Prakash Saudagar ·
Timir Tripathi *Editors*

Protein Folding Dynamics and Stability

Experimental and Computational Methods

Springer

Protein Folding Dynamics and Stability

Prakash Saudagar • Timir Tripathi
Editors

Protein Folding Dynamics and Stability

Experimental and Computational Methods

Springer

Editors

Prakash Saudagar
Department of Biotechnology
National Institute of Technology
Warangal, Telangana, India

Timir Tripathi
Department of Biochemistry
North-Eastern Hill University
Shillong, Meghalaya, India

ISBN 978-981-99-2078-5 ISBN 978-981-99-2079-2 (eBook)
https://doi.org/10.1007/978-981-99-2079-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The biological roles of most proteins are determined by their three-dimensional structure, while their interaction with substrates, cofactors, other proteins, and biomolecules is determined by their conformation and dynamics. In rational structure-based drug discovery, small compounds are identified and developed to interact selectively with the target protein and modulate its function. Thus, an atomic-level understanding of the protein structure, folding, and conformational dynamics is essential. With several advancements in the field in the last few decades, researchers have developed sensitive and accurate methods to study and characterize these interactions and folding and, eventually, determine the structure–function relationship of proteins. Spectroscopy and calorimetry are techniques that have strengthened their roots in protein science. Equipped with the optics and rapid sensors, spectroscopy has helped understand proteins' dynamics and transient interactions, which otherwise would have been highly challenging. Coupled with spectroscopic methods, computational algorithms have also been utilized to provide valuable information at the atomic level with the advantage of time. With this book, we introduce readers to state-of-the-art advancements in the field of spectroscopic and calorimetric techniques and how they have been integrated with computational methods to study protein folding, interaction, and dynamics.

The book comprises 13 chapters dedicated to understanding protein dynamics using spectroscopic and computational tools. Chapter "Applications of Circular Dichroism Spectroscopy in Studying Protein Folding, Stability, and Interaction" focuses on studying protein folding, stability, and interactions using circular dichroism spectroscopy. Chapter "Fluorescence Spectroscopy-Based Methods to Study Protein Folding Dynamics" concentrates on fluorescence spectroscopy to study protein folding dynamics with special attention to amyloid fibril aggregation in Alzheimer's and Parkinson's disease. Chapter "Applications of Differential Scanning Calorimetry in Studying Folding and Stability of Proteins" presents an understanding of the folding and stability of proteins with the help of differential scanning calorimetry. Chapter "Nuclear Magnetic Resonance Spectroscopy to Analyze Protein Folding and Dynamics" discusses the use of nuclear magnetic resonance for protein folding and dynamics. The following five chapters introduce the readers to different computational methods available to decipher protein folding, interactions, and dynamics. Chapter “Molecular Dynamics Simulation Methods to Study Protein Structural Dynamics” familiarizes the molecular dynamics simulation methods to study the structural dynamics of a protein. To study the dynamic interactions between proteins and ligands, chapter “Molecular Dynamics Simulation to Study Protein Conformation and Ligand Interaction” discusses the use of molecular dynamics simulation to study ligand-based transitions in protein folding. Chapter “Monte Carlo Approaches to Study Protein Conformation Ensembles” analyses the use of Monte Carlo approaches to understand protein dynamics. The employment of Markov state models of molecular simulations to study protein folding and dynamics is detailed in chapter “Markov State Models of Molecular Simulations to Study Protein Folding and Dynamics”. The enhanced sampling and free energy methods are efficient and advanced methods described in chapter “Enhanced Sampling and Free Energy Methods to Study Protein Folding and Dynamics.” After this, the following four chapters discuss how spectroscopic methods and computational tools can be integrated to study protein folding and dynamics. The use of chaotropic agents to investigate the protein unfolding and stability using spectroscopic methods and molecular dynamics simulations is presented in chapter “Investigating Protein Unfolding and Stability Using Chaotropic Agents and Molecular Dynamics Simulation.” Chapters “pH-Based Molecular Dynamics Simulation for Analyzing Protein Structure and Folding” and “Molecular Dynamics Simulation to Study Thermal Unfolding in Proteins” discuss the methods available to computationally study the effect of pH and temperature on protein structure and folding. Finally, chapter “Principles, Methods, and Applications of Protein Folding Inside Cells” provides an insight into the principles, methods, and application of protein folding inside the cells.

The book introduces the readers to the use of both biophysical and computational tools to analyze protein folding, stability, and dynamics. The chapters result from meticulous research and critical contribution from eminent researchers of several decades. As editors, we believe that students, faculties, and researchers will find this book comprehensive, resourceful, and practical for the field of protein science. Finally, we thank all the authors and contributors for their time and effort in bringing out this book.

Warangal, India Prakash Saudagar
Shillong, India Timir Tripathi

Contents

Applications of Circular Dichroism Spectroscopy in Studying Protein Folding, Stability, and Interaction 1
Preeti Gupta, Asimul Islam, Faizan Ahmad, and Md Imtaiyaz Hassan

Fluorescence Spectroscopy-Based Methods to Study Protein Folding Dynamics 25
Ritesh Kumar, Timir Tripathi, and Prakash Saudagar

Applications of Differential Scanning Calorimetry in Studying Folding and Stability of Proteins 37
Banesh Sooram, Neharika Gupta, Vihadhar Reddy Chethireddy,
Timir Tripathi, and Prakash Saudagar

Nuclear Magnetic Resonance Spectroscopy to Analyse Protein Folding and Dynamics 61
Nikita V. Saibo, Soumendu Boral, Rituparna Saha, Amit K. Das,
and Soumya De

Molecular Dynamics Simulation Methods to Study Structural Dynamics of Proteins 83
Anil Kumar and Krishna Kumar Ojha

Molecular Dynamics Simulation to Study Protein Conformation and Ligand Interaction 107
Santanu Sasidharan, Vijayakumar Gosu, Timir Tripathi,
and Prakash Saudagar

Monte Carlo Approaches to Study Protein Conformation Ensembles . . . 129
Nidhi Awasthi, Rohit Shukla, Devesh Kumar, Arvind Kumar Tiwari,
and Timir Tripathi

