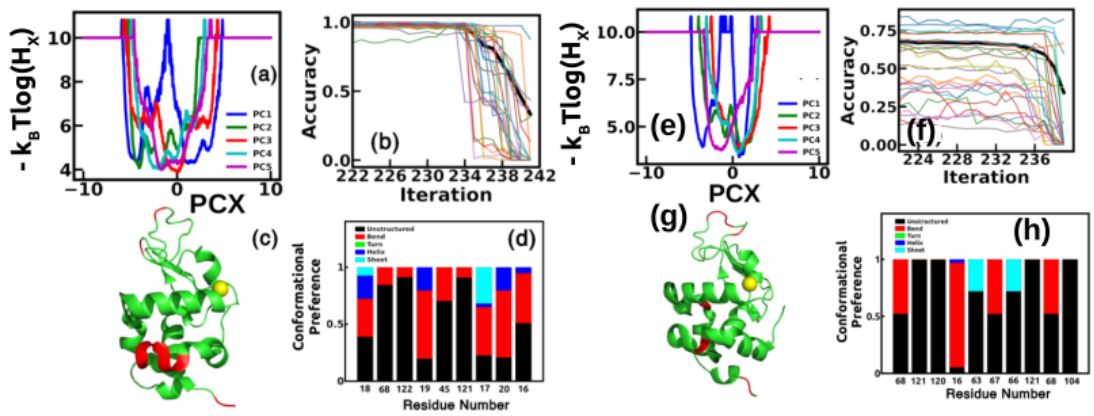

Figure 3.9: Identification of essential coordinate for apo (Ca^{2+} ion is shown for better understanding) protein applying constant pH simulation at pH7.(a) Principal component obtained from dPCA+ analysis. PCs 1-5 are shown in figure. (b) Accuracy loss plot of XGBoost classifier. The figure is shown as a function of number of discarded coordinate. Accuracy of all metastable states drops drastically upon removing of mostly last 10 coordinates., (c) Residues having essential coordinates are marked in initial crystal structure. They are colored in red. (d) Conformational preference of those residues having essential coordinates. Similar analysis for Identification of essential coordinate for holo protein using unbiased molecular dynamics simulation at neutral pH. (e) PCs 1-5 obtained using dPCA+, (f) Accuracy loss plot of XGBoost classifier, (g) Residues having essential coordinates are marked in initial crystal structure. All non-essential residues belong to loop region. (h) Conformational preference of those residues having essential coordinates.

3.3.5 Implication for functionality

It is known that in the MG state, aLA binds to fatty acids, like oleic acid (OLA) with negatively charged carboxylate (COO^-) head groups and long hydrophobic tail to form cytotoxic complex.[28,30,131] Experiments[28] suggest that the binding sites for OLA binding lies between A1 and A2 helices (Fig. 3.8(a)) and the cleft region. Table.3.6 shows putative binding residues of OLA with nature. Among the ECs, VAL99 of the cleft region and GLY17 of the A1 helix are putative binding residues. Thus, some residues directly participating in the OLA binding bear the essential coordinates. We correlate the fluctuations of the ECs in Ca^{2+} binding loop to the SASA of the OLA binding residues. As ligand binding is mainly due to hydrophobic interaction, we consider hydrophobic residues of cleft region. Fig.3.10(a) shows a correlation plot between SASA of ILE89 and dihedral ψ fluctuations of PHE80 (Fig.3.10(a)). The correlation plot shows different clusters corresponding to different structural elements of PHE80. We also check the correlation plot between SASA of ILE89 and dihedral ϕ fluctuations

Residue	Nature	Residue	Nature
VAL8	Hydrophobic	PHE53	Hydrophobic
PHE9	Hydrophobic	ILE55	Hydrophobic
ARG10	Basic	LYS58	Basic
LEU12	Hydrophobic	ILE59	Hydrophobic
LYS13	Basic	TRP60	Hydrophobic
LYS16	Basic	ILE89	Hydrophobic
GLY17	Hydrophobic	MET90	Hydrophobic
GLY19	Hydrophobic	VAL92	Hydrophobic
GLY20	Hydrophobic	LYS93	Basic
VAL21	Hydrophobic	LYS94	Basic
LEU23	Hydrophobic	ILE95	Hydrophobic
TRP26	Hydrophobic	LEU96	Hydrophobic
VAL27	Hydrophobic	LYS98	Basic
PHE31	Hydrophobic	VAL99	Hydrophobic
HIS32	Basic	GLY100	Hydrophobic
GLY51	Hydrophobic	ILE101	Hydrophobic
LEU52	Hydrophobic	TRP104	Hydrophobic

Table 3.6: Putative binding sites of Oleic acid (OLA) with nature.

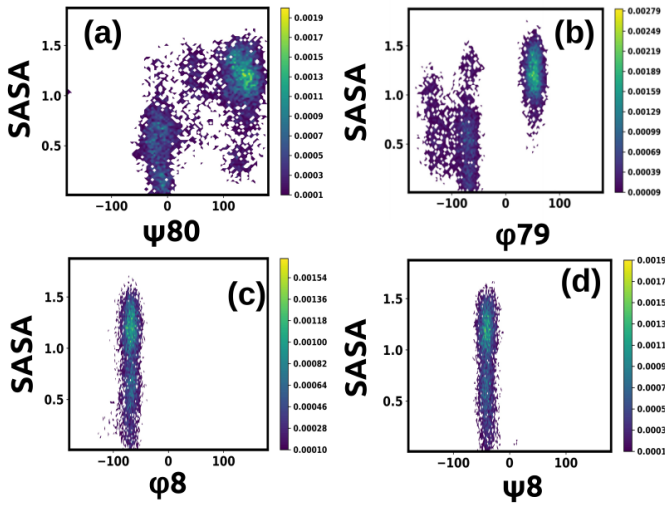

Figure 3.10: Correlation plot between SASA value of Isoleucine89 with (a) dihedral ψ fluctuations of Phenylalanine80, (b) dihedral ϕ fluctuations of Lysine79, (c) ϕ of Valine8 and (d) ψ of Valine8.

of LYS79(Fig.3.10(b)). Correlation plot shows different clusters in two different regions of conformational space which is similar as PHE80. It may be noted that PHE80 and LYS79 are distant from ILE89, suggesting allostery[38,132,133] between these residues. The correlation plots for other ECs and putative binding residue shows different clusters as the conformations of EC changes. We contrast the SASA of ILE89 with dihedral ϕ (Fig.3.10(c)) and ψ (Fig.3.10(d)) fluctuations of VAL8, which has only non-essential coordinates. The correlation plot shows

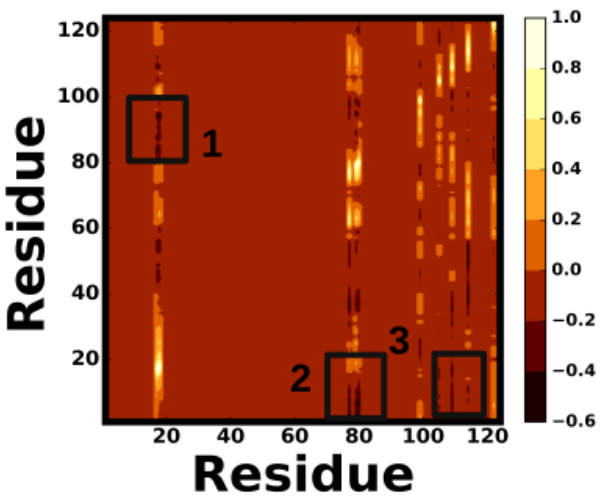

Figure 3.11: DCCM map between residues having essential coordinates with all other residues. Box 1 represents the DCCM map between GLY17 and TYR18 with putative binding residues of the interfacial cleft, box 2 represents the DCCM map between LYS79 and PHE80 with putative binding residues of the A1 helix, and box 3 represents the DCCM map between LEU105, ALA109, and LYS114 with residues of A1 and A2 helices

that the SASA fluctuations are not correlated to those of the dihedrals. Thus, the fluctuations in Ca^{2+} binding loop residues help the OLA binding residues to get exposed.

We further elucidate the connection between putative binding residues and the EC via the dynamical cross correlation (Eq.3.5) map of C_α atom fluctuations belonging to these residues (Fig.3.11). DCCM plot shows that GLY17 and TYR18 are anti-correlated with putative binding residues of the inter-facial cleft (marked as 1). Similarly, LYS79 and PHE80 are dynamically anti-correlated with residues of the A1 helix (marked as 2). Residues LEU105, ALA109 and LYS114 show dynamic anti-correlation with residues of A1 and A2 helices (marked as 3). Such anti-correlated motion between residues results in the opening of the inter-facial cleft required for OLA binding to aLA in an acidic medium.

3.4 Conclusions

In conclusion, we explore the conformation fluctuations and meta-stability in the MG state of aLA in terms of essential coordinates using density based clustering and a machine learning approach. We find metastability in the free energy landscape of apo-aLA at MG state. ECs responsible for metastability in the MG state prefer unstructured or bend conformations, although in crystal structure they possess a stable secondary structure. Residues participating in the coordination of Ca^{2+} ion (PHE80 and LYS79) act as essential coordinates of the system. Thus, the removal of Ca^{2+} initiates metastability in the protein upon lowering of pH. The ECs play a major role in opening up the putative fatty acid

binding sites. These features in the MG state are similar to those of IDP. It will be worthwhile to understand the binding of fatty acids in such a disordered state of protein. Our study will be helpful to understand the functionality of a protein in partly denatured conditions, as in the MG state.

Appendix

A1. Dihedral principal component analysis

Protein is well described by the backbone dihedral angles ϕ and ψ. Hence, we use principal component analysis on dihedral angles based on newly developed dPCA+ method by Sittel et al.[23] Traditional principal component analysis sometimes produces errors in the computation of the covariance matrix and the projection of circular coordinates on the eigenvector. The dPCA+ method minimizes these errors by shifting the dihedrals periodically to set the maximal sampling gap at the periodic boundary. This method primarily makes use of the well-known fact that dihedral angles, due to steric hindrance, do not cover the entire space of dihedral space $[-\pi : \pi]$ but are bounded in a specific region. One can obtain free energy landscape by applying the dPCA+ method to MD data of a protein. Based on the shape of a one dimensional projection of the free energy landscape, further important principal components can be chosen for analysis. In our analysis, we exclude the terminal residues.

We also analyze the PCs based on time evolution of auto correlation functions which is defined as,

$$C_X^{PC}(\Delta t) = \frac{\langle \delta V_i(t + \Delta t)\delta V_i(0)\rangle}{\delta V_i^2} \tag{3.6}$$

with $\delta V_i(t) = V_i(t) - \langle V_i \rangle$, X-denotes PC number.

A2. Density based clustering

We perform a robust density based geometrical cluster analysis[22] over a landscape in the hyperspace spanned by the dihedral PCs. For every structure in the trajectory, we count the number of frames within a fixed radius R from the given frame inside the hypersphere. Normalization of the count gives the density of sampling probability P. Free energy is estimated using the equation $\Delta G = -k_B T ln P$. To begin, an energy cut off value at relatively low free energy (F < 0.1 $k_B T$) is defined. All structures below the cut-off are considered, and the others are ignored. Selected frames that are closer than a certain lumping radius (d_{lump}) are assigned to the same cluster. The energy cut-off is increased gradually at a step of 0.1 $k_B T$ until all clusters are converged at the energy barrier. In this way, all structures get specific cluster membership. Nagel et al.[25] shows that in most cases, d_{lump} itself is a good choice for clustering radius R. Here, we consider

both R and d_{lump} equal to 0.521. At the end of density based clustering, when all data points are considered, a minimal population (P_{min}) of each state is given as percentage of all data points. This prevents from considering various small microstates within the same minimum, which may form due to local free energy fluctuations. Because P_{min} affects the number of microstates, it should be chosen based on the desired coarse graining level.

A3. Dynamical clustering

Density based geometrical clustering is expected to give a valid description of the primary description of the system if the conformational states of different geometrical states are separated by large barriers. But, sometimes small structural changes are observed. Even in the case of a rare transition between states, geometrically different but dynamically close states are artificially separated. On the other hand, in the case of low sampling, dynamically distinct states are considered as geometrically close and thus may be allotted wrongly. This error can be minimized by considering a dynamic clustering method that combines MD frames that are near in time evolution instead of geometrical evolution. Here, we use the most probable path (MPP) algorithm developed by Jain et al.[20,21]for dynamical clustering. At first, given a set of microstates, the transition matrix of these states is calculated. For a given state, if self transition probability is lower than certain metastability criterion $Q_{min} \in (0,1]$, then the state will be merged with the state having the highest transition probability and lower free energy. The process is reiterated until for a given Q_{min} no further transitions happen. Here, we choose $Q_{min} = 0.92$ for further calculations.

A4. Essential internal coordinates (EC)

We have used supervised machine learning techniques to find essential coordinates for meta-stability in the MG state. We identify an essential coordinate of the system in the meta-stable state using the extreme gradient boosting (XGBoost) algorithm.[134] In this algorithm, given a trajectory of MD coordinates in terms of dihedral angle and a meta-stable states obtained from clustering, a machine learning model is constructed by minimizing a loss function. The overall accuracy of the model can be estimated by dividing the available data into train and test sets. The importance of a coordinate is given by the gain of the loss function value. Any MD coordinate which has a higher gain in loss function is more important for characterizing the state than others. Given a trained model, all the

dihedrals are sorted as per their importance. When a non-essential dihedral is discarded, the accuracy of the model does not change significantly. Thus, one can identify essential internal coordinates that are necessary to discriminate the states explicitly. Here, all XGBoost parameters are chosen as in Ref.[8]

Chapter 4 Fatty acid binding with α-lactalbumin in MG state

4.1 Introduction

Previous chapter shows that α-lactalbumin (aLA) protein found in milk, has the ability to adopt the MG state (MG-aLA) in acidic pH condition.[15,99–103] The conformation fluctuations in MG-aLA are similar to those in the Intrinsically disordered proteins (IDP). Although it is known that the inherent dynamic nature of IDPs play a key role in rapid ligand recognition,[92,135,136] nothing analogous is known about ligand recognition by proteins in the MG state.

It is observed that bovine Holo-aLA is unable to bind oleic acid (OLA). But, MG-aLA binds to OLA with binding free energy -9.45 kcal/mol.[26] This MG-aLA-OLA complex, popularly known as XAMLET, X species (stands for human, bovine, goat or other species) α-lactalbumin Made lethal to tumor cells, performs cyto-toxic activities. aLA in MG state acts as carrier of the cyto-toxic factor OLA.[104–106] It has been proposed that OLA stabilizes the MG state in XAMLET.[137] Fatty acid like OLA is amphipathic in nature, i.e., they have negatively charged carboxylate (COO^-) head-groups and long hydrophobic tail. Experimental studies[28] suggest that the putative binding site of OLA lie between the A1 and A2 helices and the inter-facial cleft (See, chapter 3, Fig.3.1). However, no experimentally probed structure of the XAMLET complex has been reported so far.

Here, we explore binding of bovine MG-aLA with OLA using the all atom molecular dynamics (MD) simulations. Since there is no experimental data available on the binding mode, first we determine the binding residues of MG-aLA to OLA with the help of thermodynamic cost of conformation changes.

Recent studies show that the free energy and entropy costs associated with conformation changes can be calculated based on the dihedral fluctuations of a protein over simulated trajectories in different conformations.[7,30,138] Conformationally destabilized residues with increasing free energy and disordered residues with increasing entropy in a given state with respect to a reference state are shown to participate in binding events in the given conformation state.[30] We simulate the protein at neutral pH and pH=2 using the constant pH molecular dynamics (CpHMD) technique,[111] and find out the conformationally destabilized and disordered residues in MG-aLA with respect to the Holo-aLA. We use these residues as bias to dock OLA to a representative conformation of the MG-aLA over the simulated trajectory. We further carry out MD simulations on the docked complex to generate the conformations of the MG-aLA-OLA complex and compute the conformational thermodynamics with respect to MG-aLA from the dihedral angle distributions. We further estimate the binding free energy using the umbrella sampling (US) method[29] for MG-aLA-OLA using the separation between the centre of mass of the protein and the ligand as reaction coordinate. We also explore the kinetics where we study conformational fluctuations for different separations between the protein and the ligand binding using the steered molecular dynamics and the machine learning techniques.[8]

Our conformation thermodynamics data show that residues near the cleft region are prone to binding, in agreement with the earlier study.[28] Further, OLA binding to the MG-aLA in the MG-aLA-OLA complex gives conformational stability and order with respect to MG-aLA. The binding energy is comparable with earlier experimental results[26] in MG state. The conformational fluctuations show that the most essential coordinates(EC) in the conformation fluctuations[31] belong to the ligand binding residue LEU52, LYS94, ILE95, GLY100, ILE101. As the ligand goes further away, the EC is shifted towards the residues of Ca^{2+} binding region, suggesting that the Ca^{2+} binding residues have allosteric control on the ligand binding as suggested in earlier works.[30,31] The average water number around the residues having the essential coordinates increases as the ligand goes further away. Thus, water molecules around the essential coordinates are perturbed in complex formation, suggesting their key role in OLA binding at MG state. We also check the dynamics of water molecules near the protein surface. It shows that the water molecules close to both protein and ligand have sub diffusive behavior.

4.2 Methods & analysis

4.2.1 Constant pH molecular dynamics

The protein considered in this study is Bovine-α-lactalbumin in both holo (with Ca^{2+} ion, RCSB PDB ID: 1F6R) and apo (without Ca^{2+} ion, RCSB PDB ID: 1F6S)[113] form as discussed in previous chapter. For constant pH molecular dynamics (CpHMD) we follow similar protocol as discussed in chapter 3.

4.2.2 Conformational thermodynamics

The free energy and entropy associated with conformational changes are computed from the histograms of the dihedral angles as reported earlier.[7] The conformational free energy and entropy cost for an individual protein residue can be estimated by taking the sum of contribution of free energy and entropy of each dihedral in that residue. We compute this with our in house tools, which can be obtained from GitHub (https://github.com/snbsoftmatter/confthermo). Residues having ΔG,$T\Delta S > 0.0$ are identified as destabilized and disordered residues. These residues are active in ligand binding. This yields the binding mode between the ligand and the protein.

4.2.3 Protein-ligand complex preparation

Clustering & docking

As no crystal structure is available for MG-aLA, we use K-means clustering[117,139] to identify representative structure over the trajectory of CpHMD run at pH=2.0. In molecular simulation, clustering represents the grouping of similar conformations together. Similarity is determined by a distance metric, where the smaller distance represents more similar structures. Here, coordinate root-mean-square deviation (RMSD) is used as distant metric. Clustering is computed using CPPTRAJ tools of AMBER.[117] Next, we dock protein structure obtained from clustering of MG-aLA using the HADDOCK.[140,141] We dock the ligand using conformationally destabilized and disordered residues of cleft region as bias. The docking protocol consists of following stages, at first a rigid body energy minimization and then MD base refinement process provides model structures. Basis on Z score, therefore, the top cluster is chosen. Details of docking are in Table 4.1. The OLA molecule is taken from the crystal structure of liver fatty

acid binding protein–oleate complex (PDB ID: 1LFO). We use similar protocol to construct the model for Holo-aLA-OLA complex.

HADDOCK score	-34.4 +/- 2.4
Cluster size	19
RMSD from the overall lowest-energy structure	1.9 +/- 0.0
Van der Waals energy	-21.8 +/- 1.1
Electrostatic energy	-93.6 +/- 1.4
Desolvation energy	-6.7 +/- 0.2
Restraints violation energy	34.9 +/- 19.2
Buried Surface Area	613.0 +/- 11.9
Z-Score	-2.0

Table 4.1: Docking study on MG-aLA-OLA system.

MD simulations of protein-ligand complex

The MD simulations (See chapter 2, Appendix A1) of the docked complex are performed using GROMACS[76] simulation package with Amber99Sb force field[75] and TIP3P[77] water model. The GROMACS parameter for OLA are generated using Antechamber and Acpype. Antechamber parametrizes the molecule using general AMBER force field (GAFF).[142] Acpype is python interface to antechamber, which provide GROMACS topologies.[143]

AMBER CpHMD method provides fraction of time titrable residues protonated during the simulation. Table.3.1 of the previous chapter shows fraction of time titrable residues remain protonated during simulations. Based on this, we set protonation state of titrable residues in GROMACS.

We further perform a total of 2 sets of simulations : (i) MG-aLA-OLA and (ii) Holo-aLA-OLA complexes. At first, systems are immersed in cubic box of dimension 7.221x7.221x7.221 nm^3 and 7.725x7.725x7.725 nm^3 respectively. 11 Cl^- ion is required for protein-ligand complex at MG state. Holo-aLA required 6 Na^+ ion to neutralize the complex structure. Minimization is done for 50,000 steps using the steepest descent algorithms. All bonds in the protein are constrained using the LINCS algorithm. Equations of motion are integrated using leap-frog algorithm with an integration time step of 2fs. Systems are equilibrated through 2 steps (NVT and NPT) using position restraints to heavy atoms. The NVT and NPT equilibration is carried out at 300K Temperature and 1 Bar pressure. Afterward the full all-atom simulations are performed for 1 μs with 2 femtosecond time step integration employing periodic boundary conditions in all directions.

To address the dynamics of hydration water, we save the trajectory at 0.1 ps resolution for post-processing analysis. Dynamical quantities are calculated by separating 100 ns trajectory after equilibration, into 10 blocks, with a 1 ns duration for every block and maintaining a gap of 10 ns between two consecutive blocks.

4.2.4 Steered MD & Umbrella sampling(US) simulation

The equilibrated structure of MG-aLA-OLA complex is selected for US study. The initial system is prepared using GROMACS software and protein-OLA complex at MG state is made parallel to Y axis. The dimension of the box is 6.53x11.95x4.34 nm^3. The prepared box is solvated, neutralized, minimized following previous simulation protocol. The box is further equilibrated at a specific temperature (NVT) and specific pressure (NPT) similar to the previous MD simulation setup. The US method initiates with steered molecular dynamics(SMD) method using distance between center of mass as reaction coordinates. The OLA molecule is pulled from binding residues of protein towards the bulk solvent for 500 picosecond (ps) with 1000 kJ/mol-nm force. OLA molecule is pulled at the rate of 0.01 nm per ps. Snapshots are saved at each picosecond during the course of pulling, hence total 500 configurations are generated. We extract 22 required frames from the SMD trajectory to prepare the umbrella sampling windows, where the distance between each configuration is 0.2 nm. These configurations serve as initial structure of US simulation and each frame is independently equilibrated, performing NPT equilibration for 100 ps. Next, MD is performed for each individual configurations. The potential of mean force (PMF) is calculated from equilibrated trajectory of US simulation using weighted histogram analysis method (WHAM),[144] included in GROMACS. One can calculate binding energy from PMF curve by subtracting the PMF value at the position of ligand at the maximum distance from the PMF value at the position of ligand at minimum distance from the protein. We also run separate umbrella sampling where complex structure obtained from equilibrium MD trajectory is used as initial configuration. In total, we performed 4.3 μs of US simulation. The details of the method of steered molecular dynamics, umbrella sampling and WHAM is in Appendix A1.

4.2.5 Identification of essential coordinates (EC)

We identify essential coordinate of the system for various position of OLA molecule from aLA at MG state. For this, we use SMD techniques as discussed

earlier, where distance between center of mass between protein and ligand is used as reaction coordinate. Essential coordinate is identified using the method described by Brandt et al.[8] The method is used earlier to identify essential coordinates in the MG state of aLA (MG-aLA) in Abhik et al.[31] The details of the method is in Appendix A1-A4 of chapter 3.

4.2.6 Dynamical cross-correlation analysis

The method of calculating dynamical cross correlation function, $C(i,j)$ between various C_α atom of the protein is already discussed in previous chapter.

4.2.7 Radial distribution function

We calculate radial distribution function (rdf), g(r) to understand the arrangement of water around protein surface. The rdf is calculated using the following equations:

$$g(r) = \langle \frac{1}{\rho_N} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(r_{ij} - r) \rangle \tag{4.1}$$

where denotes the total number of water molecules, ρ_N is the particle density and the angular bracket represents average over time origins.

4.2.8 Dynamical parameters of water

Residence time(τ_R)

The residence time of a water molecule is calculated from the survival time correlation function.[145,146] The survival time correlation function is defined as

$$C_S(t) = \frac{\langle P_i(t) P_i(0) \rangle}{\langle P_i(0) P_i(0) \rangle} \tag{4.2}$$

where, P_i=1 if the i-th water molecule continuously present within the solvation shell for a time t and 0 otherwise. The angular bracket signifies that the averaging is done over both time and water molecule. We fit the decay curve of $C_R(t)$ can be fitted with bi-exponential function of the form

$$C_S(t) = \sum_{i=1}^{2} A_i \exp\left(-\frac{t}{\tau_i}\right) \tag{4.3}$$

to calculate average residence time ($\langle\tau_R\rangle$) of water molecule where, τ_i corresponds to different time constants and A_i corresponds to their relative contributions respectively. $\langle\tau_R\rangle$ is related to $\langle\tau_i\rangle$ via the equation

$$\langle\tau_R\rangle = \sum_i A_i.\tau_i \tag{4.4}$$

Translation diffusion coefficient(D_E)

The mean square displacement (MSD) is defined as,

$$\langle\Delta r^2\rangle = \langle|r_i(t) - r_i(0)|^2\rangle \tag{4.5}$$

where, $r_i(t)$ and $r_i(0)$ represents the position vector of i-th water oxygen atom at time t and at time t=0. Here, angular brackets signify that the averaging is computed over all tagged water molecule at different time origins.

Rotational orientation

The rotational motion of water molecule can be probed through reorientation dynamics of water dipole, $\vec{\mu}$ (vector joining the central oxygen atom of water molecule and specific hydrogen atom attached with oxygen atom). We calculate the dipole-dipole reorientation time correlation function (TCF), $C_\mu(t)$, defined as

$$C_\mu(t) = \frac{\langle\hat{\mu}_i(t).\hat{\mu}_i(0)\rangle}{\langle\hat{\mu}_i(0).\hat{\mu}_i(0)\rangle} \tag{4.6}$$

Here $\hat{\mu}(t)$ is the i-th water molecule's unit dipole moment vector at time t. Here also averaging is done over all tagged water molecule at different time origin. Decay curve of $C_\mu(t)$ can be fitted with tri-exponential functions to calculate amplitude-weighted average reorientation time constant, ($\langle\tau_\mu\rangle$) as earlier.

4.3 Results & Discussions

4.3.1 MG-aLA-OLA complex

First, we consider the equilibrium aspects of OLA binding to MG-aLA. We identify conformationally destabilized and disordered residues of the cleft region in MG-aLA compared to the Holo-aLA. This set of residues consists of a large number of hydrophobic residues and a few basic residues. Fig.4.1(a) shows data

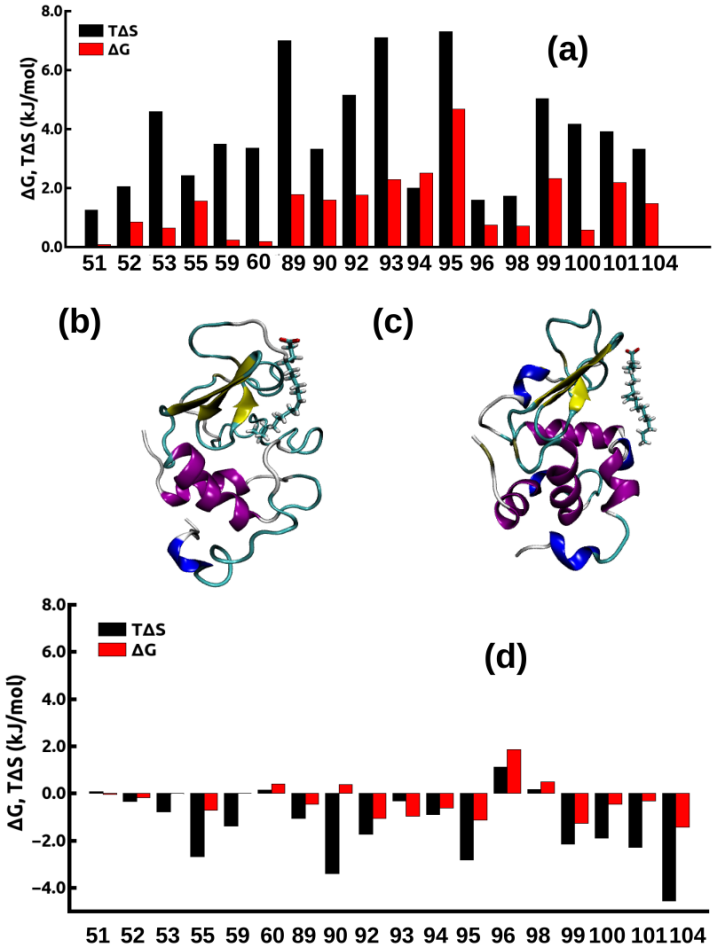

Figure 4.1: (a) Conformational thermodynamics change at molten globule state with respect to neutral state, (b) Equilibrated structure of MG-aLA-OLA complex. Hydrophobic tail of the OLA goes into the cleft region, (c) Equilibrated structure of Holo-aLA-OLA complex. OLA remain outside the protein cleft region all over the simulation, (d) Conformational thermodynamics change of active residues at MG-aLA-OLA conformations w.r.t MG-aLA conformations.

for changes in conformational thermodynamics, free energy (ΔG) and entropy ($T\Delta S$) of those residues. Here, the change in free energy and entropy consist of contributions due to the change in dihedral distributions of backbone dihedral angle ϕ, ψ and side chain dihedral angle $\chi_i, i = 1, .., 5$. We consider them as active residues to dock OLA to a representative structure of the protein in the MG state which is subject to MD simulations. An equilibrium structure from MD trajectory of MG-aLA-OLA complex is shown in Fig.4.1(b). The long hydrophobic tail of OLA extends towards the interfacial cleft. Fig.4.1(c) shows an equilibrium structure from MD trajectory of Holo-aLA-OLA. It shows that OLA does not bind into cleft region of Holo-aLA. It suggest that at neutral

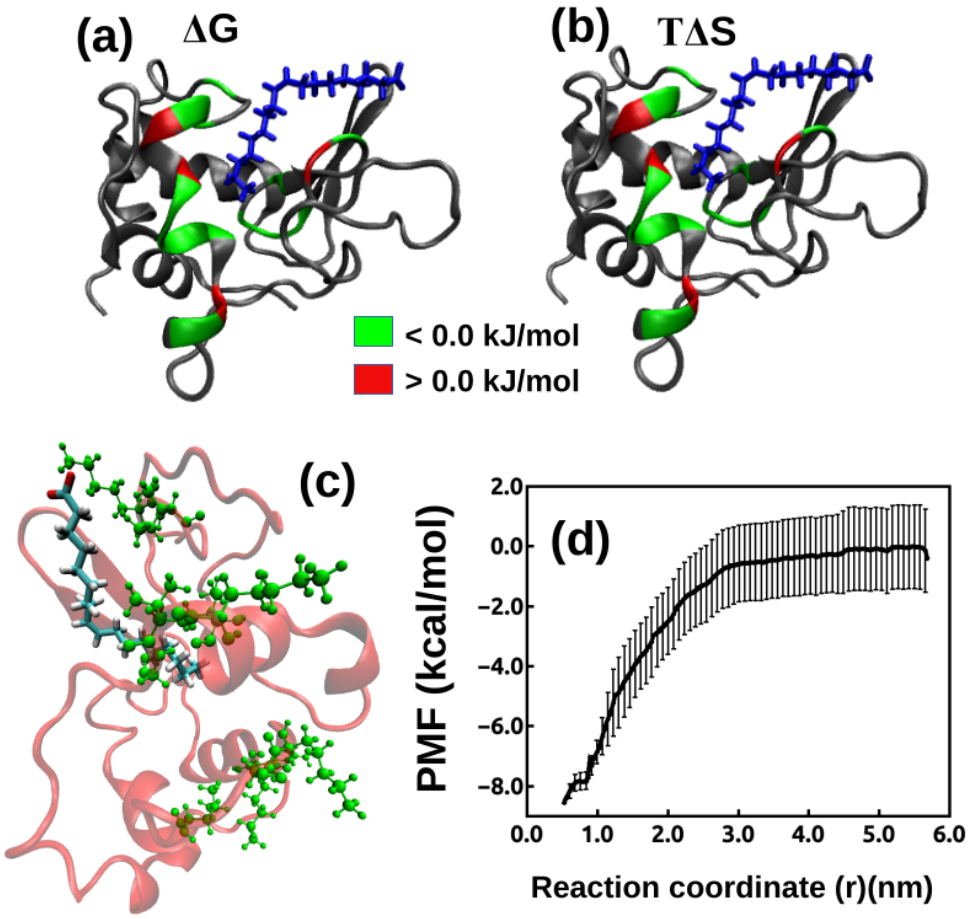

Figure 4.2: (a) ΔG (b) $T\Delta S$ of active residues in equilibrated structure. Green color shows stabilized/ordered residues, and red corresponds to destabilized/disordered residues. (c) Residues of aLA involved to form binding interface with OLA are marked in green, (d) PMF curve of protein-ligand complex obtained from umbrella sampling method for MG-aLA-OLA conformations

condition binding does not happen.