Markov State Models of Molecular Simulations to Study Protein Folding and Dynamics 147
Vivek Junghare, Sourya Bhattacharya, Khalid Ansari, and Saugata Hazra

Enhanced Sampling and Free Energy Methods to Study Protein Folding and Dynamics 165
Muthuraja Arun Pravin and Sanjeev Kumar Singh

Investigating Protein Unfolding and Stability Using Chaotropic Agents and Molecular Dynamics Simulation 181
Rohit Shukla and Timir Tripathi

pH-Based Molecular Dynamics Simulation for Analysing Protein Structure and Folding 203
Santanu Sasidharan, Rohit Shukla, Timir Tripathi, and Prakash Saudagar

Molecular Dynamics Simulation to Study Thermal Unfolding in Proteins 221
Md Imtaiyaz Hassan, Mohd. Umair, Yash Mathur, Taj Mohammad, Afreen Khan, Md Nayab Sulaimani, Afsar Alam, and Asimul Islam

Principles, Methods, and Applications of Protein Folding Inside Cells . . 251
Subhashree Sahoo, Kummari Shivani, Amrita Arpita Padhy, Varsha Kumari, and Parul Mishra

About the Editors

Prakash Saudagar is an active researcher and a sterling classroom teacher, currently working as an Associate Professor at the Department of Biochemistry, National Institute of Technology Warangal, India. He obtained his Ph.D. from the Indian Institute of Technology, Guwahati, India, in 2013. His research has immensely contributed to exploring potential drug target proteins and inhibitors. His research interests include molecular and biochemical parasitology, infectious disease, and protein biochemistry. He has a strong command over computational and in vitro techniques used to study proteins. He has published more than 40 research articles in reputed journals like the International Journal of Biological Molecules, FEBS, FEBS OpenBio, PLOS one, Scientific Reports, Biological Chemistry, Parasitology International, Molecular Simulation, etc., and book chapters in Elsevier and Springer to his name. He has been PI/Co-PI in research grants from SERB and DST. He has been awarded the B.S. Narasinga Rao award by SBC, India (2011), Best Presentation Award, ICIDN, Nepal (2015), Young Faculty Award, VIF India (2016), and Young Scientist Award, Telangana Academy of Science (2018). He is an associate fellow of the Telangana Academy of Science (2018) and a life member of the Indian Science Congress and Society of Biological Chemists. He has guided several Ph.D. and M.Tech. students, postdoctoral fellows, project fellows, and trainees. He has many interdisciplinary collaborating partners in prestigious institutions in India and abroad.

Timir Tripathi is the Regional Director of Indira Gandhi National Open University (IGNOU), Regional Centre Kohima, Nagaland, India. Earlier, he served as a Senior Assistant Professor and Principal Investigator at the Department of Biochemistry, North-Eastern Hill University, Shillong, India. He holds a Ph.D. from the Central Drug Research Institute, Lucknow, India. He was a visiting faculty at ICGEB, New Delhi, India, and Khon Kaen University, Thailand. He is known for his research in the fields of protein biophysics, biochemistry, structural biology, and drug discovery. He has over 18 years of experience in teaching and research related to protein structure, function, and dynamics at the post-graduate and doctoral levels. He has developed and improved methods to investigate and analyze proteins. His research areas include protein interaction dynamics and understanding the roles of non-catalytic domains in regulating the catalytic activity of proteins. He has handled 12 research projects as a principal investigator from various national and international funding agencies. He is an elected member of the National Academy of Sciences, India, and the Royal Society of Biology, UK. He has published over 100 research papers, reviews, commentaries, viewpoints, and editorial articles in international journals and published several book chapters. He has edited five books and authored a textbook on spectroscopic methods for UG and PG students. Currently, he serves as the editor of the International Journal of Biological Macromolecules and is the editorial board member of Scientific Reports, PLoS One, and Acta Tropica.

Applications of Circular Dichroism Spectroscopy in Studying Protein Folding, Stability, and Interaction

Preeti Gupta, Asimul Islam, Faizan Ahmad, and Md Imtaiyaz Hassan

Abstract Circular dichroism (CD) spectroscopy has been extensively used to determine the structure and folding of proteins. It provides valuable information about the protein folding phenomenon, especially the molten globule or other intermediates of the folding/unfolding pathway. This technique is beneficial in characterizing protein obtained via recombinant techniques or isolated from tissues. In addition, the effect of mutations on the folding and conformational stability of the protein can be readily assessed using CD spectroscopy. Unlike X-ray crystallography and NMR spectroscopy, the two primary powerful structure determination techniques, the ease and the requirement of low protein concentrations, make CD spectroscopy a desirable and demanding method of choice. This chapter discusses applications of CD spectroscopy in measuring protein structure and stability. The CD spectroscopic investigation of conformational changes and protein stability studied through steady-state and time-resolved CD measurements have been further highlighted. This chapter will provide a better understanding of CD spectroscopy and its uses in biomolecular studies.

Keywords Protein folding · Protein stability · Circular dichroism · Spectroscopy · Folding intermediate

1 Introduction

The characterization of recombinant proteins provides valuable information about their structure, proper folding, and stability which is invaluable for fundamental research and biopharmaceutical industries. There are many frequently used techniques to monitor the conformational changes and stability of proteins in solution,

P. Gupta · A. Islam · M. I. Hassan (✉)
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: mihassan@jmi.ac.in

F. Ahmad
Department of Biochemistry, Jamia Hamdard, New Delhi, India

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_1

Table 1 A comparison of CD spectroscopy, NMR spectroscopy, and X-ray crystallography

CD	NMR	X-ray crystallography
Measurements can be performed in solution	Measurements can be performed in solution	High-quality pure protein crystals are required for structure determination
0.05–1 mg/mL unlabeled protein is required depending upon the cuvette size	Structure determination typically requires a protein (labeled) concentration of 0.5 mM or greater	Approximately 10 mg of pure protein is needed to get crystals, though as little as 1 mg may now be sufficient for some proteins
Easy sample preparation	Difficult sample preparation and need for high sample purity	Difficulty in crystallizing some proteins
Molecules of any size can be studied	There is a size limitation (MWs below 40–50 kDa)	No size limitation
Does not give residue-specific information	Provide residue-specific information	Provide residue-specific information; however, protein dynamic study is not possible with the crystal
	Investigation of protein dynamics in solution is possible	Direct determination of secondary structures and especially domain movements is not possible
	Provide information on a kinetic basis, such as the internal movement of proteins over multiple time scales and their binding mechanism to ligands	Examination of small parts in the molecule is difficult
Does not provide atomic-level details	Very powerful for atomic-level structural analysis in solution	Provide atomic-level structural details in the crystalline state

such as differential scanning calorimetry (DSC), fluorescence spectroscopy, circular dichroism (CD) spectroscopy, nuclear magnetic resonance (NMR), spectroscopy, and X-ray crystallography [1–11]. Although widely used, all those techniques mentioned above have certain limitations, and hence one method is seldom enough for a detailed study of complex protein characterization. DSC requires high protein concentrations, which are challenging to achieve in the case of aggregation-prone proteins and at a bulk manufacturing scale in industries [12]. On the other hand, fluorescence spectroscopy relies on the presence of intrinsic fluorescence, making it an inefficacious technique to study proteins wherein the presence of a prosthetic group (covalently or non-covalently bound to the protein) in the close vicinity of fluorophore quenches the protein's intrinsic fluorescence. Moreover, it provides no detailed information about the global folding of the protein but instead gives an idea about the local conformational changes around the fluorophores [13, 14]. However, there are certain limitations associated with NMR and X-ray crystallography in the structural determination of proteins. The ease of CD measurements and the requirement of low protein concentrations make it a demanding technique in structural biology [7, 8]. Table 1 highlights the comparison between CD, NMR, and X-ray crystallography. This chapter aims to discuss the uses of CD spectroscopy to obtain insights into the secondary/tertiary structures and stability/conformational dynamics of proteins. We also highlight methodological approaches in performing the CD method and data analysis tools in detail.

2 Determination of Secondary and Tertiary Structures of Proteins Using CD Spectroscopy

Far-UV CD and UV/Vis CD are absorption spectroscopy techniques that investigate the secondary structures of proteins and charge-transfer transitions in metal–protein complexes, respectively. The near-infrared CD is used to study geometric and electronic structures by probing metal $d \rightarrow d$ transitions, while the vibrational CD is used for structural studies of small organic molecules, proteins, and DNA [15]. CD utilizes the differential absorption of the right-handed and left-handed components of the circularly polarized light by the chiral molecules to study their structural aspects. The difference in the absorption of the left-handed and right-handed circularly polarized light is measured and quantified in CD experiments [16, 17]. The homochirality of amino acids imparts chirality to proteins [18]. All amino acids (except glycine) carry at least one chiral center at C_α; threonine and isoleucine have an additional chiral center at C_β [19, 20]. The CD signal is observed when a chromophore is optically active (chiral) either (1) intrinsically by its structure, (2) by being covalently linked to a chiral center, or (3) by being placed in an asymmetric environment [8]. CD is widely used to rapidly determine the secondary and tertiary structure of proteins. The CD spectrum is divided into three wavelength regions based on the electronic transitions that predominate in the given wavelength range (Fig. 1). These include:

1. The far-UV range (190–250 nm), where the contribution from peptide bonds dominates, is used to determine the secondary structure of proteins. A weak but broad $n \rightarrow \pi^*$ transition region is present around 220 nm, and a stronger and sharper $\pi \rightarrow \pi^*$ transition is centered at around 190 nm.
2. The near-UV range (250–300 nm), where the aromatic side chains contribute significantly, gives details about the tertiary structure of proteins.

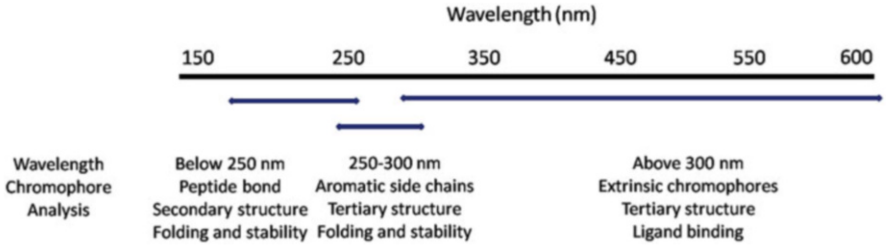

Fig. 1 CD spectral regions in proteins with their respective contributing chromophores

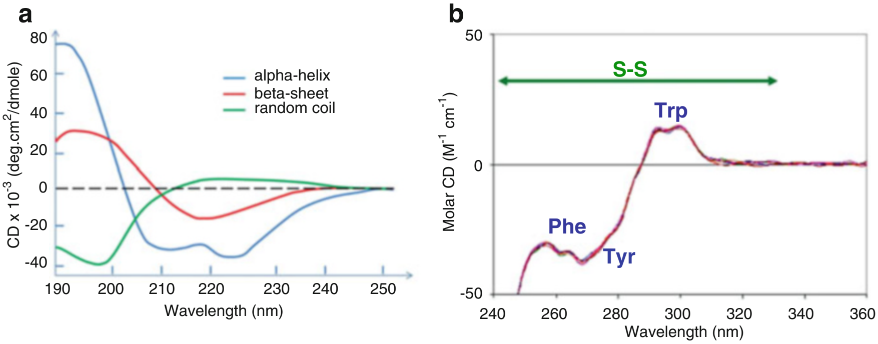

Fig. 2 (**a**) The standard far-UV spectra are associated with different secondary structures in proteins. Adapted with permission from Corrêa et al. [24]. (**b**) The near-UV CD spectra of IgG monoclonal antibody. The characteristic peaks corresponding to Trp, Tyr, and Phe signals are shown. Source: https://www.chiralabsxl.com/Circular_Dichroism/CD_App_Protein_NUV.html

3. The near UV-visible range (300–700 nm), where the extrinsic chromophores contribute, is used to monitor metal ion protein interactions [8, 21, 22].