Next, we check the conformational free energy and entropy costs of the MG-aLA-OLA complex with respect to MG-aLA are shown in Fig.4.1(d). We show the conformational thermodynamics change ΔG and ΔS of those active residues for OLA binding by adding all dihedral contributions. We find that all binding residues have ΔG is negative except TRP60, MET90, LEU96, ILE98 . Those residues are marked in red color in the equilibrated structure of MG-aLA-OLA in Fig.4.2(a). Most of the active residue are ordered except TRP60, LEU96 and ILE98. Those residues are colored in red in Fig.4.2(b). The total change in ΔG and ΔS of the binding region of aLA at MG-aLA conformations are 26.20 kJ/mol and 68.84 kJ/mol. But due to presence of OLA binding at MG-aLA-OLA

Complex		**OLA at 1.34 nm**		**OLA at 1.93 nm**		**OLA at 2.5 nm**	
EC	Residue Name	EC	Residue Name	EC	Residue Name	EC	Residue Name
ϕ**95**	**ILE**	ϕ**63**	**ASP**	ϕ**39**	**GLN**	ψ**17**	**GLY**
ψ18	TYR	ϕ23	LEU	ϕ84	ASP	ψ13	LYS
ϕ100	GLY	ψ107	HIS	ψ54	GLN	ψ24	PRO
ϕ20	GLY	ψ13	LYS	ψ64	ASP	ψ81	LEU
ϕ19	GLY	ϕ88	ASP	ψ91	CYS	ψ40	ALA
ψ52	LEU	ϕ49	GLU	ψ39	GLN	ϕ24	PRO
ψ94	LYS	ϕ54	GLN	ϕ56	ASN	ϕ13	LYS
ψ101	ILE	ϕ24	PRO	ϕ64	ASP	ϕ87	ASP
ψ19	GLY	ϕ56	ASN	ψ53	PHE	ϕ25	GLU
ϕ94	LYS	ϕ53	PHE	ψ32	HIS	ψ55	ILE

Table 4.2: Essential coordinate for 4 different cases

conformations, the change in ΔG and ΔS of the binding region becomes -5.47 kJ/mol and -24.8 kJ/mol. It suggests that overall binding region of aLA become stabilized and ordered due to OLA binding.

We identify binding interface in MG-aLA-OLA conformations based on minimum distance between any two pair of atom (d_{min}). Here, we consider protein C_α atom and main chain carbon atom of OLA. Residues responsible to form binding interface are marked in green color over the protein in equilibrated structure in Fig.4.2(c). We find that active residues like ILE59, ILE95, LYS98, VAL99, GLY100 form binding interface along with some other residues like ARG10, LYS13, LYS16, GLY17, LEU23 of A1-A2 helices and LYS58 of interfacial cleft. It is to be noted that experimental study suggest that putative binding site of OLA lies between the A1 and A2 helices and the interfacial cleft.[28]

We calculate binding energy between protein-ligand using umbrella sampling techniques (Details are in Method section) where the reaction coordinate is the center-of-mass (COM) distance between the ligand and the protein. Fig.4.2(d) shows PMF curve of protein-ligand complex for MG-aLA-OLA. For, MG-aLA-OLA case, the binding energy of protein-ligand complex is approximately -8.3 kcal/mol, which is comparable with earlier experimental observations.[26] We calculate the error bar in PMF by averaging multiple PMF obtained from various independent umbrella sampling simulations.

Next, we check metastability for MG-aLA-OLA complex. Fig.4.3(a)-(c) shows free energy landscape (FEL) of principal components (PC1-3) obtained using dPCA+ method. FEL plot for MG-aLA-OLA conformations are marked in blue,

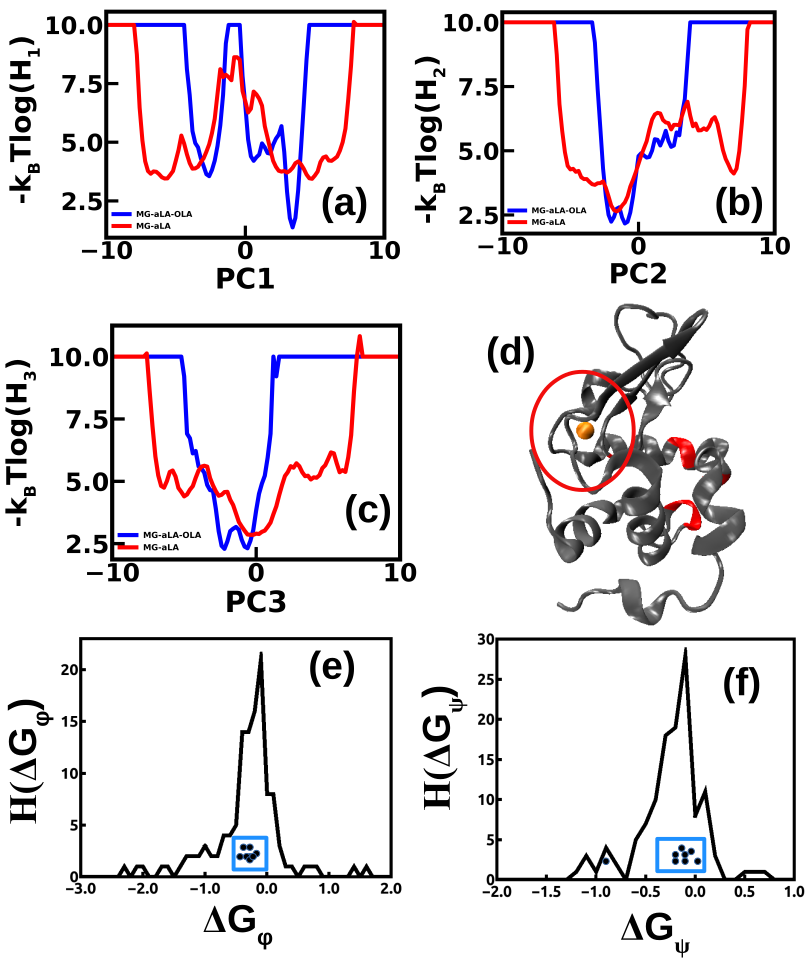

Figure 4.3: Free energy landscape obtained from dPCA+ along (a) PC1, (b) PC2, (c) PC3 for MG-aLA-OLA (blue) and MG-aLA (red). Y axis represents negative log of population of PCs (H), (d) Residues having essential coordinate in MG-aLA-OLA complex. Histogram of ΔG value for all residues considering (e) ϕ dihedral, (f) ψ dihedral angle. Value of ΔG for essential residues are marked inside the blue box.

whereas MG-aLA conformations are marked in red. We find that metastability is changed in complex formation. Along PC1-PC2, metastability is reduced for MG-aLA-OLA conformations as compared to MG-aLA conformations. Along, PC3 metastability completely vanished in MG-aLA-OLA complex, whereas for MG-aLA metastability still present. Reduction in metastability suggests that at MG-aLA-OLA conformations, protein has more stable state as compared to MG-aLA conformations which is consistent with the conformational thermodynamics data.

We identify essential coordinates (EC) of the system using clustering and supervised machine learning base analysis.[8] Residue having essential coordinates in protein are colored (red) over the crystal structure of the protein in Fig.4.3(d). Ca^{2+} binding region is marked within a red circle. ECs are also tabulated in

Table.4.2 where most EC which is obtained in the final iteration of XGBoost, removing all other coordinates, is marked in bold. In complex MG-aLA-OLA, ECs mostly belong to the putative binding site of OLA. $\phi 95$ acts as most essential coordinate, which belongs to the residue ILE95. It is to be noted that residues having EC like ILE95, LYS94, GLY100, ILE101, LEU52 also acts as active residue for ligand binding. It suggests that at complex state the dynamics of the system mostly governed by some of the binding residues.

Next, we discuss the free energy change of ΔG of ECs at MG-aLA-OLA conformations with respect to MG-aLA conformation based on conformational thermodynamics. Fig.4.3(e)-(f) shows histogram of free energy change for ϕ and ψ degree of freedoms ($H(\Delta G_\phi)$ and $H(\Delta G_\psi)$) of all residues. We check where the free energy change of residues having essential coordinate fall in the histogram. Fig.4.3(e) shows for ϕ degree of freedom, residues having ECs, ΔG values are clustered near 0 value. Data points are marked within a blue box. Similarly, for ψ degree of freedom (Fig.4.3(f)), most of the residues fall neighborhood zero value. It suggests that residues having EC clustered near the low energy change region.

The secondary structural section forming the binding region of protein are known to reveal correlated motions during the recognition process of ligand.[147] We compute to this end the dynamic cross correlation map C(i,j)[119,147] between protein residues, (see in the method section). Fig.4.4(a) shows binding region of the protein at MG state in color, where green color corresponds to hydrophobic and basic residues of A1-A2 region and red color represents hydrophobic and basic residues of cleft region. We compare DCCM map for two different cases: (i) MG-aLA (Fig.4.4(b)), and (ii) MG-aLA-OLA complex (Fig.4.4(c). In MG-aLA, the residues of ligand binding region are negatively correlated(Fig.4.4(b)). The region of interest is marked within the box. The region involved in anti-correlated motion in the protein segment primarily belongs to green and red region of the protein in Fig.4.4(a). Anti-correlated motions between residues of protein decrease in MG-aLA-OLA(Fig.4.4(c)). This suggests that the long ranged anti-correlated motion between protein segments are involved in ligand binding.

4.3.2 Kinetics of OLA binding to MG-aLA

The presence of the long-ranged correlations are also revealed from the fluctuation spectra for various separation between the protein and the ligand. We carry out all-atom MD simulations on various configurations with fixed the distance between centers of mass of the protein and the ligand obtained during the steered

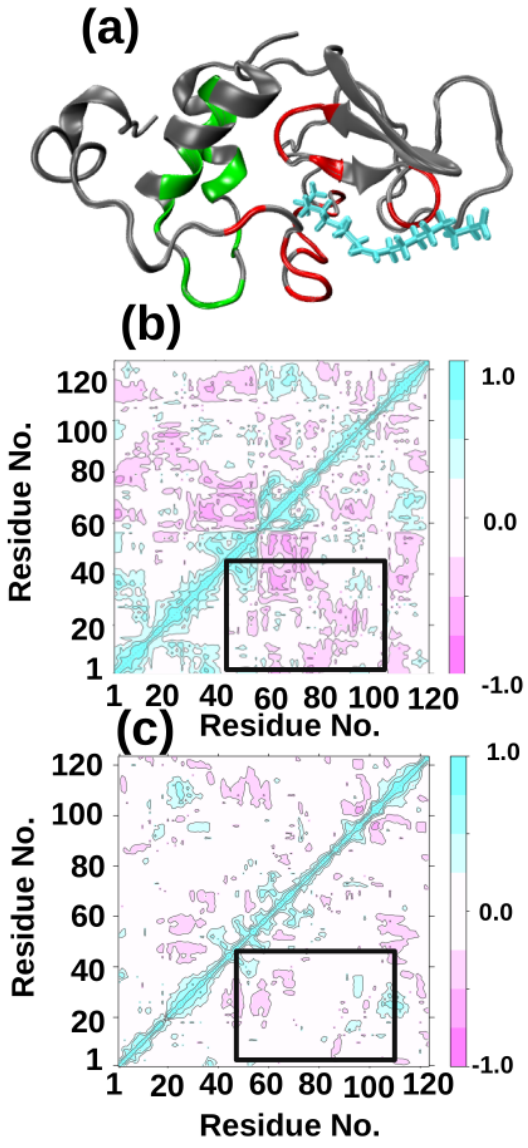

Figure 4.4: a) Binding region of the protein at MG state in color, where blue color corresponds to hydrophobic and basic residues of A1-A2 region and red color represents hydrophobic and basic residues of cleft region. Dynamic cross correlation map (DCCM) for two different cases, (b) MG-aLA and (c) MG-aLA-OLA complex.

molecular dynamics simulations. Residue having essential coordinates in protein are colored (red) over the crystal structure of the protein for 3 different cases in Fig.4.5(a)-(c). Ca^{2+} binding region is marked within a red circle. Table.4.2 shows that essential coordinates and corresponding amino acid name for different cases. Fig.4.3(d) already shows residues having ECs in the complex. When ligand is fixed at a distance of 1.34 nm, most EC shifted to ASP63 and corresponding degree of freedom is, $\phi 63$ which is near one of the binding residue TRP60. In these conformations, ECs are started to shift towards one of the residue of Ca^{2+} binding region (Fig.4.5(a)) i.e. ASP88 and corresponding EC is $\phi 88$. When ligand is at 1.93 nm (Fig.4.5(b)), most EC belongs to GLN39. In this conformation, ϕ

dihedral angle of ASP84 acts as one of the EC. Due to further pulling of ligand up to 2.5 nm, most EC shifted to GLY17 and residue LEU81 acts as one of the essential coordinates, Fig.4.5(c). It is noted that ASP82, ASP87 and ASP88 directly participate in Ca^{2+} coordination. We already discuss in previous chapter[31] that at MG-aLA conformations, EC belongs to $\psi 80$ and $\phi 79$. Thus, there is shifting of EC from ligand binding residues when ligand is in complex to Ca^{2+} binding loop region as ligand-protein distance increased. This suggests allostery between OLA binding and the Ca^{2+} binding loop residues.

We calculate average water number ($\langle N_w^R(d)\rangle$) around the R th residue, considering a 5 Å cutoff around the residues for separation d between COM of the protein and ligand. We consider, in particular $\langle N_w^{ILE95}(d)\rangle$, $\langle N_w^{LYS94}(d)\rangle$ and $\langle N_w^{TRP60}(d)\rangle$. It is important to point out that ILE95 contains EC $\phi 95$, LYS94 contains EC both $\phi 94$ and $\psi 94$. Fig.4.5(d) shows $\langle N_w^R(d)\rangle$ around residue ILE95 and LYS 94 for different position of ligand from protein. As ligand goes further from protein, $\langle N_w^R(d)\rangle$ increases around their residues. We fit the dataset for ILE95 and LYS94 with $y = x^\alpha$ to find the dependency of $\langle N_w^{Res}(d)\rangle$ with distance between ligand and protein. We find that the exponent α value for ILE95 is 0.71 and for LYS94 is 0.24. On the contrary, TRP60, which acts as one of the binding residue of protein but not as EC, does not show any change. Thus, the hydration level of the residues having ECs change markedly upon displacement of the ligand. In other words, when the hydrophobic tail of OLA approaches the binding residues having EC, it replaces water molecules and goes into the binding pocket, making a stable complex.