Due to "exciton" interactions, the optical transitions of the chromophores of the polypeptide chain get split into multiple transitions when aligned in arrays. This gives characteristic CD spectra of different structural elements in the protein [6, 23]. For instance, α-helix rich proteins show two negative bands at 222 and 208 nm of comparable magnitude and a strong positive band close to 193 nm. Proteins dominated with antiparallel β-pleated sheet structure have a negative band at 218 nm and a positive band at 195 nm. Disordered proteins rich in random coil structures show a strong negative band near 195 nm [6, 24, 25]. The far-UV CD spectra of various secondary structural elements in proteins are shown in Fig. 2a.

Phenylalanine (Phe), tyrosine (Tyr), tryptophan (Trp), and disulfide bonds contribute to the near-UV CD of proteins in the wavelength region 250–300 nm. This is the region in which these chromophores absorb. The denatured protein has a weak CD signal. However, if these chromophores are buried in the folded native protein, they give strong CD signals. The intensity of the CD signal of each chromophore depends on how tightly it is held in the asymmetric environment. The near-UV CD spectrum of proteins cannot be interpreted in terms of protein structure, unlike the far-UV CD spectrum is interpreted in terms of secondary structure. However, detailed studies can decompose the CD spectrum into bands attributed to different chromophores. For instance, Trp exhibits a fine-structured peak between 290 and 305 nm. Tyr displays a peak in the range of 275–285 nm, while Phe shows a weak but intense peak at 255–270 nm. These characteristic peaks of amino acid residues emerge due to the vibronic transitions occurring in different vibrational levels of the excited state [8, 16, 26, 27]. The local tertiary structure of the protein can be used for quality control as it often reveals subtle changes from batch to batch not reflected in the far-UV region. Disulfide bonds also contribute to the CD spectrum in the near-UV region [28, 29].

Simple proteins (i.e., proteins devoid of any prosthetic group) do not absorb above 300 nm, and hence they do not exhibit CD signals in wavelengths above 300 nm. However, many prosthetic groups (non-protein chromophores or extrinsic chromophores), including flavins, pyridoxal, and heme moieties, absorb above 300 nm. In the free state, extrinsic chromophores are either achiral or present as enantiomeric mixtures, so they do not show any optical activity. However, upon interaction with the chiral environment of the protein, they generate optical activity [7, 8]. The heme group is a classic example that shows no CD signal alone but exhibits a strong positive band (Soret band) with a wavelength maximum of 412 nm when incorporated in the apoprotein of hemoglobin and myoglobin. The interaction between heme moiety and aromatic residues of the protein is thought to be the reason for heme chirality and hence the CD signal in the Soret region [30].

2.1 Servers to Estimate the Secondary Structure of Proteins from CD Data

Various web servers are available to estimate the secondary structure of the protein from CD spectroscopic data, including DichroWeb [31], BeStSel [32], and K2D3 [33]. They usually employ reference datasets consisting of a set of proteins with known structures to calculate secondary structure information that best matches the experimental (query) spectrum. The CD contribution at each wavelength is weighted, providing the correct secondary structure of the protein as the output. These servers use a range of deconvolution methods, including the simple least square method and more complicated singular value deconvolution and ridge regression method. Generally, the more diverse the components in the reference database are, the more accurate the estimation of secondary structure elements in the query spectrum [34]. It must be noted that the specialized datasets specifically designed for the integral membrane proteins are to be used for their analysis as they tend to have transitions at somewhere different wavelengths compared to soluble proteins [35].

2.1.1 DichroWeb

DichroWeb (http://dichroweb.cryst.bbk.ac.uk) is a freely available web server for determining the secondary structure of a protein based on CD and SRCD spectra. The server facilitates analyses utilizing five different algorithms, including CDSSTR, SELCON3, and VARSLEC (SVD methods with variable selection functions), CONTINLL (a regression restraint method), and K2D (a neural network method now replaced by the stand-alone K2D3 method). The server accepts data in a wide range of formats, including those output from both CD and SRCD instruments, and uses seven reference databases for structure prediction depending upon the protein to be analyzed. It generates an output file containing calculated secondary structures, a tabular and graphical display of experimental, calculated, and difference spectra, and a goodness-of-fit parameter (normalized root mean squared deviation or NRMSD) for the analyses [34].

NRMSD is an important parameter that tells us about the correspondence between the experimental and calculated spectra and is thus used to judge the accuracy of the results. It is important to note that a low value of NRMSD is required but is insufficient to conclude the correctness of the result obtained from the analysis [34, 36]. DichroWeb highlights the importance of precise protein concentration and path length and requires the input data to be down to at least 190 nm and properly subtracted baselines before submission for accurate analysis. It also emphasizes that the best NRMSD is not always the precise solution and that the reference databases do not work well for peptides and membrane proteins [36].

2.1.2 BeStSel

BeStSel server (https://bestsel.elte.hu/index.php) is explicitly designed to analyze β-sheet rich proteins. However, it can be utilized for structural analysis of any protein class, including membrane proteins, amyloid fibrils, and protein aggregates. A comprehensive structural analysis of different secondary structure elements that includes parallel and antiparallel β-sheets and three types of twists, viz. left-handed, relaxed, and right-handed twisted sheets, is performed by the BeStSel server. Based on the structural analysis, it speculates the protein fold down to the topology level organization of the CATH protein fold database [32, 34]. Although BeStSel provides precise secondary structure estimation for a wide range of proteins, the analysis of some special structure types is unsuitable for this server. Such structures include polyproline-II helix (a characteristic structure of collagen-like fibrillar proteins), 3_{10}-helices (present in high amounts in some globular proteins), and various types of turns that are the major structural components of short peptides [32]. Also, BeStSel produces large RMSD values for intrinsically disordered proteins (IDPs) and is not a useful tool for studying this class of proteins [37].

2.1.3 K2D3

The K2D3 server (http://cbdm-01.zdv.uni-mainz.de/~andrade/k2d3/) is based on a neural network approach and a successor to the K2D method. The theoretical CD spectra of a non-redundant set of structures representing most proteins in the PDB are calculated using DichroCalc (https://comp.chem.nottingham.ac.uk/dichrocalc/). These theoretical CD spectra then serve as a reference dataset which is directly applied to predict protein secondary structure. Using the most similar CD spectra in the reference database and weighing their distances from the query spectrum, a predicted CD spectrum is generated. The output contains the query spectrum overlaid on the back-calculated spectrum along with the estimated values of α-helix and β-sheet. No parameter that depicts the fit quality is presented in the K2D3 server. However, if the distance between the query spectrum and the most similar spectrum in the database is greater than a threshold value, a warning signal is displayed [33].

3 Determination of Conformational Changes in the Protein Using CD Spectroscopy

The conformation and structural stability are key determinants of the physiological functions of proteins. Structural perturbation in protein is one of the main reasons for the onset and progression of several diseases, including neurodegenerative disorders and cancer. For example, misfolding of α-synuclein and amyloid β leads to protein oligomerization and fibrillation, resulting in Parkinson's and Alzheimer's diseases, respectively [38, 39]. The structural alteration in the prion protein in a cell membrane environment with subsequent deposition of amyloid plaques is known to cause prion disease [40]. Genomic instability that results from mutations in crucial genes is a hallmark of almost all cancers [41]. The phenotypic outcomes of mutations on proteins include activity, binding mode and interactions, complex stability, and turnover rate.

Proteins bind to their specific targets in a precise manner, and the specificity of these interactions is predominantly defined by the structural and physicochemical properties of binding interfaces [42–44]. Any structural alteration in the protein due to genetic mutation disrupts the binding with the intracellular target, hindering the functionality of the protein. For instance, missense mutations in the BRCT domain inhibit the ability of BRCA1 for substrate recognition. Consequently, the functional role of BRCA1 in the DNA damage repair pathway is hindered and responsible for most hereditary breast and ovarian cancer cases [45]. Protein misfolding/unfolding and degradation also play crucial roles in developing lung diseases, particularly COPD (chronic obstructive pulmonary disease) and idiopathic pulmonary fibrosis and their associated clinical complications [46, 47].

Protein denaturation and aggregation are major problems during manufacturing, storage, and transport in biotechnological and pharmaceutical industries [48]. For instance, therapeutic proteins like antibodies and insulin denature in the bulk solution or at different interfaces during mass production in pharmaceutical companies. The functionality of a protein in the physiological environment or industrial applications is highly dependent on its native conformation. Thus, it is imperative to monitor conformational changes in the protein due to mutations, pH fluctuations, heat, denaturants, or binding interactions with ligands and analyze their functional consequences. CD is a reliable and convenient spectroscopic technique to detect conformational changes in the protein. Moreover, the time dependence of protein structural changes can be determined using the time-resolved CD measurements. CD is also essential in studying peptides that are not feasible by X-ray crystallography

[7]. A classic example of such a study is the switching between α-helix and β-sheet structures in prion peptides [49].

4 Analyzing the Conformational Changes in a Polypeptide Sequence upon Mutations Using CD

CD measurements can easily detect structural alterations in the protein upon mutations. Figure 3a shows the far-UV CD spectra of wild-type CopR protein and its mutants (Dim1-7). The wild-type protein shows an α-helical structure as depicted by a positive band at around 192 nm and two negative bands near 208 and 222 nm. The far-UV CD spectra of mutant proteins (Dim1, Dim3, Dim4, and Dim5) show minor deviations in the ellipticity pattern compared to the wild-type protein. This indicates that mutations did not drastically perturb the secondary structure of the proteins. In contrast, the CD spectrum of Dim6 shows a drastic reduction in α-helical content, pointing toward the significant structural perturbations upon single point mutation. The CD spectrum of Dim7 also shows conformational changes, but they are less pronounced than those in Dim6 [50].

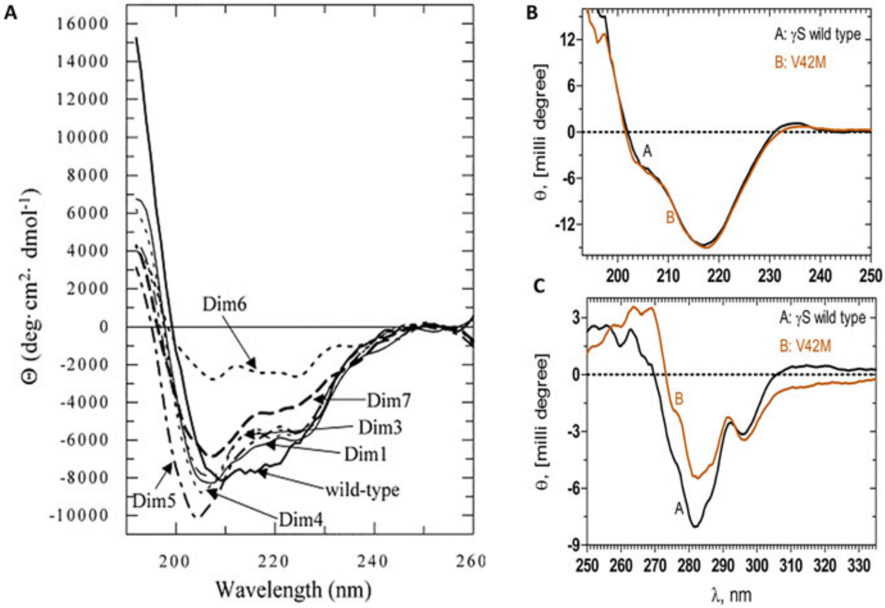

Fig. 3 A transcriptional activator protein. (**a**) Far-UV CD spectra of wild-type and mutant forms (Dim 1–7) of CopR. Adapted with permission from Steinmetzer et al. [50]. (**b**) Far-UV CD and (**c**) near-UV CD spectra of wild-type and mutant γS-crystallin (V42M). Adapted with permission from Vendra et al. [51]

Another example of monitoring mutation-induced structural changes in the protein is γS-crystallin. Figure 3b, c compares the far-UV and near-UV CD spectra of the wild-type and mutant γS-crystallin (V42M). The far-UV CD spectrum of the wild-type γS-crystallin displays two spectral bands, a negative band near 218 nm and a positive band at 195 nm, signifying the β-sheet secondary structural fold of the polypeptide chain. It should be noted that far-UV CD spectral profiles of wild-type and mutant protein were almost identical, indicating that the V42M mutation does not significantly affect the secondary structure of the protein. Strikingly, the near-UV CD spectra reveal that the tertiary structure around the aromatic residues is moderately altered in the mutant protein [51].