4.3.3 Dynamics of water near protein surface

We explore the relaxation of water in terms of various dynamical quantities like solvent residence time, translational and rotational diffusion at the interface. To identify the interface, we calculate radial distribution function (g(r)) of water molecule around C_α atom of amino acid residues as discussed in methods. Fig.4.6(a) shows g(r) plot for MG-aLA (black) and MG-aLA-OLA(red). The peak is mostly unaffected due to complex formation in MG-aLA-OLA as compared to MG-aLA. We consider the water molecules up to 6 Å to probe the dynamics of the water at the interface. We consider the dynamics of hydration water around the protein in the MG-aLA-OLA (system 1), the ligand in MG-aLA-OLA (system 2), the protein in MG-aLA (system 3) and common to both protein and ligand (system 4).[32] The residence time of water molecule are calculated from

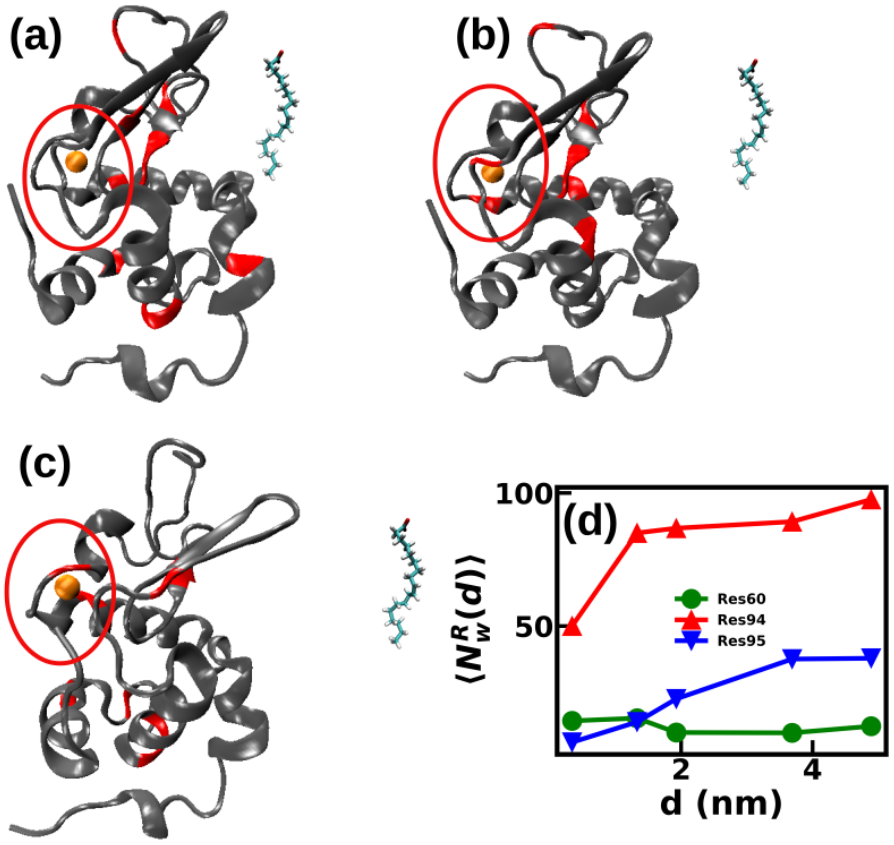

Figure 4.5: Essential coordinates are colored in red over crystal structure of aLA for different position of OLA i.e. (a) OLA at 1.34 nm, (b) OLA at 1.93 nm, (c) OLA at 2.5 nm. The ligand is shown in the figure. For better understanding, we show Ca^{2+} binding region of protein aLA along with Ca^{2+}. (d) Average number of water molecule ($\langle N_w^R(d) \rangle$) around ILE95 (blue), LYS94(red) and TRP60 (green).

the survival time auto-correlation function $C_S(t)$ (See in Methods).[146] Fig.4.6(b) shows that $C_S(t)$ for hydration water common to both protein and ligand exhibit slower decay of $C_S(t)$ than all other cases. We fit the $C_S(t)$ with bi-exponential functions. $\langle \tau_R \rangle$ for different systems are given in Table.4.3 that confirms that hydration water molecules common to both protein and ligand are the slowest. The dynamics of water around the protein and the ligand are not affected much by the presence of the other (system 1, 2 and 3).

This is further reflected in the mean-square-displacement (MSD) curve (See in Methods) where water molecule is tracked up to the survival time. Fig.4.6(c) shows MSD curve for different systems, where common water molecules exhibit the slowest motion as compared to other systems. The nature of the MSD suggest that water molecules present near the interface of protein exhibit a sub

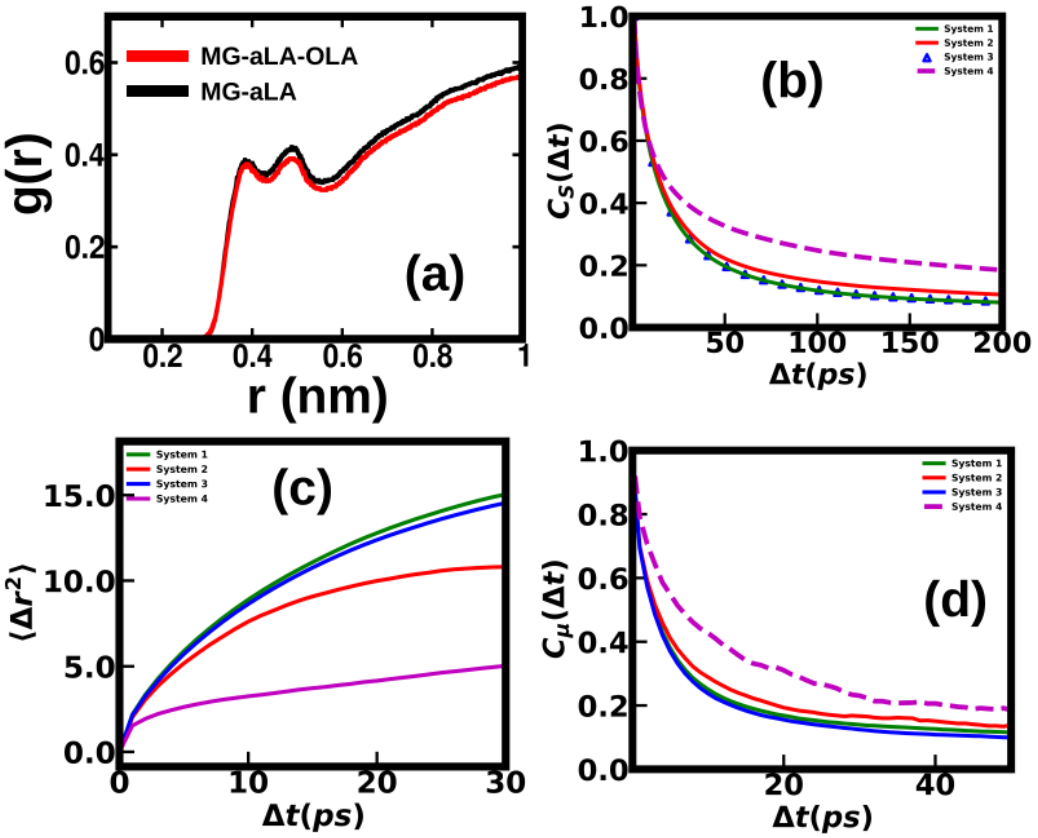

Figure 4.6: (a) Radial distribution function,g(r) of water molecules as a function of distance from the protein at MG-aLA and MG-aLA-OLA conformations, ((b) Survival time correlation function ($C_S(\Delta t)$), (c) mean square displacement ($\langle \Delta r^2 \rangle$), (d) Reorientation time correlation function ($C_\mu(\Delta t)$) for different systems. Systems are defined as follows: system 1: hydration water around protein in complex, system 2: hydration water around ligand in complex, system 3: hydration water around protein in free state (without ligand) and system 4: water molecules which are simultaneously present within a distance of 6Å from protein and ligand in complex i.e. common to both protein and ligand. Color is different for different systems.

linear time dependency.[32] We fit the MSD data with $\langle \Delta r^2 \rangle \sim t^\alpha$. We calculate α value for different system and tabulated in Table.4.3. It shows that $\alpha < 1$, confirming the sub linear dependency. The water molecules common to both protein and OLA have the least α value. It suggests that those water molecules exhibit slower dynamics, as suggested from residence time. The dipole-dipole reorientation time correlation function (TCF), $C_\mu(t)$ (See in Methods) in Fig.4.6(d) shows that the common water molecules possess slow decay which is also confirmed by amplitude-weighted average reorientation time constant, $\langle \tau_\mu \rangle$ (see Table.4.3), obtained from the fitted decay curve with a multi (tri-)exponential dependence. We find that common water molecules exhibit higher $\langle \tau_\mu \rangle$ value

	System 1	**System 2**	**System 3**	**System 4**
$\langle\tau_R\rangle$ (ps)	48.74	62.81	50.41	102.64
α	0.58	0.56	0.57	0.41
$\langle\tau_\mu\rangle$ (ps)	20.52	22.59	16.5	282.22

Table 4.3: Residence time, exponents and average reorientational time constants for different systems.

as compared to others. Hydration water molecules in system 1, 2 and 3 show similar dynamics. We find that lower α value corresponds to higher $\langle\tau_\mu\rangle$. Water molecules close to both protein and ligand (system 4) exhibit the lowered value of α and thus higher $\langle\tau_\mu\rangle$.

4.4 Conclusions

In conclusion, we have related the conformational change of protein aLA at MG state due to binding with fatty acid like OLA. The investigation reveals that protein MG state become stabilized upon OLA binding. The fluctuations in the complex is governed by degrees of freedom of some binding residues, which are eventually transferred to the Ca^{2+} binding region as ligand is completely separated out from the protein. The free energy change for those binding residues having EC along ϕ and ψ dihedral lies near the zero value. It suggests by small cost of free energy they are able to bind with OLA and form XAMLET complex. This low value of free energy might help in drug release. Our study is helpful in the microscopic understanding of OLA binding with protein at MG state and in stabilizing XAMLET for cancer treatment.

Appendix

A1. Steered molecular dynamics & umbrella sampling

For any kind of complex biological or chemical system, if the phase space is separated by high energy barrier then conventional molecular dynamics is not able to sample the phase space properly. Now to deal with this sampling problem, one can use some kind of biased molecular dynamics or enhanced sampling techniques. Steered molecular dynamics (SMD) and Umbrella Sampling (US) are two commonly used techniques in complex biological or chemical system to explore the kinetics of processes like protein ligand binding, protein folding e.t.c.[148] In this methods, at first a suitable reaction coordinate(ξ) is identified which represents the progress of the phenomena.

Here in protein-ligand system we choose centre of mass between protein-ligand as reaction coordinates. At first SMD techniques is used to pull the ligand by applying a constant force keeping the protein fixed. The position of ligand is maintained using a bias potential. Thus one can generate a series of configurations along the reaction coordinates. Out of this, some conformations are used as starting conformations of umbrella sampling. Here the bias potential applied on the system is $V_i(\xi)$ which is only function of ξ. Hence, biased energy of the system is

$$U_i^b(r) = U_i(r) + V_i(\xi) \tag{4.7}$$

Now, one can obtain the unbiased probability distributions ($P_i^u(\xi)$),

$$P_i^u(\xi) = \frac{\int exp[-\beta U_i(r)]\delta(\xi - \xi_i)dr}{\int exp[-\beta U_i(r)dr} \tag{4.8}$$

Here one can get biased probability distribution along reaction coordinate, $P_i^b(\xi)$ as follows

$$P_i^b(\xi) = \frac{\int exp[-\beta U_i(r) + V_i(\xi)]\delta(\xi - \xi_i)dr}{\int exp[-\beta U_i(r) + V_i(\xi)]dr} \tag{4.9}$$

therefore,

$$P_i^b(\xi) = exp(-\beta V_i(\xi))\frac{\int exp[-\beta U_i(r)]\delta(\xi - \xi_i)dr}{\int exp[-\beta U_i(r) + V_i(\xi)]dr} \tag{4.10}$$

Now, equation 4.10 can be written as equation 4.8 as follows,

$$P_i^u(\xi) = P_i^b(\xi) exp[\beta V_i(\xi)] \langle exp[-\beta V_i(\xi)] \rangle \tag{4.11}$$

From here one can get the free energy as follows,

$$A_i(\xi) = -\frac{1}{\beta} ln(P_i^b(\xi)) - V_i(\xi) + F_i \tag{4.12}$$

where, $F_i = -\frac{1}{\beta} ln(\langle exp[-\beta V_i(\xi)] \rangle)$ depends on applied bias potential which is need to be determined.

Now to obtain a global potential mean force(PMF) curve, one need to combine free energy of each window. So, calculation of F_i is necessary to get global PMF curve. The Weighted Histogram Analysis Method (WHAM) is popular technique to calculate PMF from set of US simulations.

A2. Weighted Histogram Analysis Method (WHAM)

WHAM[144] is a popular method to obtain the unbiased PMF by combining the data from biased probability distributions obtained from US simulation in multiple windows along reaction coordinate. The unbiased probability distribution is as follows,

$$P^u(\xi) = \sum_{i=1}^{N} w_i(\xi) P_i^u(\xi) \tag{4.13}$$

Now, for a one dimensional reaction coordinate, the WHAM is given by

$$P^u(\xi) = \frac{\sum_i^N n_i(\xi) P_i^b(\xi)}{\sum_j^N n_j exp[-\beta(V_j(\xi) - F_j)]} \tag{4.14}$$

and

$$exp[-\beta F_j] = \int exp[-\beta V_j(\xi)] P^u(\xi) d\xi \tag{4.15}$$

Here, n_i is defined as total number of data points in i-th histogram. One can get unbiased probability distributions by iteratively solving eqn. 4.14 and 4.15 until convergence is obtained.

Chapter 5 Coarse-grained model of protein with structural informations

5.1 Introduction

In previous chapters, we show that dihedral angles provide insight to the function of protein. Typically, dihedral angle information extracted from the all atom description of the protein. We treat each atom including solvent explicitly and interactions between atoms are described based on validated force-field. In such calculations, we get detailed description of each atom and chemical bonds of the system. The major drawback is that the all atom approach is computationally very expensive. It requires high computational resources due to involvement of large number of degree of freedoms. Considering more than a single protein is propitiously difficult.[34]

Often, various cellular phenomena involve ample number of bio-molecules ranging from water, small and medium-size oligomers and co-polymers (peptides, proteins, RNA, etc.) to huge co-polymers, such as DNA. For instance, obtaining all atom description in the process of membraneless organelles via aggregation of intrinsically disordered protein[149] are computationally challenging. Lowering the representation from all explicit atoms to coarse grained (CG) model is called for to study systems involving large numbers of bio-molecules.[35,36]

Coarse grained representation of protein is already discussed in earlier literatures.[35,36,150] The main purpose of those studies was to reduce the number of degrees of freedom. The simple lattice protein-like HP models are the examples of such kind.[151–154] In this model, each amino acid is either represented as hydrophobic or polar which ignore protein backbone information. This model is widely used to understand folding pathway of the protein. Bagchi et al.[155]

represent a CG model of protein where amino acid of protein is represented in terms of two atoms. One atom represents the backbone C_α atom, while the other one represents the whole side chain residues. The model investigates the folding dynamics and energy landscape picture of protein conformations for two different protein HP-36 and amyloid beta using extensive Brownian dynamics simulations.

Due to reduction in degrees of freedom, conformational space become restricted. It is known that that protein functionality largely depends on its structure.[37] Earlier studies suggest that protein structural and functional aspects could be understood in terms of protein dihedral angles.[7,12,41] The main drawback of earlier CG model is these kinds of analysis does not provide any structural information in. Motivated by this, we study a simple polymer model to make a CG description of protein with dihedral angle information. We consider each amino acid of protein as a bead. Beads have both bonded and non-bonded interactions between them. Bonded interactions are mainly governed by bending of two consecutive beads and stretching of three consecutive beads. Non-bonded interaction is given by Lennard-Jones(L-J) 12-6 potential. We model the solvent as follows: Solvophobic beads interact with solvent particles via repulsive interactions. Solvophilic beads interact with solvent via L-J potential. Each bead is assigned with two additional degrees of freedom to represent two dihedral angles. Coupling between the dihedral in the model is given using free energy landscape obtained from all atom molecular dynamics simulation. We use Monte Carlo (MC) method to generate conformations based on the model interactions. We consider protein GB3 which have 56 amino acid residues. Our simulation study shows good structural agreement of average dihedral angle obtained from coarse-grained simulations with the crystal structure. We also check whether one can reproduce structural similarity for other protein based on our CG model using energy profile generated for GB3. We test this on protein ubiquitin (Ub) and disordered Bacteriophage λN protein. We find similarity in secondary structure element for ubiquitin. For λN protein, we find similarity during comparison with the crystal structure.