5 Analysis of Protein–Ligand Interactions

The binding of proteins with specific ligands such as cofactors, substrates, or regulatory molecules leads to structural changes vital for their physiological function. Such conformational changes may be monitored by the far-UV CD, near-UV CD, or both [52]. However, if the binding occurs near the aromatic amino acid residues, then small structural changes are easier to detect in the near-UV region since the CD contributions of the aromatic residues are highly sensitive to their environment. In contrast, the major structural changes in the protein's backbone are usually reflected in the far-UV CD spectrum [53, 54].

The study of ligand binding to a macromolecule is amenable if the signal from the complex is different from the sum of the signals from the components. Protein–protein interaction is a widespread and important biological regulatory phenomenon within cells. CD is a valuable technique for studying protein–protein interactions as changes occur in the secondary or tertiary structure of one or both components. The binding of small-molecule ligands such as metal ions or drugs is often accompanied by changes in the CD signal due to changes in the secondary or tertiary of the protein or the ligand. Small-molecule ligands usually have no or weak CD signal when free in solution. Still, they can show notable ellipticity when bound in the asymmetric environment of the binding pocket on the protein. Apart from protein–protein and protein–small-molecule interactions, the CD is specifically applicable to investigate protein–nucleic acid interactions as nucleic acids have strong signals in the near-UV region (250–300 nm), where proteins usually absorb weakly [55].

Human polynucleotide kinase (hPNK) acts by transferring the γ-phosphate of ATP to the 5′ ends of nucleic acids. Hence, the binding of ATP to hPNK is crucial for the proper functioning of the hPNK. Figure 4 shows the conformational changes occurring in hPNK upon binding to ATP. Two negative CD bands at 218 and 209 nm are observed in the CD spectrum of hPNK, a characteristic of the mixed α/β structure, with the band near 218 nm being attributed to the presence of β-structure in the protein. The binding of ATP to the activity of hPNK induces a substantial conformational change, as suggested by a decreased ellipticity value. The analysis of CD data indicated an increase in β-sheet structure and decreased α-helical

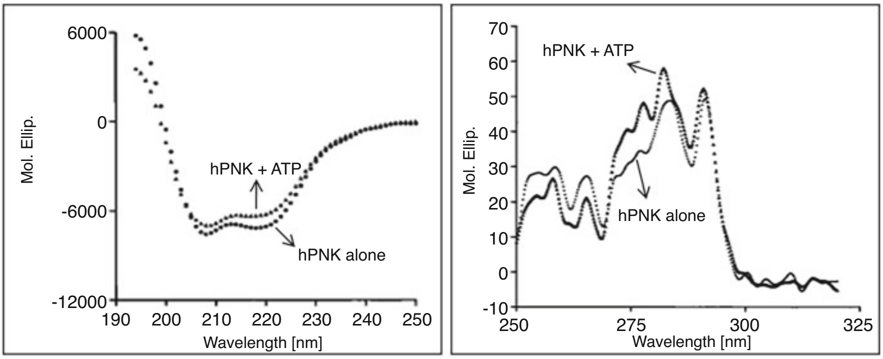

Fig. 4 Far-UV (left panel) and near-UV (right panel) CD spectra of hPNK alone (●) and in the presence of ATP (▲). Adapted with permission from Mani et al. [56]

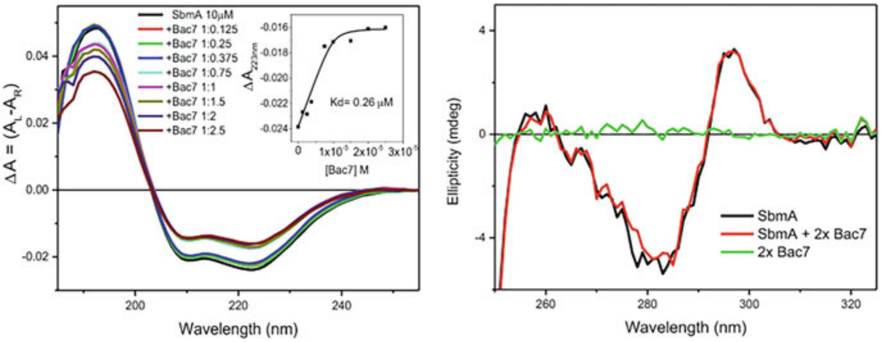

Fig. 5 Far-UV (left panel) and near-UV (right panel) CD spectra of SbmA with and without Bac7. Adapted with permission from Hussain et al. [54]. ΔA (absorption unit) $= \theta$ (ellipticity in millidegree)/3298.2

content in hPNK upon ATP binding [56]. The near-UV CD spectra also showed the perturbed environment of aromatic amino acid residues upon adding ATP. A substantial increase in the CD signal at the Trp peak near 291 nm is observed in the presence of ATP. Strikingly, two well-defined peaks attributed to tyrosine residues are observed at 284 and 278 nm instead of a broad shoulder around 279 nm in hPNK alone. The CD bands corresponding to Phe residues also show reduced ellipticity values in the apoprotein [56].

Another example where CD spectroscopy was used to monitor structural changes upon ligand binding is a bacterial inner membrane protein, SbmA, of a Gram-negative bacterium. SbmA is required to directly uptake the eukaryotic glycopeptides and antimicrobial peptides. The far-UV CD spectroscopy study showed that SbmA interacts with a proline-rich peptide, Bac7, and induces conformational changes, as revealed by the decrease in the CD signal in the wavelength range of 190–250 nm (Fig. 5). The dissociation constant (K_d) calculated after fitting the CD data was 0.26 μM showing the high binding affinity of Bac7 to SbmA. In contrast to far-UV CD spectra, no significant changes are observed in the near-UV CD spectral range, indicating that no aromatic residues are present in close vicinity at the binding interface of protein and ligand [54].

6 Determination of Protein Folding Pathways

The proper folding of a polypeptide chain into its biologically functional native structure is one of the fundamental processes of biology. Misfolding proteins in the cellular milieu often leads to fatal human and animal diseases. In the industrial context, overexpression of recombinant proteins leads to misfolding and aggregation, causing significant loss of the final product. CD is one of the many biophysical techniques routinely used to understand various aspects of the protein folding process, including the kinetic and thermodynamic properties of folding intermediates. Investigating protein folding mechanisms in vitro also provides valuable information about cellular processes such as protein trafficking and degradation.