5.2 Model and simulation method

5.2.1 All atom simulation to generate dihedral interaction

All-Atom molecular dynamics (See chapter 2, Appendix A1) simulations are used to generate free energy profile of conformational variables ϕ and ψ dihedral

angle of protein. The dihedral couplings are estimated as follows: (i)The dihedral angles, ϕ and ψ, are computed from atomic coordinates of the residues. (ii) For the intra-residue coupling, joint probability distribution for dihedral $P(\phi, \psi)$ are calculated. $P(\phi, \psi)$ gives the probability of ϕ and ψ occurring simultaneously. We obtain free energy profile from the joint probability distributions based on formula: $G(\phi, \psi) = -k_B T ln(P(\phi, \psi))$. Finally, for intra-residual coupling we got two energy profile, one corresponds to all hydrophobic amino acid residues and other hydrophilic amino acid residues. (iii) We use similar method to calculate free energy profile for inter residual dihedral coupling. For the inter residue dihedral couplings, a 7Å cut-off is taken about a given residue in each frame. Then dihedral angles of corresponding residues are characterized based on hydrophobicity and hydrophilicity of the residues. Finally, three possible combinations of data sets are calculated: hydrophobic-hydrophobic, hydrophobic-hydrophilic/hydrophilic-hydrophobic, hydrophilic-hydrophilic. For each set, 4 possible dihedral combinations are possible and negative logarithmic of each joint distribution gives a corresponding free energy profile.

We consider GB3 (PDB id: 2OED)[73] with 56 amino acids in our studies. Amber99sb force field (ff)[75] is used in the GROMACS software package[76] for simulation. The TIP3P water model is considered for solvent molecules, and counter-ions are added for electroneutrality. Particle Mesh Ewald method is used to assess long ranged electrostatic energy.[86] 10 Angstrom (Å) is considered as truncation limit for both Lennard-Jones and short range interactions. GB3 protein is solvated in a cubic box of dimensions 5.6x5.6x5.6 nm^3 with 5524 water molecules 2 Na^+ ions are added to attain electrical neutralization. Minimization is done for 50,000 steps using the steepest descent algorithms. All bonds in the protein are constrained using the LINCS method. Equations of motion are integrated using leap-frog algorithm with an integration time step of 2 femtosecond (fs). Systems are equilibrated through NVT and NPT simulations using position restraints on heavy atoms at 300K Temperature and 1 Bar pressure, respectively. The production NPT runs are executed for 1.05 μs with 2 fs time step integration employing periodic boundary conditions in all directions. The average radius of gyration of the residues (4Å) is taken to be the bead diameter.

We also consider two other proteins: (i) protein ubiquitin (PDB ID: 1UBQ) and (ii) disordered Bacteriophage λN protein (PDB ID: 1QFQ).[156] The initial model structure for disordered protein is taken without RNA from PDB structure 1QFQ. We perform all atom simulation of both protein using GROMACS. Detailed of ubiquitin simulation is already discussed in chapter 2. For, λN simulations we

used a cubic water box of dimensions 6.87249x6.87249x6.87249 nm^3 with the periodic boundary conditions in all directions. Electroneutrality is maintained with adding 6 Cl^- atom. The system is energy minimized for 50,000 steps using the steepest descent algorithms. Lennard-Jones and short-range electrostatic interactions are truncated at 10 Å. Long Range electrostatic energy is calculated using PME method. For the equilibration of the system, NVT equilibration phase is followed by a longer NPT phase equilibration. Temperature and pressure are kept at 300K and 1 bar. Production MD is run for 1 microsecond using 2 femtosecond time step integration.

5.2.2 Coarse-grained model

The protein is modelled as a flexible polymer chain by a bead spring model, where each beads correspond to an amino acid of protein. The bonded interaction between beads corresponds to stretching between neighboring beads with cutoffs, $V_{bond} = k_b(l - l_0)^2$ where l is the distance between two successive beads, k_b is force constant and l_0 is equilibrium bond length.[157] Here, we choose $l_0 = 0.7\sqrt{2}\sigma$ according to Ref[157] and σ is the diameter of each bead. The change in bond angle costs energy: $V_{angle}(\theta) = \frac{1}{2}k_\theta \cos^2\theta$, where k_θ is force constant and $\theta = \cos^{-1}(\frac{\vec{r_{ij}}.\vec{r_{jk}}}{|\vec{r_{ij}}||\vec{r_{jk}}|})$ is the angle between three consecutive monomers i, j, k. Force constants k_b and k_θ are chosen as $30.0\epsilon/\sigma^2$ and $10.0\epsilon/\sigma^2$ respectively. Non-bonded interaction between two monomer is governed by Lennard-Jones(L-J) 12-6 potential, $V_{lj}(r) = 4\epsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6]$, $r < 2^{\frac{1}{6}}\sigma$. The solvent interaction is of L-J type i.e. $V_{lj}(r) = 4\epsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6]$, the bead-solvent interactions are of two types: (i) solvophobic beads interact with solvent beads via repulsive interaction, $V_{Solvophobic}(r) = 4\epsilon[(\frac{\sigma}{r})^{12}]$; and (ii) solvophilic-solvent interactions with L-J type interaction, $V_{Solvophilic}(r) = 4\epsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6]$.

We perform Monte-Carlo (MC) simulation on CG model of protein in canonical (NVT) ensemble utilizing metropolis algorithm where each bead is assigned its position and two dihedral angles. When we construct MC move, we compute the energy cost due to different interactions like bonded, non-bonded and solvent contributions. We also change the dihedral degree of freedom and calculate the energy cost from potential generated from all-atom simulation using linear interpolation. The interaction between dihedrals of a solvophobic bead is taken from free energy profile of hydrophobic beads. The interaction between dihedrals of a solvophilic bead is taken from free energy profile of hydrophilic beads. Details of MC simulation is described in Appendix A1. MC simulations are

carried out in a cubic box of length L=10.6 in reduced unit with periodic boundary conditions in all direction. The system temperature is set as $k_BT = 1$ with k_B being the Boltzmann's constant and T is the temperature. The total number of particle is 1000 where 56 beads correspond to protein residue and the remaining are solvent particles. For inter-residual dihedral coupling, we set a cut-off 1.6σ which satisfies that inter-residual coupling is considered only if two residues are specific cutoff. We perform 1,00,000 number of MC steps out of which first 30,000 MC steps are discarded for equilibration, as judged by potential energy of the system.

Different quantities like solvent distributions, dihedral angle are averaged for configurations[85] of last 70,000 steps of multiple independent trajectories.

5.3 Results and discussion

We perform all-atom simulation on GB3 to calculate the free energy profile. The coarse-grained model of GB3 is considered as per the sequence of the protein where the hydrophilic amino acids are represented by solvophilic beads while hydrophobic amino acids are represented by solvophobic beads. The initial value of dihedral angle of each bead is taken from protein crystal structure data.

5.3.1 Free energy profile

Intra residual dihedrals: Fig.5.1(a)-(b) shows free energy landscape (FEL) (ϕ vs ψ) for intraresidual dihedral coupling. Fig.5.1(a) shows FEL plot for hydrophobic residues and Fig.5.1(b) shows FEL plot for hydrophilic residues. For hydrophobic residues, two deep minima (A1 and A2 region in Fig.5.1(a)) and a shallow minima (A3 region in Fig.5.1(a)) are observed in the free energy surface. On the other hand, for hydrophilic residues deep minima region B1 and B2 (Fig.5.1(b)) remain same as hydrophobic FEL. Although, another deep minima is observed in FEL (Region B3) for hydrophilic residues, which is absent in the FEL plot for hydrophobic residues.

Inter-residual dihedral: Next, we discuss on FEL (ϕ vs ψ) for inter-residual dihedral coupling in Fig.5.2. Fig.5.2(a)-(d) shows FEL plot for various combination of dihedral angle of two different solvophobic resiudes. Fig.5.2(a) shows FEL plot for dihedral ϕ_i and ϕ_j where i,j corresponds to two different residue within a cut off of 7 Å $C_\alpha - C_\alpha$ distances. Deep minima along with some shallow minima is observed in FEL. The FEL plot is changed for dihedral ϕ_i and ψ_j combination

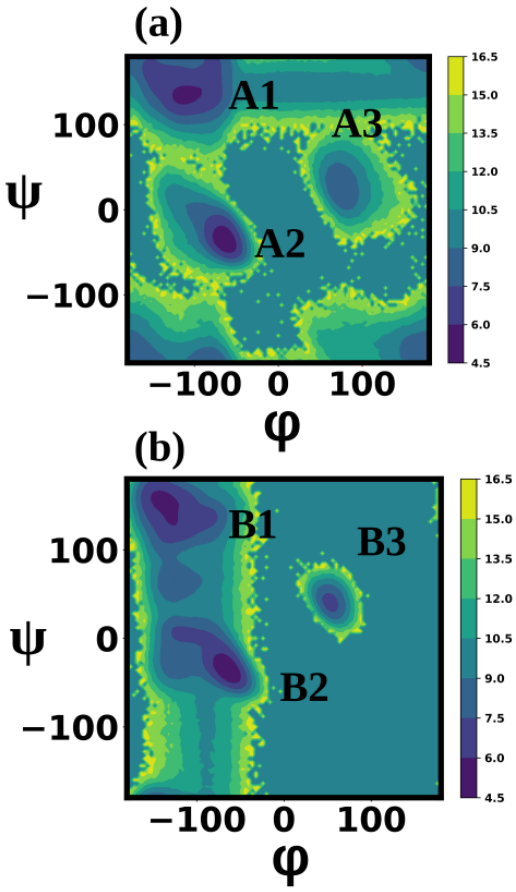

Figure 5.1: Free energy landscape for intraresidual dihedral coupling, considering all (a) hydrophobic residues and (b) hydrophilic residues. Minimum energy region are marked.

(Fig.5.2(b)) due to the presence of some extra deep minima. The position of one deep minima remain same as dihedral ϕ_i and ϕ_j combination. The position of minimae is changed for ψ_i and ϕ_j combination (Fig.5.2(c)). Fig.5.2(d) shows FEL plot for ψ dihedral for two different residues. We observe that the positions of minima is changed. If we compare FEL plots Fig.5.2(a)-(d), we find that for every possible combination of dihedral angle, there is a deep minima which remain same in all FEL plot (Region M1 in the figure).

Next, Fig.5.2(e)-(h) shows similar FEL plot if one residue is hydrophobic other one is hydrophilic or vice versa. Fig.5.2(e) shows FEL plot for dihedral ϕ. The nature of FEL is comparable with Fig.5.2(a) i.e. ϕ_i and ϕ_j FEL plot for hydrophobic residues. Deep minima remain in the same position as earlier hydrophobic-hydrophobic case. Similarly, position of energy minima for ϕ-ψ dihedral combination (Fig.5.2(e)) mostly remain the same as ϕ-ψ dihedral combination for hydrophobic-hydrophobic residue combination(Fig.5.2(f)). Fig.5.2(g)-(h) shows FEL plot for $\psi_i - \phi_j$ and $\psi_i - \psi_j$ combination. A common region of deep minima is present in same position for both cases.

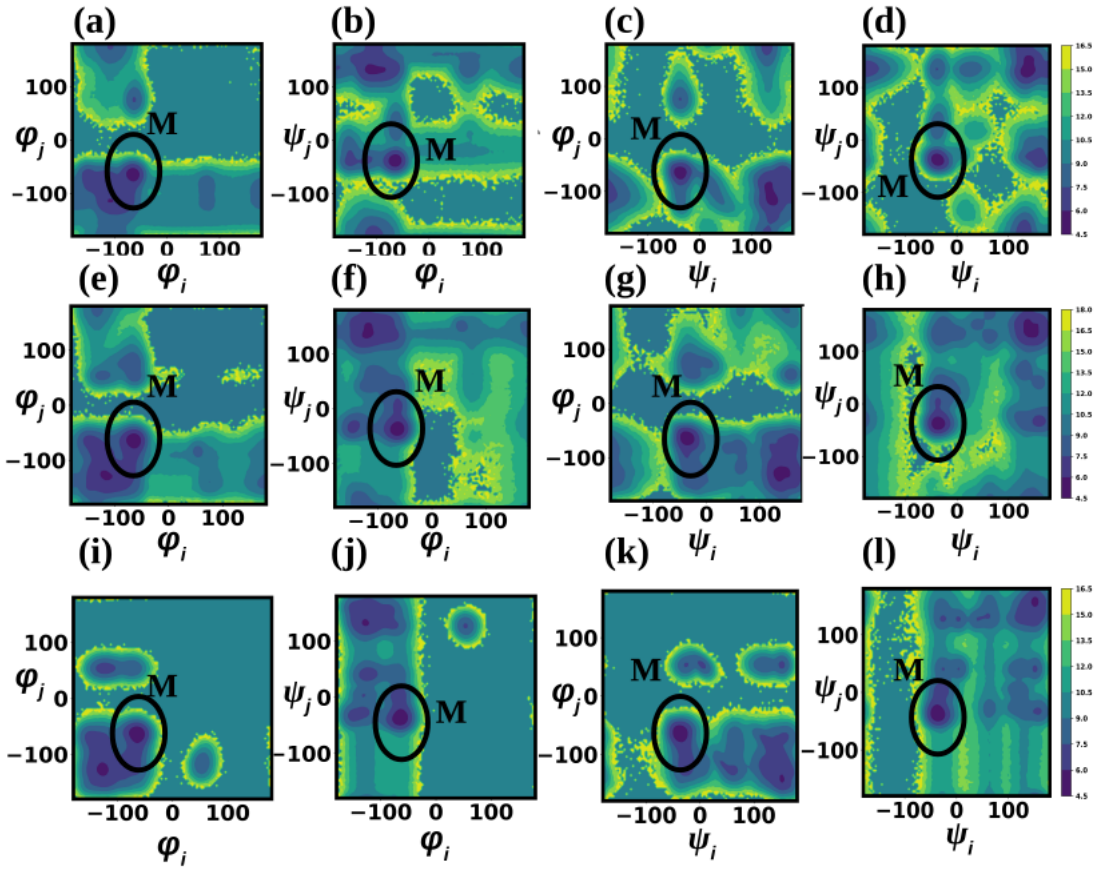

Figure 5.2: Free energy landscape for interresidual dihedral coupling considering (a-d) hydrophobic-hydrophobic residues and (e-h) hydrophobic-hydrophilic/hydrophilic-hydrophobic residues, (i-l) hydrophilic-hydrophilic residues. Common region of deep minima, M is marked by a circle in all figure.

Next, we discuss on FEL for hydrophilic-hydrophilic residue combination, Fig.5.2(i)-(l). Here we also find similarity with earlier cases. For ϕ-ϕ dihedral angle, FEL plot in hydrophilic-hydrophilic residue combination, is similar with earlier cases (Fig.5.2(a) and (e)). The position of most of the deep minima in other cases ϕ_i-ψ_j (Fig.5.2(j)), ψ_i-ψ_j (Fig.5.2(k)) and ψ_i-ψ_j (Fig.5.2(l)) are different. Although in all cases for hydrophilic-hydrophilic residues, one deep minima is in the same position.

Overall, the analysis on FEL of inter residual dihedral coupling suggest that position of a deep minima is fixed in all possible combination. We marked that portion, M by a circle in each plot.

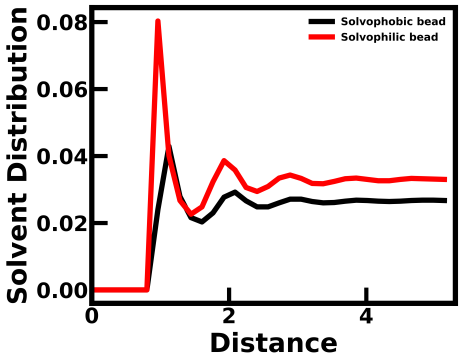

Figure 5.3: Solvent distributions around solvophobic and solvophilic bead

5.3.2 Comparison of dihedral distributions for CG and all atom simulations

Fig.5.3(a) shows solvent distribution around solvophobic bead (black) and solvophilic bead (red). The distribution shows that first peak of solvent distributions is higher for solvophilic beads as compared to solvophobic beads. It suggests the solvent distributions as per the interactions.

Next, we compare the structural element of each residue obtained from the coarse-grained MC simulation with initial crystal structure and average structure based on all atom simulations. Table.5.1 tabulates secondary structural element for each residue of GB3 for three different cases, namely: (i) initial crystal structure, (ii) average structure obtained from all atom MD simulations. (iii) average structure obtained from CG MC simulation. Here average conformation is obtained based on averaging over all equilirbated conformations. We identify helix (H), sheet (S) and unstructured (U) region based on the dihedral angles (ϕ and ψ) from known range.[158] Unstructured region corresponds to other than helix and sheet region of protein. At first we check similarity of secondary structural element obtained from MC simulation with initial crystal structure. We find that secondary structural element remain same for 73% residues. Next we compare secondary structural element obtained from MC to average structure obtained from all atom simulations. We found that secondary structure remain intact for approx 70 % residues.

5.3.3 Ramachandran plot (RC) analysis

We further consider $\phi - \psi$ Ramachandran plot (RC) to compare average structure obtained from MC run with crystal structure and average structure obtained

Res	Crystal Structure	Average Structure (MD)	Average Structure (MC)	Res	Crystal Structure	Average Structure (MD)	Average Structure (MC)
1	U	S	U	29	H	H	H
2	S	S	S	30	H	H	H
3	S	S	S	31	H	H	H
4	S	S	S	32	H	H	H
5	S	S	S	33	H	H	H
6	S	S	S	34	H	H	H
7	S	S	S	35	H	H	H
8	S	S	S	36	H	H	H
9	U	U	U	37	H	H	U
10	U	U	H	38	U	U	U
11	U	U	H	39	U	U	S
12	U	U	S	40	U	U	S
13	S	S	S	41	U	U	H
14	S	S	U	42	S	S	S
15	S	S	U	43	S	S	S
16	S	S	U	44	S	S	S
17	S	S	U	45	S	S	S
18	S	S	S	46	S	S	S
19	S	S	S	47	U	U	H
20	U	S	U	48	U	U	H
21	U	U	H	49	U	U	U
22	H	U	U	50	U	S	U
23	H	H	H	51	S	S	S
24	H	H	H	52	S	S	S
25	H	H	H	53	S	S	S
26	H	H	H	54	S	S	S
27	H	H	H	55	S	S	S
28	H	H	H	56	S	S	S

Table 5.1: Comparison of secondary structural element for GB3 protein. MD denotes trajectory of all atom molecular dynamics simulations and MC denotes conformations based on monte carlo simulations. 'H' signifies Helix, 'S' signifies sheet and 'U' corresponds to other than element helix or sheet i.e. loop/coil/turn/bend region of the protein.

from all atom molecular dynamics simulations. Fig.5.4(a) shows a correlation plot where green circles represent crystal structure and solid triangles represent average CG structure. The average dihedral angle obtained from MD simulation is marked in $\phi - \psi$ plot as black circle. Secondary structure is comparable to the crystal structure and average structure obtained from MD of the protein GB3.

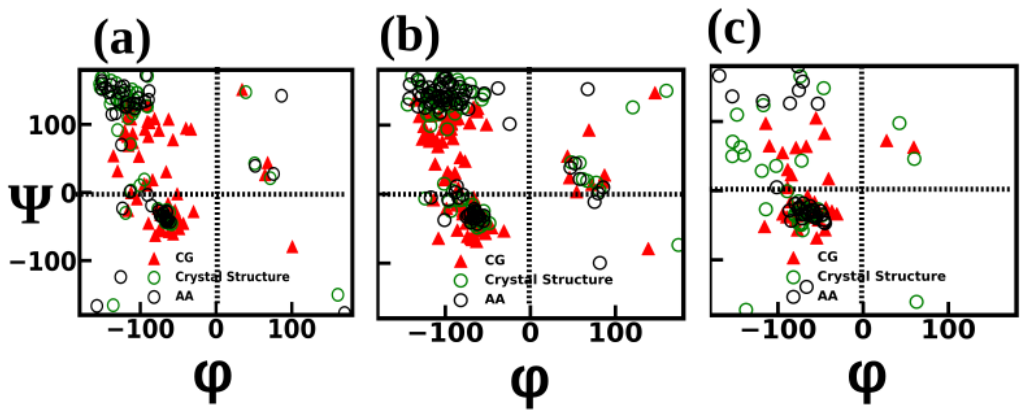

Figure 5.4: Comparison of Ramachandran plot for different structure obtained from crystal structure, average structure based on MD simulations and average structure based on MC simulations for (a) protein GB3, (b) protein Ub, (c) λN protein. Triangle represents average dihedral angle obtained from MC and green circle represents initial dihedral angle of amino acids obtained from crystal structure and black circle represents average dihedral angle of amino acids obtained from all atom MD simulations.

5.3.4 Transferability of the coarse-grained model to other proteins

Next we check transferability of the model where we use free energy profile obtained from all atom simulation of GB3 to Ub. At first coarse grained model of Ub are arranged according to the sequence of amino acids in Ub based on hydrophobic and hydrophilic nature as discussed earlier. The bonded and non bonded interactions are all same as GB3 protein. Table.5.2 shows secondary structural element for each residue for three different cases. Average structure generated using CG Monte Carlo method shows 70% structural similarity with initial crystal structure. The secondary structure remain intact for 68.4% residues when secondary structure element is compared with average structure generated using all atom molecular dynamics simulations. Fig.5.4(b) shows comparison of RC plot for ubiquitin. RC plot for Ub protein suggests that average structure obtained from CG Monte Carlo is comparable with crystal structure as well as average structure generated from all atom molecular dynamics simulations.

Next, we consider protein λN keeping the bonded and non-bonded interaction same as GB3. Beads are arranged according to hydrophilic and hydrophobic nature of amino acids present in the protein sequence of λN. Table.5.3 tabulates secondary structural element for λN protein. We observe that 65% residues possess the same secondary structure element in average structure obtained from MC simulations as compared to initial crystal structure. We get a lower ($\sim$)50% similarity percentage in comparison with average structure based on all atom

MD simulations. The low value of percentage might be due to the disordered nature of protein and choice of proper force field during GB3 simulation. RC plot for λN protein is in Fig.5.4(c). Thus, the energy profile generated using GB3, can be used to generate conformations for other proteins like ubiquitin which shows high structural similarity with all atom descriptions. It attributes the transferability of model effectively.

5.4 Conclusions

In conclusion, we build up a CG representation of protein where the amino acids are treated as beads with dihedral angle as additional degrees of freedom. We find structural similarity for each residue with all atom descriptions. The transferability of the model is validated using FEL obtained from all atom molecular dynamics of GB3 to other protein like Ub and λN protein. This can open the way for the CG model of protein with structural information and may be useful to study phenomena involving many protein molecules, like protein aggregations. A combined molecular dynamics and Monte Carlo approach might be useful to obtain time dependent information.

Res	Crystal Structure	Average Structure (MD)	Average Structure (MC)	Res	Crystal Structure	Average Structure (MD)	Average Structure (MC)
1	S	S	U	39	U	H	H
2	S	S	S	40	S	H	H
3	S	S	S	41	S	S	S
4	S	S	S	42	S	S	S
5	S	S	S	43	S	S	S
6	S	S	S	44	S	S	S
7	S	S	S	45	S	U	S
8	U	U	U	46	U	U	U
9	U	U	U	47	U	U	U
10	S	S	U	48	S	U	S
11	S	S	S	49	S	U	S
12	S	S	S	50	S	U	S
13	S	S	S	51	U	U	S
14	S	S	S	52	U	U	U
15	S	S	S	53	U	U	H
16	S	S	S	54	U	U	U
17	S	S	H	55	U	U	S
18	U	U	U	56	H	H	H
19	U	U	H	57	H	H	H
20	U	U	U	58	H	H	H
21	U	U	S	59	H	H	H
22	U	U	S	60	U	U	U
23	H	H	H	61	U	U	S
24	H	H	H	62	U	U	S
25	H	H	H	63	U	U	S
26	H	H	H	64	S	U	U
27	H	H	H	65	S	U	H
28	H	H	H	66	S	S	S
29	H	H	H	67	S	S	S
30	H	H	H	68	S	S	S
31	H	H	H	69	S	S	S
32	H	H	H	70	S	S	S
33	H	H	H	71	S	S	S
34	H	H	U	72	S	U	S
35	U	U	S	73	U	U	S
36	U	U	S	74	U	U	S
37	U	U	S	75	U	U	U
38	U	H	H	76	U	U	U

Table 5.2: Comparison of secondary structural element for Ub protein.

Res	Crystal Structure	Average Structure (MD)	Average Structure (MC)	Res	Crystal Structure	Average Structure (MD)	Average Structure (MC)
2	U	U	U	20	H	H	U
3	U	U	H	21	H	H	U
4	H	U	H	22	U	U	H
5	H	U	H	23	U	U	H
6	H	U	H	24	U	U	U
7	H	U	H	25	U	U	S
8	H	U	H	26	U	U	U
9	H	U	H	27	U	U	U
10	H	U	U	28	U	U	U
11	H	H	H	29	U	U	S
12	H	H	H	30	U	U	S
13	H	H	H	31	U	U	U
14	H	H	H	32	U	U	H
15	H	H	H	33	U	U	S
16	H	H	H	34	U	U	S
17	H	H	H	35	U	U	S
18	H	H	H	36	U	U	U
19	H	H	H				

Table 5.3: Comparison of secondary structural element for λN protein.

Appendix

A1. Monte Carlo simulations (MC)

MC[85] is a popular method of molecular simulation which is used to obtain consequence of stochastic process using random number generation and probabilistic statistics. The method of simulation is closely related to random experiment where outcome is not known apriori. In this method random walk algorithm is used to perform equilibrium sampling over the statistical ensemble. For a system of N particle, the partition function Q is defined as:

$$Q = c \int d\mathbf{p}^N d\mathbf{r}^N exp[-\frac{H(\mathbf{r}^N, \mathbf{p}^N)}{k_B T}] \tag{5.1}$$

where, **q** and **p** signifies coordinate and momenta of each particle, Hamiltonian $H(\mathbf{r}^N, \mathbf{p}^N)$ represents total energy of the system, k_B is Boltzmann's constant and T is temperature. The pre-factor c is defined as $c = \frac{1}{h^{3N} N!}$. The expectation value of any variable A ($\langle A \rangle$) is defined as:

$$\langle A \rangle = \frac{\int d\mathbf{p}^N d\mathbf{r}^N A(\mathbf{r}^N, \mathbf{p}^N) exp[-\frac{H(\mathbf{r}^N, \mathbf{p}^N)}{k_B T}]}{\int d\mathbf{p}^N d\mathbf{r}^N exp[-\frac{H(\mathbf{r}^N, \mathbf{p}^N)}{k_B T}]} \tag{5.2}$$

Now, k, kinetic energy solely depends on momentum hence above equation can be solved analytically only for the momentum term.

Rather directly computing the integral $\int d\mathbf{r}^N A(\mathbf{r}^N) exp[-\frac{u(\mathbf{r}^N)}{k_B T}]$, it is computed as;

$$\int d\mathbf{r}^N A(\mathbf{r}^N) exp[-\frac{u(\mathbf{r}^N)}{k_B T}] / \int d\mathbf{r}^N exp[-\frac{u(\mathbf{r}^N)}{k_B T}] \tag{5.3}$$

The term $\frac{exp[-\frac{u(\mathbf{r}^N)}{k_B T}]}{\int d\mathbf{r}^N exp[-\frac{u(\mathbf{r}^N)}{k_B T}]}$ is defined as the probability density of finding the system in a configuration space around $\mathbf{r}^N$. Hence, here the relative probability of visiting different point of the configuration space is considered rather consideration of absolute probability.

Now the sampling around probable configuration space is done using metropolis algorithm. At first, a random particle at random position is chosen and corresponding potential energy ($u(\mathbf{r}^N)$) is computed. Therefore a random displacement, Δ is given to the particle so that new position is $r' = r + \Delta$ and new potential energy is $u(\mathbf{r'}^N)$. If change of energy $\Delta U = u(\mathbf{r'}^N) - u(\mathbf{r}^N)$ is negative then new configuration is considered as updated one. Now, the probability

of finding the particle in the new position is, $p_n \propto exp[-\frac{u(\mathbf{r'}^N)}{k_B T}]$ and in the old position is $p_o \propto exp[-\frac{u(\mathbf{r}^N)}{k_B T}]$. If the ratio $(\frac{p_n}{p_i})$ greater than a random number between 0 to 1 then also the new position is updated. The process continues for all particles in the system.

Periodic boundary condition and minimum image convention is applied along all direction. A cut off is considered for the truncation of the interaction. Now, if the amount of random displacement is small then all moves will be accepted, which results poor sampling of the configuration space. Similarly, if the displacement is large then all moves will be rejected. If the total number of trial move is N_{trial} displacing all the particle of number N and the number for which the MC move is accepted, N_{accept} then the value of displacement is controlled throughout the simulation such a way so that the acceptance ratio (N_{accept}/N_{trial}) reaches to a optimal value of 0.5. It ensures proper sampling of equilibrium phase space. Therefore, one can compute various thermodynamical quantity over the conformations.

Bibliography

[1] P. L. Di and A. Giuliani, "Protein structures as complex systems: a simplification conundrum," *Adv Syst Biol*, vol. 3, no. 1, pp. 7–9, 2014.

[2] H. Frauenfelder, *The physics of proteins: an introduction to biological physics and molecular biophysics*. Springer Science & Business Media, 2010.

[3] X. Salvatella, "Understanding protein dynamics using conformational ensembles," *Protein Conformational Dynamics*, pp. 67–85, 2014.

[4] A. Ramanathan, A. Savol, V. Burger, C. S. Chennubhotla, and P. K. Agarwal, "Protein conformational populations and functionally relevant substates," *Accounts of chemical research*, vol. 47, no. 1, pp. 149–156, 2014.

[5] K. Lindorff-Larsen, P. Maragakis, S. Piana, and D. E. Shaw, "Picosecond to millisecond structural dynamics in human ubiquitin," *The Journal of Physical Chemistry B*, vol. 120, no. 33, pp. 8313–8320, 2016.

[6] S. Dutta, M. Ghosh, and J. Chakrabarti, "Spatio-temporal coordination among functional residues in protein," *Scientific Reports*, vol. 7, no. 1, p. 40439, 2017.

[7] A. Das, J. Chakrabarti, and M. Ghosh, "Conformational contribution to thermodynamics of binding in protein-peptide complexes through microscopic simulation," *Biophysical journal*, vol. 104, no. 6, pp. 1274–1284, 2013.

[8] S. Brandt, F. Sittel, M. Ernst, and G. Stock, "Machine learning of biomolecular reaction coordinates," *The journal of physical chemistry letters*, vol. 9, no. 9, pp. 2144–2150, 2018.

[9] A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, "Dihedral angle principal component analysis of molecular dynamics simulations," *The Journal of chemical physics*, vol. 126, no. 24, p. 244111, 2007.

[10] R. B. Fenwick, S. Esteban-Martin, B. Richter, D. Lee, K. F. A. Walter, D. Milovanovic, S. Becker, N. A. Lakomek, C. Griesinger, and X. Salvatella, "Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition," *Journal of the American Chemical Society*, vol. 133, pp. 10336–10339, jul 2011.