A CD spectrophotometer coupled with the stopped-flow system is regarded as one of the best tools to study the mechanism of unfolding and refolding of proteins. This system provides the structural data of protein during the refolding process at a sub-millisecond time scale which can be used to explore the mechanism of the protein folding pathway [57, 58]. Figure 6 shows the refolding measurements of the denatured cytochrome *c* in the far-UV and near-UV regions using a stopped-flow system attached to the CD spectrophotometer. The changes in the CD signal at 222 nm are faster and occur within the time scale of 200 ms, suggesting the fast refolding of the secondary structure of the protein. However, the ellipticity changes at 289 nm, reflecting the environment around aromatic residues is relatively slower. Overall, the refolding kinetics measurements indicate the brief existence of an intermediate state with a folded secondary structure and flexible aromatic residue

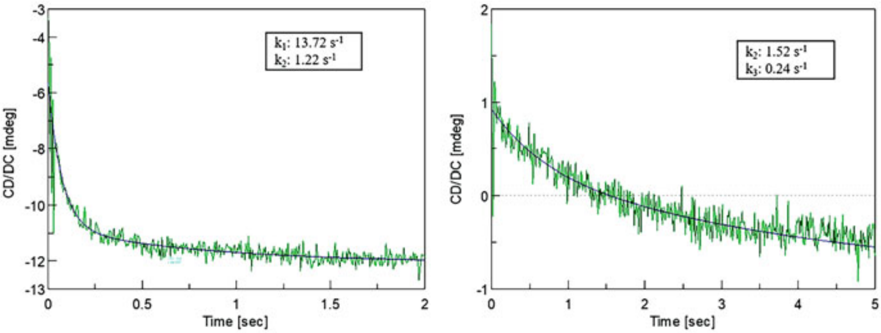

Fig. 6 Refolding kinetics of cytochrome *c* monitored at 222 nm (left panel) and 289 nm (right panel) using a stopped-flow CD system. Adapted with permission from [57]

side chains in the tertiary structure during the early stages of the refolding process [57].

7 Determination of Protein Stability

The stability of a protein is the fundamental property defined by the physicochemical conditions under which the protein is optimally functional. Hence, it is important to identify conditions that maximize the structural stability of the protein not only from the view of basic protein research but also to have a good yield of therapeutic proteins and other protein-based formulations in biotechnological industries. By improving the structural stability of the protein, off-pathway processes such as denaturation and aggregation could be prevented. However, it is important to note that the conditions (e.g., ionic strength, pH) that optimize the protein's physical stability might have deleterious effects on its chemical stability (e.g., deamination, oxidation). Therefore, the most stable protein formulation is achieved while considering all aspects of product quality, even with a bit of compromise with the physical and chemical stability of the protein [59]. Various methods that can be employed to measure the stability of the protein require the disruption of the native protein structure either by chemical or physical means. The conformational stability is essentially proportional to the resistance of the protein toward perturbation. The physical denaturation tools used to assess protein stability include temperature, high pressure, mechanical agitation, ultrasound, and ultraviolet radiations [60–64]. In comparison, the chemical denaturation of protein can be achieved by strong acids and bases, high concentrations of inorganic salts, salinity, organic solvents, and heavy metal salts [65–69].

7.1 Thermal Denaturation

Temperature is the most widely used tool for the physical denaturation of protein. Ideally, the thermal denaturation of a protein should be studied at its isoelectric point. However, the native protein is the least soluble at this pH, and the denatured protein is more prone to aggregation. Another disadvantage is that proteins usually get denatured far above the physiological or storage temperature. This requires long extrapolations of data to lower temperatures during thermodynamic analysis, which is often error-prone [70]. Furthermore, the thermal denaturation of proteins near their isoelectric points is usually an irreversible process that makes the calculation of stability parameters from the analysis of thermodynamic data highly unreliable and ambiguous. In such a case, physical stability rankings are presented only on T_m values, representing only a small part of the protein conformational stability curve as a function of temperature [71]. Irreversible protein denaturation is mostly followed by aggregation, affecting the accuracy of the measured T_m values [72–

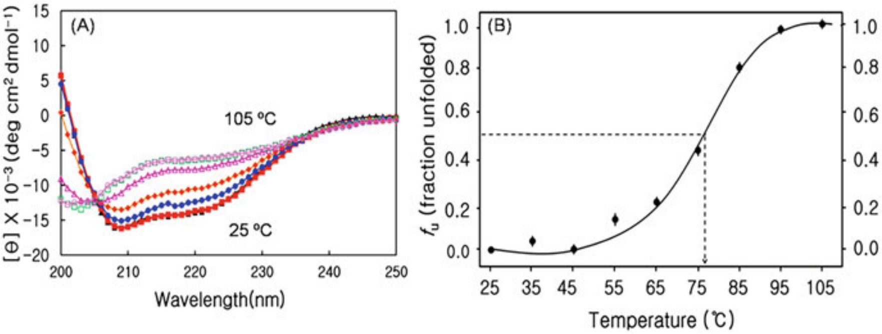

Fig. 7 Thermal unfolding of MTH1880. (**a**) Far-UV CD spectra measured at 10 °C intervals during the heating cycle (25–105 °C); 25 °C (black triangle), 45 °C (red square), 65 °C (blue circle), 75 °C (yellow square), 85 °C (pink triangle), 95 °C (green square), and 105 °C (pink circle). (**b**) A plot of fraction unfolded (f_u) derived from molar ellipticity measured at 222 nm as a function of temperature. The midpoint temperature of the unfolding transition (T_m) is 76 °C. Adapted with permission from Kim et al. [79]

74]. Additionally, the T_m value depends on the rate with which the temperature increases during the measurement, further complicating the stability extrapolations to lower temperatures.