[11] D. Long and R. Brüschweiler, "Structural and entropic allosteric signal transduction strength via correlated motions," *The Journal of Physical Chemistry Letters*, vol. 3, pp. 1722–1726, jun 2012.

[12] A. Das, M. Ghosh, and J. Chakrabarti, "Time dependent correlation between dihedral angles as probe for long range communication in proteins," *Chemical Physics Letters*, vol. 645, pp. 200–204, 2016.

[13] D. Yang, R. Konrat, and L. E. Kay, "A multidimensional nmr experiment for measurement of the protein dihedral angle ψ based on cross-correlated relaxation between 1hα- 13cα dipolar and 13c '(carbonyl) chemical shift anisotropy mechanisms," *Journal of the American Chemical Society*, vol. 119, no. 49, pp. 11938–11940, 1997.

[14] B. Reif, M. Hennig, and C. Griesinger, "Direct measurement of angles between bond vectors in high-resolution nmr," *Science*, vol. 276, no. 5316, pp. 1230–1233, 1997.

[15] R. L. Hill and K. Brew, "Lactose synthetase," *Adv Enzymol Relat Areas Mol Biol*, vol. 43, pp. 411–490, 1975.

[16] D. Barrick and R. L. Baldwin, "The molten globule intermediate of apomyoglobin and the process of protein folding," *Protein Science*, vol. 2, no. 6, pp. 869–876, 1993.

[17] N. Bhattacharjee, P. Rani, and P. Biswas, "Capturing molten globule state of α-lactalbumin through constant ph molecular dynamics simulations," *The Journal of Chemical Physics*, vol. 138, no. 9, p. 03B601, 2013.

[18] J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky, "Introducing protein intrinsic disorder," *Chemical reviews*, vol. 114, no. 13, pp. 6561–6588, 2014.

[19] M. J. Kronman and G. D. Fasman, "Metal-ion binding and the molecular conformational properties of α lactalbumi," *Critical reviews in biochemistry and molecular biology*, vol. 24, no. 6, pp. 565–667, 1989.

[20] A. Jain and G. Stock, "Identifying metastable states of folding proteins," *Journal of chemical theory and computation*, vol. 8, no. 10, pp. 3810–3819, 2012.

[21] A. Jain and G. Stock, "Hierarchical folding free energy landscape of hp35 revealed by most probable path clustering," *The Journal of Physical Chemistry B*, vol. 118, no. 28, pp. 7750–7760, 2014.

[22] F. Sittel and G. Stock, "Robust density-based clustering to identify metastable conformational states of proteins," *Journal of chemical theory and computation*, vol. 12, no. 5, pp. 2426–2435, 2016.

[23] F. Sittel, T. Filk, and G. Stock, "Principal component analysis on a torus: Theory and application to protein dynamics," *The Journal of chemical physics*, vol. 147, no. 24, p. 244101, 2017.

[24] F. Sittel and G. Stock, "Perspective: Identification of collective variables and metastable states of protein dynamics," *The Journal of chemical physics*, vol. 149, no. 15, p. 150901, 2018.

[25] D. Nagel, A. Weber, B. Lickert, and G. Stock, "Dynamical coring of markov state models," *The Journal of chemical physics*, vol. 150, no. 9, p. 094111, 2019.

[26] C. Barbana, M. Pérez, L. Sánchez, M. Dalgalarrondo, J.-M. Chobert, T. Haertlé, and M. Calvo, "Interaction of bovine α-lactalbumin with fatty acids as determined by partition equilibrium and fluorescence spectroscopy," *International dairy journal*, vol. 16, no. 1, pp. 18–25, 2006.

[27] J. K. Marzinek, P. J. Bond, G. Lian, Y. Zhao, L. Han, M. G. Noro, E. N. Pistikopoulos, and A. Mantalaris, "Free energy predictions of ligand binding to an α-helix using steered molecular dynamics and umbrella sampling simulations," *Journal of Chemical Information and Modeling*, vol. 54, no. 7, pp. 2093–2104, 2014.

[28] S. Tolin, G. De Franceschi, B. Spolaore, E. Frare, M. Canton, P. Polverino de Laureto, and A. Fontana, "The oleic acid complexes of proteolytic fragments of α-lactalbumin display apoptotic activity," *The FEBS journal*, vol. 277, no. 1, pp. 163–173, 2010.

[29] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977.

[30] S. Sikdar, J. Chakrabarti, and M. Ghosh, "A microscopic insight from conformational thermodynamics to functional ligand binding in proteins," *Molecular Biosystems*, vol. 10, no. 12, pp. 3280–3289, 2014.

[31] A. G. Moulick and J. Chakrabarti, "Conformational fluctuations in the molten globule state of α-lactalbumin," *Phys. Chem. Chem. Phys.*, vol. 24, pp. 21348–21357, 2022.

[32] M. Jana and S. Bandyopadhyay, "Restricted dynamics of water around a protein–carbohydrate complex: Computer simulation studies," *The Journal of chemical physics*, vol. 137, no. 5, p. 055102, 2012.

[33] J. D. Harper and P. T. Lansbury Jr, "Models of amyloid seeding in alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins," *Annual review of biochemistry*, vol. 66, no. 1, pp. 385–407, 1997.

[34] M. Vendruscolo and C. M. Dobson, "Protein dynamics: Moore's law in molecular biology," *Current Biology*, vol. 21, no. 2, pp. R68–R70, 2011.

[35] J. Trylska, "Coarse-grained models to study dynamics of nanoscale biomolecules and their applications to the ribosome," *Journal of Physics: Condensed Matter*, vol. 22, no. 45, p. 453101, 2010.

[36] M. G. Saunders and G. A. Voth, "Coarse-graining of multiprotein assemblies," *Current opinion in structural biology*, vol. 22, no. 2, pp. 144–150, 2012.

[37] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[38] V. J. Hilser, J. O. Wrabl, and H. N. Motlagh, "Structural and energetic basis of allostery," *Annual review of biophysics*, vol. 41, pp. 585–609, 2012.

[39] D.-W. Li, D. Meng, and R. Bruschweiler, "Short-range coherence of internal protein dynamics revealed by high-precision in silico study," *Journal of the American Chemical Society*, vol. 131, pp. 14610–14611, oct 2009.

[40] J. P. Hansen and I. R. McDonald, "Theory of simple liquids," *Physics Today*, vol. 41, p. 89, 1988.

[41] S. Dutta, M. Ghosh, and J. Chakrabarti, "Spatio-temporal coordination among functional residues in protein," *Scientific Reports*, vol. 7, jan 2017.

[42] J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton, *Protein NMR spectroscopy: principles and practice*. Academic press, 1996.

[43] N. Tjandra and A. Bax, "Direct measurement of distances and angles in biomolecules by nmr in a dilute liquid crystalline medium," *Science*, vol. 278, no. 5340, pp. 1111–1114, 1997.

[44] D. Yang, K. H. Gardner, and L. E. Kay, "A sensitive pulse scheme for measuring the backbone dihedral angle ψ based on cross-correlation between 13 c α-1 hα dipolar and carbonyl chemical shift anisotropy relaxation interactions," *Journal of biomolecular NMR*, vol. 11, no. 2, pp. 213–220, 1998.

[45] D. Yang and L. E. Kay, "Determination of the protein backbone dihedral angle ψ from a combination of nmr-derived cross-correlation spin relaxation rates," *Journal of the American Chemical Society*, vol. 120, no. 38, pp. 9880–9887, 1998.

[46] P. Pelupessy, E. Chiarparin, R. Ghose, and G. Bodenhausen, "Simultaneous determination of ψ and ϕ angles in proteins from measurements of cross-correlated relaxation effects," *Journal of Biomolecular NMR*, vol. 14, no. 3, pp. 277–280, 1999.

[47] E. Chiarparin, P. Pelupessy, R. Ghose, and G. Bodenhausen, "Relaxation of two-spin coherence due to cross-correlated fluctuations of dipole- dipole couplings and anisotropic shifts in nmr of 15n, 13c-labeled biomolecules," *Journal of the American Chemical Society*, vol. 121, no. 29, pp. 6876–6883, 1999.

[48] R. Sprangers, M. Bottomley, J. Linge, J. Schultz, M. Nilges, and M. Sattler, "Refinement of the protein backbone angle ψ in nmr structure calculations," *Journal of biomolecular NMR*, vol. 16, no. 1, pp. 47–58, 2000.

[49] N. R. Skrynnikov, R. Konrat, D. Muhandiram, and L. E. Kay, "Relative orientation of peptide planes in proteins is reflected in carbonyl- carbonyl chemical shift anisotropy cross-correlated spin relaxation," *Journal of the American Chemical Society*, vol. 122, no. 29, pp. 7059–7071, 2000.

[50] J. R. Tolman, H. M. Al-Hashimi, L. E. Kay, and J. H. Prestegard, "Structural and dynamic analysis of residual dipolar coupling data for proteins," *Journal of the American Chemical Society*, vol. 123, no. 7, pp. 1416–1424, 2001.

[51] J. Meiler, J. J. Prompers, W. Peti, C. Griesinger, and R. Brüschweiler, "Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins," *Journal of the American Chemical Society*, vol. 123, no. 25, pp. 6098–6107, 2001.

[52] W. Peti, J. Meiler, R. Brüschweiler, and C. Griesinger, "Model-free analysis of protein backbone motion from residual dipolar couplings," *Journal of the American Chemical Society*, vol. 124, no. 20, pp. 5822–5833, 2002.

[53] B. Vögeli and K. Pervushin, "Trosy experiment for refinement of backbone ψ and φ by simultaneous measurements of cross-correlated relaxation rates and 3, 4 j hαhn coupling constants," *Journal of biomolecular NMR*, vol. 24, no. 4, pp. 291–300, 2002.

[54] K. Kloiber, W. Schüler, and R. Konrat, "Automated nmr determination of protein backbone dihedral angles from cross-correlated spin relaxation," *Journal of biomolecular NMR*, vol. 22, no. 4, pp. 349–363, 2002.

[55] D. Früh, E. Chiarparin, P. Pelupessy, and G. Bodenhausen, "Measurement of long-range cross-correlation rates using a combination of single-and multiple-quantum nmr spectroscopy in one experiment," *Journal of the American Chemical Society*, vol. 124, no. 15, pp. 4050–4057, 2002.

[56] P. Pelupessy, S. Ravindranathan, and G. Bodenhausen, "Correlated motions of successive amide nh bonds in proteins," *Journal of biomolecular NMR*, vol. 25, no. 4, pp. 265–280, 2003.

[57] B. Vögeli, J. Ying, A. Grishaev, and A. Bax, "Limits on variations in protein backbone dynamics from precise measurements of scalar couplings," *Journal of the American Chemical Society*, vol. 129, no. 30, pp. 9377–9385, 2007.

[58] L. Yao, B. Vögeli, D. A. Torchia, and A. Bax, "Simultaneous nmr study of protein structure and dynamics using conservative mutagenesis," *The Journal of Physical Chemistry B*, vol. 112, no. 19, pp. 6045–6056, 2008.

[59] L. E. Kay, "Nmr studies of protein structure and dynamics," *Journal of magnetic resonance*, vol. 213, no. 2, pp. 477–491, 2011.

[60] K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo, "Simultaneous determination of protein structure and dynamics," *Nature*, vol. 433, no. 7022, pp. 128–132, 2005.

[61] G. M. Clore and C. D. Schwieters, "Concordance of residual dipolar couplings, backbone order parameters and crystallographic b-factors for a small α/β protein: a unified picture of high probability, fast atomic motions in proteins," *Journal of molecular biology*, vol. 355, no. 5, pp. 879–886, 2006.

[62] P. R. Markwick, G. Bouvignies, and M. Blackledge, "Exploring multiple timescale motions in protein gb3 using accelerated molecular dynamics and nmr spectroscopy," *Journal of the American Chemical Society*, vol. 129, no. 15, pp. 4724–4730, 2007.

[63] M. Goldman, "Interference effects in the relaxation of a pair of unlike spin-12 nuclei," *Journal of Magnetic Resonance (1969)*, vol. 60, no. 3, pp. 437–452, 1984.

[64] V. A. Daragan and K. H. Mayo, "Motional model analyses of protein and peptide dynamics using 13c and 15n NMR relaxation," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 31, pp. 63–105, jul 1997.

[65] A. Kumar, R. Grace, and P. Madhu, "Prog nucl magn reson spectrosc," *Prog Nucl Magn Reson Spectrosc*, vol. 37, pp. 191–319, 2000.

[66] B. Vogeli and L. Yao, "Correlated dynamics between protein HN and HC bonds observed by NMR cross relaxation," *Journal of the American Chemical Society*, vol. 131, pp. 3668–3678, mar 2009.

[67] B. Vögeli and L. Vugmeyster, "Distance-independent cross-correlated relaxation and isotropic chemical shift modulation in protein dynamics studies," *ChemPhysChem*, vol. 20, no. 2, pp. 178–196, 2019.

[68] R. B. Fenwick and B. Vögeli, "Detection of correlated protein backbone and side-chain angle fluctuations," *ChemBioChem*, vol. 18, no. 20, pp. 2016–2021, 2017.

[69] R. B. Fenwick, C. D. Schwieters, and B. Vogeli, "Direct investigation of slow correlated dynamics in proteins via dipolar interactions," *Journal of the American Chemical Society*, vol. 138, no. 27, pp. 8412–8421, 2016.

[70] C. Kauffmann, I. Ceccolini, G. Kontaxis, and R. Konrat, "Detecting anisotropic segmental dynamics in disordered proteins by cross-correlated spin relaxation," *Magnetic Resonance*, vol. 2, no. 2, pp. 557–569, 2021.

[71] A. G. Moulick and J. Chakrabarti, "Correlation between protein bond vector and dihedral fluctuations," in *AIP Conference Proceedings*, vol. 2265, p. 030036, AIP Publishing LLC, 2020.

[72] A. Villa and G. Stock, "What nmr relaxation can tell us about the internal motion of an rna hairpin: a molecular dynamics simulation study," *Journal of chemical theory and computation*, vol. 2, no. 5, pp. 1228–1236, 2006.

[73] T. S. Ulmer, B. E. Ramirez, F. Delaglio, and A. Bax, "Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal nmr spectroscopy," *Journal of the American Chemical Society*, vol. 125, no. 30, pp. 9179–9191, 2003.

[74] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 åresolution," *Journal of molecular biology*, vol. 194, no. 3, pp. 531–544, 1987.

[75] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 3, pp. 712–725, 2006.

[76] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "Gromacs: fast, flexible, and free," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[77] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of chemical physics*, vol. 79, no. 2, pp. 926–935, 1983.

[78] G. Lipari and A. Szabo, "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity," *Journal of the American Chemical Society*, vol. 104, no. 17, pp. 4546–4559, 1982.

[79] D. M. Korzhnev, I. V. Ibraghimov, M. Billeter, and V. Y. Orekhov, "Munin: application of three-way decomposition to the analysis of heteronuclear nmr relaxation data," *Journal of biomolecular NMR*, vol. 21, no. 3, pp. 263–268, 2001.

[80] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[81] B. Vögeli, "Cross-correlated relaxation rates between protein backbone h–x dipolar interactions," *Journal of biomolecular NMR*, vol. 67, no. 3, pp. 211–232, 2017.

[82] G. Abrusán and J. A. Marsh, "Alpha helices are more robust to mutations than beta strands," *PLOS Computational Biology*, vol. 12, pp. 1–16, 12 2016.

[83] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, *et al.*, "Charmm: the biomolecular simulation program," *Journal of computational chemistry*, vol. 30, no. 10, pp. 1545–1614, 2009.

[84] M. A. González and J. L. Abascal, "The shear viscosity of rigid water models," *The Journal of chemical physics*, vol. 132, no. 9, p. 096101, 2010.

[85] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*. Oxford university press, 2017.

[86] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh ewald method," *The Journal of chemical physics*, vol. 103, no. 19, pp. 8577–8593, 1995.

[87] P. A. Jennings and P. E. Wright, "Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin," *Science*, vol. 262, no. 5135, pp. 892–896, 1993.

[88] J. Balbach, V. Forge, N. A. van Nuland, S. L. Winder, P. J. Hore, and C. M. Dobson, "Following protein folding in real time using nmr spectroscopy," *Nature structural biology*, vol. 2, no. 10, pp. 865–870, 1995.

[89] H. A. Mckenzie and F. H. White Jr, "Lysozyme and α-lactalbumin: structure, function, and interrelationships," *Advances in protein chemistry*, vol. 41, pp. 173–315, 1991.

[90] B. Ramakrishnan, P. S. Shah, and P. K. Qasba, "α-lactalbumin (la) stimulates milk β-1, 4-galactosyltransferase i (β4gal-t1) to transfer glucose from udp-glucose to n-acetylglucosamine: Crystal structure of β4gal-t1· la complex

with udp-glc," *Journal of biological chemistry*, vol. 276, no. 40, pp. 37665–37671, 2001.

[91] Q. Qiao, G. R. Bowman, and X. Huang, "Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation," *Journal of the American Chemical Society*, vol. 135, no. 43, pp. 16092–16101, 2013.

[92] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nature reviews Molecular cell biology*, vol. 16, no. 1, pp. 18–29, 2015.

[93] C. Rischel, P. Thyberg, R. Rigler, and F. M. Poulsen, "Time-resolved fluorescence studies of the molten globule state of apomyoglobin," *Journal of molecular biology*, vol. 257, no. 4, pp. 877–885, 1996.

[94] E. A. Permyakov, V. V. Yarmolenko, L. P. Kalinichenko, L. A. Morozova, and E. A. Burstein, "Calcium binding to α-lactalbumin: Structural rearrangement and association constant evaluation by means of intrinsic protein fluorescence changes," *Biochemical and Biophysical Research Communications*, vol. 100, no. 1, pp. 191–197, 1981.

[95] E. A. Permyakov, L. A. Morozova, and E. A. Burstein, "Cation binding effects on the ph, thermal and urea denaturation transitions in α-lactalbumin," *Biophysical Chemistry*, vol. 21, no. 1, pp. 21–31, 1985.

[96] P. J. Anderson, C. L. Brooks, and L. J. Berliner, "Functional identification of calcium binding residues in bovine α-lactalbumin," *Biochemistry*, vol. 36, no. 39, pp. 11648–11654, 1997.

[97] T. Hendrix, Y. V. Griko, and P. L. Privalov, "A calorimetric study of the influence of calcium on the stability of bovine α-lactalbumin," *Biophysical Chemistry*, vol. 84, no. 1, pp. 27–34, 2000.

[98] Y. V. Griko and D. P. Remeta, "Energetics of solvent and ligand-induced conformational changes in α-lactalbumin," *Protein Science*, vol. 8, no. 3, pp. 554–561, 1999.

[99] O. B. Ptitsyn, "Protein folding: hypotheses and experiments," *Journal of Protein Chemistry*, vol. 6, no. 4, pp. 273–293, 1987.

[100] K. Kuwajima, "The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure," *Proteins: Structure, Function, and Bioinformatics*, vol. 6, no. 2, pp. 87–103, 1989.

[101] C. M. Dobson, P. A. Evans, and S. E. Radford, "Understanding how proteins fold: the lysozyme story so far," *Trends in biochemical sciences*, vol. 19, no. 1, pp. 31–37, 1994.

[102] C. M. Dobson, "Unfolded proteins, compact states and molten globules: Current opinion in structural biology 1992, 2: 6–12," *Current Opinion in Structural Biology*, vol. 2, no. 1, pp. 6–12, 1992.

[103] M. Arai and K. Kuwajima, "Role of the molten globule state in protein folding," *Advances in protein chemistry*, vol. 53, pp. 209–282, 2000.

[104] K. M. Cawthern, M. Narayan, D. Chaudhuri, E. A. Permyakov, and L. J. Berliner, "Interactions of α-lactalbumin with fatty acids and spin label analogs," *Journal of Biological Chemistry*, vol. 272, no. 49, pp. 30812–30816, 1997.

[105] M. Svensson, A.-K. Mossberg, J. Pettersson, S. Linse, and C. Svanborg, "Lipids as cofactors in protein folding: Stereo-specific lipid–protein interactions are required to form hamlet (h uman α-lactalbumin m ade le thal to t umor cells)," *Protein Science*, vol. 12, no. 12, pp. 2805–2814, 2003.

[106] Y.-B. Zhang, W. Wu, and W. Ding, "From hamlet to xamlet: the molecular complex selectively induces cancer cell death," *African Journal of Biotechnology*, vol. 9, no. 54, pp. 9270–9276, 2010.

[107] K. Kuwajima, Y. Hiraoka, M. Ikeguchi, and S. Sugai, "Comparison of the transient folding intermediates in lysozyme and. alpha.-lactalbumin," *Biochemistry*, vol. 24, no. 4, pp. 874–881, 1985.

[108] B. A. Schulman, C. Redfield, Z.-y. Peng, C. M. Dobson, and P. S. Kim, "Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human α-lactalbumin," 1995.

[109] L. J. Smith, C. M. Dobson, and W. F. van Gunsteren, "Molecular dynamics simulations of human α-lactalbumin: Changes to the structural and dynamical properties of the protein at low ph," *Proteins: Structure, Function, and Bioinformatics*, vol. 36, no. 1, pp. 77–86, 1999.

[110] M. Shimizu, Y. Kajikawa, K. Kuwajima, C. M. Dobson, and Y. Okamoto, "Determination of the structural ensemble of the molten globule state of a protein by computer simulations," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 8, pp. 635–645, 2019.

[111] J. Mongan, D. A. Case, and J. A. McCammon, "Constant ph molecular dynamics in generalized born implicit solvent," *Journal of computational chemistry*, vol. 25, no. 16, pp. 2038–2048, 2004.

[112] R. R. Ernst, G. Bodenhausen, and A. Wokaun, *Principles of nuclear magnetic resonance in one and two dimensions*. No. BOOK, 1987.

[113] E. D. Chrysina, K. Brew, and K. R. Acharya, "Crystal structures of apo-and holo-bovine α-lactalbumin at 2.2-å resolution reveal an effect of calcium on inter-lobe interactions," *Journal of Biological Chemistry*, vol. 275, no. 47, pp. 37021–37029, 2000.

[114] J. M. Swails, D. M. York, and A. E. Roitberg, "Constant ph replica exchange molecular dynamics in explicit solvent using discrete protonation states: implementation, testing, and validation," *Journal of chemical theory and computation*, vol. 10, no. 3, pp. 1341–1352, 2014.

[115] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The amber biomolecular simulation programs," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.

[116] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of computational physics*, vol. 23, no. 3, pp. 327–341, 1977.

[117] D. R. Roe and T. E. Cheatham III, "Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data," *Journal of chemical theory and computation*, vol. 9, no. 7, pp. 3084–3095, 2013.

[118] P. Chatterjee, S. Bagchi, and N. Sengupta, "The non-uniform early structural response of globular proteins to cold denaturing conditions: A case study with yfh1," *The Journal of Chemical Physics*, vol. 141, no. 20, p. 11B615_1, 2014.

[119] S. Swaminathan, W. Harte Jr, and D. L. Beveridge, "Investigation of domain structure in proteins via molecular dynamics simulation: application to hiv-1 protease dimer," *Journal of the American Chemical Society*, vol. 113, no. 7, pp. 2717–2721, 1991.

[120] J. Luo and T. C. Bruice, "Ten-nanosecond molecular dynamics simulation of the motions of the horse liver alcohol dehydrogenase phch2o- complex," *Proceedings of the National Academy of Sciences*, vol. 99, no. 26, pp. 16597–16600, 2002.

[121] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. Caves, "Bio3d: an r package for the comparative analysis of protein structures," *Bioinformatics*, vol. 22, no. 21, pp. 2695–2696, 2006.

[122] M. Kataoka, F. Tokunaga, K. Kuwajima, and Y. Goto, "Structural characterization of the molten globule of α-lactalbumin by solution x-ray scattering," *Protein Science*, vol. 6, no. 2, pp. 422–430, 1997.

[123] K. Gast, D. Zirwer, H. Welfle, V. Bychkova, and O. Ptitsyn, "Quasielastic light scattering from human α-lactalbumin: comparison of molecular dimensions in native and 'molten globule'states," *International Journal of Biological Macromolecules*, vol. 8, no. 4, pp. 231–236, 1986.

[124] C. Redfield, B. A. Schulman, M. A. Milhollen, P. S. Kim, and C. M. Dobson, "α-lactalbumin forms a compact molten globule in the absence of disulfide bonds," *nature structural biology*, vol. 6, no. 10, pp. 948–952, 1999.

[125] E. Paci, L. J. Smith, C. M. Dobson, and M. Karplus, "Exploration of partially unfolded states of human α-lactalbumin by molecular dynamics simulation," *Journal of molecular biology*, vol. 306, no. 2, pp. 329–347, 2001.

[126] D. Dolgikh, R. Gilmanshin, E. Brazhnikov, V. Bychkova, G. Semisotnov, S. Y. Venyaminov, and O. Ptitsyn, "α-lactalbumin: compact state with fluctuating tertiary structure?," *FEBS letters*, vol. 136, no. 2, pp. 311–315, 1981.

[127] J. Baum, C. M. Dobson, P. A. Evans, and C. Hanley, "Characterization of a partly folded protein by nmr methods: studies on the molten globule state of guinea pig. alpha.-lactalbumin," *Biochemistry*, vol. 28, no. 1, pp. 7–13, 1989.

[128] A. K. Lala and P. Kaul, "Increased exposure of hydrophobic surface in molten globule state of alpha-lactalbumin. fluorescence and hydrophobic photolabeling studies.," *Journal of Biological Chemistry*, vol. 267, no. 28, pp. 19914–19918, 1992.

[129] S. Improta, H. Molinari, A. Pastore, R. Consonni, and L. Zetta, "Probing protein structure by solvent perturbation of nmr spectra: Photochemically induced dynamic nuclear polarization and paramagnetic perturbation techniques applied to the study of the molten globule state of α-lactalbumin," *European journal of biochemistry*, vol. 227, no. 1-2, pp. 87–96, 1995.

[130] A. G. Moulick and J. Chakrabarti, "Correlated dipolar and dihedral fluctuations in a protein," *Chemical Physics Letters*, p. 139574, 2022.

[131] B. Spolaore, O. Pinato, M. Canton, M. Zambonin, P. Polverino de Laureto, and A. Fontana, "α-lactalbumin forms with oleic acid a high molecular weight complex displaying cytotoxic activity," *Biochemistry*, vol. 49, no. 39, pp. 8658–8667, 2010.

[132] Y. Fu, V. Kasinath, V. R. Moorman, N. V. Nucci, V. J. Hilser, and A. J. Wand, "Coupled motion in proteins revealed by pressure perturbation," *Journal of the American Chemical Society*, vol. 134, no. 20, pp. 8543–8550, 2012.

[133] N. Smolin, R. Biehl, G. Kneller, D. Richter, and J. C. Smith, "Functional domain motions in proteins on the 1–100 ns timescale: comparison of neutron spin-echo spectroscopy of phosphoglycerate kinase with molecular-dynamics simulation," *Biophysical journal*, vol. 102, no. 5, pp. 1108–1117, 2012.

[134] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[135] V. N. Uversky, "Intrinsic disorder, protein–protein interactions, and disease," *Advances in protein chemistry and structural biology*, vol. 110, pp. 85–121, 2018.

[136] P. E. Wright and H. J. Dyson, "Linking folding and binding," *Current opinion in structural biology*, vol. 19, no. 1, pp. 31–38, 2009.

[137] C. Svanborg, H. Ågerstam, A. Aronson, R. Bjerkvig, C. Düringer, W. Fischer, L. Gustafsson, O. Hallgren, I. Leijonhuvud, S. Linse, *et al.*, "Hamlet kills tumor cells by an apoptosis-like mechanism—cellular, molecular, and therapeutic aspects," *Advances in cancer research*, vol. 88, pp. 1–29, 2003.

[138] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, "Conformational entropy in molecular recognition by proteins," *Nature*, vol. 448, no. 7151, pp. 325–329, 2007.

[139] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[140] G. Van Zundert, J. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A. Melquiond, M. van Dijk, S. De Vries, and A. Bonvin, "The haddock2. 2 web server: user-friendly integrative modeling of biomolecular complexes," *Journal of molecular biology*, vol. 428, no. 4, pp. 720–725, 2016.

[141] R. V. Honorato, P. I. Koukos, B. Jiménez-García, A. Tsaregorodtsev, M. Verlato, A. Giachetti, A. Rosato, and A. M. Bonvin, "Structural biology in the clouds: the wenmr-eosc ecosystem," *Frontiers in Molecular Biosciences*, vol. 8, p. 729513, 2021.

[142] K. Sprenger, V. W. Jaeger, and J. Pfaendtner, "The general amber force field (gaff) can accurately predict thermodynamic and transport properties of many ionic liquids," *The Journal of Physical Chemistry B*, vol. 119, no. 18, pp. 5882–5895, 2015.

[143] A. W. Sousa da Silva and W. F. Vranken, "Acpype-antechamber python parser interface," *BMC research notes*, vol. 5, pp. 1–8, 2012.

[144] J. S. Hub, B. L. De Groot, and D. van der Spoel, "g_wham a free weighted histogram analysis implementation including robust error and autocorrelation estimates," *Journal of chemical theory and computation*, vol. 6, no. 12, pp. 3713–3720, 2010.

[145] S. G. Dastidar and C. Mukhopadhyay, "Structure, dynamics, and energetics of water at the surface of a small globular protein: a molecular dynamics simulation," *Physical Review E*, vol. 68, no. 2, p. 021921, 2003.

[146] S. Samanta and S. Mukherjee, "Deciphering complex dynamics of water counteraction around secondary structural elements of allosteric protein complex: Case study of sap-slam system in signal transduction cascade," *The Journal of Chemical Physics*, vol. 148, no. 4, p. 045102, 2018.

[147] K. Chakraborty and S. Bandyopadhyay, "Correlated conformational motions of the kh domains of far upstream element binding protein complexed with single-stranded dna oligomers," *The Journal of Physical Chemistry B*, vol. 119, no. 34, pp. 10998–11009, 2015.

[148] J. Kästner, "Umbrella sampling," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 6, pp. 932–942, 2011.

[149] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, "Biomolecular condensates: organizers of cellular biochemistry," *Nature reviews Molecular cell biology*, vol. 18, no. 5, pp. 285–298, 2017.

[150] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, "Coarse-grained protein models and their applications," *Chemical reviews*, vol. 116, no. 14, pp. 7898–7936, 2016.

[151] K. A. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, vol. 24, no. 6, pp. 1501–1509, 1985.

[152] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," *Macromolecules*, vol. 22, no. 10, pp. 3986–3997, 1989.

[153] K. A. Dill, K. M. Fiebig, and H. S. Chan, "Cooperativity in protein-folding kinetics.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 5, pp. 1942–1946, 1993.

[154] A. C. Farris, G. Shi, T. Wüst, and D. P. Landau, "The role of chain-stiffness in lattice protein models: A replica-exchange wang-landau study," *The Journal of Chemical Physics*, vol. 149, no. 12, p. 125101, 2018.

[155] A. Mukherjee and B. Bagchi, "Correlation between rate of folding, energy landscape, and topology in the folding of a model protein hp-36," *The Journal of chemical physics*, vol. 118, no. 10, pp. 4733–4747, 2003.

[156] M. Schärpf, H. Sticht, K. Schweimer, M. Boehm, S. Hoffmann, and P. Rösch, "Antitermination in bacteriophage λ: The structure of the n36 peptide-boxb

rna complex," *European Journal of Biochemistry*, vol. 267, no. 8, pp. 2397–2408, 2000.

[157] I. Gerroff, A. Milchev, K. Binder, and W. Paul, "A new off-lattice monte carlo model for polymers: A comparison of static and dynamic properties with the bond-fluctuation model and application to random media," *The Journal of chemical physics*, vol. 98, no. 8, pp. 6526–6539, 1993.

[158] A. L. Lehninger, D. L. Nelson, M. M. Cox, *et al.*, *Lehninger principles of biochemistry*. Macmillan, 2005.