The changes in the CD signal at a specific wavelength and as a function of temperature provide information about the thermodynamics of the protein unfolding process. The parameters retrieved from the CD thermodynamic data include the melting temperature or the midpoint temperature of the unfolding transition (T_m), the free energy of unfolding (ΔG), the van't Hoff enthalpy (ΔH) and entropy (ΔS) of unfolding, and the heat capacity changes (ΔC_P) of the unfolding transition [75]. Additionally, the analysis of CD spectra of protein acquired as a function of temperature provides information about the presence of intermediates in the folding pathway [7]. It should be noted that stability parameters obtained from the thermal denaturation curve of a protein must be validated by other experiments, such as the differential scanning calorimetry (DSC) measurements. In addition, an accurate determination of stability parameters from the analysis of optical denaturation curves depends on the temperature dependence of the pre- and post-transition baselines [76–78].

Figure 7 shows the thermal unfolding transition of MTH1880, a thermophilic protein from *Methanobacterium thermoautotrophicum*, probed by CD spectroscopy. Far-UV CD spectra were acquired with increasing temperature from 25 °C to 105 °C at an interval of 10 °C (Fig. 7a). MTH1880 shows the α-helical secondary structure in the temperature range of 25–45 °C, suggesting that the protein retains its native structural fold till 45 °C. The CD signal begins to decline continuously from 55 °C to 95 °C, where the protein is completely denatured and does not seem to show any further change in ellipticity with a further increase in the temperature. The thermal unfolding of MTH1880 follows a two-state transition unfolding pathway [79]. The raw CD data were converted into f_u, i.e., the fraction unfolded [77], which is used to

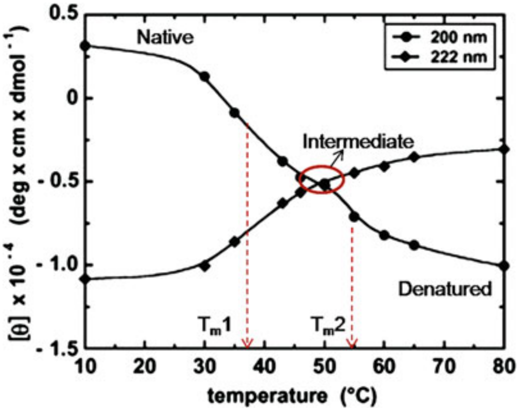

Fig. 8 Thermal unfolding of EcHUa2 followed by CD spectroscopy at 200 and 222 nm. An intermediate state is populated at around 48 °C in the melting curve. Adapted with permission from Ramstein et al. [80]

generate a f_u versus T plot (Fig. 7a). The normalized denaturation curve is fitted to the sigmoidal curve to derive thermodynamic parameters. The midpoint of unfolding transition or melting temperature (T_m) of MTH1880 was 76 °C as defined by $f_u = 0.5$.

Folding intermediates provide crucial information regarding the folding and assembly pathways. Thermal denaturation studies of protein often detect such folding/unfolding intermediates. Figure 8 shows the melting curve of *Escherichia coli* histone-like HU protein (EcHUa2) obtained by plotting the CD signals at 200 and 222 nm as a function of temperature. The thermal transition curve at 200 nm is biphasic, indicating a three-state denaturation mechanism ($N \rightarrow I \rightarrow U$) for EcHUa2 unfolding. An intermediate state is populated between N and U states at around 48 °C. The presence of two melting temperatures marks the denaturation process, i.e., 37.8 °C and 54.8 °C, corresponding to $N \rightarrow I$ and $I \rightarrow U$ transitions, respectively [80].

7.2 Chemical Denaturation

As discussed above, thermal denaturation studies of protein are often complicated and suffer from unreliable thermodynamic parameters if the unfolding process is irreversible. Different approaches are used to measure the protein stability in such a case, which employ chemical denaturants to unfold protein near-physiological temperature. Commonly used denaturants in isothermal chemical denaturation studies are urea and guanidine hydrochloride (GdnHCl) [81–83]. These chemical denaturants can prevent aggregation by keeping the unfolded protein species in stable and solubilized form, thus reversing the unfolding reaction. However, there are a few exceptions to this; for instant, low concentrations of GdnHCl fail to keep the denatured protein in the soluble form [84, 85].

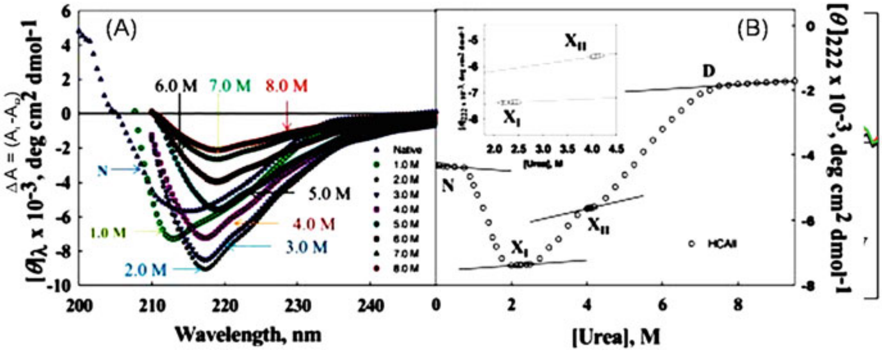

Fig. 9 Chemical denaturation of HCAII followed by CD spectroscopy. (**a**) Far-UV CD spectra of HCAII at various [Urea]. (**b**) Urea-induced denaturation is curved, followed by plotting the change in [θ] at 222 nm as a function of [Urea]. The inset shows the dependence of the optical properties of intermediates, y_{XI} and y_{XII}, on [Urea]. Adapted with permission from Wahiduzzaman et al. [95]

Measurement of typical urea (or GdnHCl)-induced denaturation curve monitored by CD involves (i) preparation of samples where the increasing concentration of denaturants are added to the protein solution followed by incubation at room temperature until the equilibrium is reached to ensure complete unfolding [59] and (ii) plotting of the CD signal at a given wavelength (θ, the raw ellipticity or [θ], the mean residue ellipticity) as a function of [denaturant], the molar concentration of the denaturant. The denaturation curve is analyzed for stability parameters, namely $\Delta G_{\mathrm{D}}{}^{0}$ (Gibbs free energy change (ΔG_{D}) associated with $N \leftrightarrow D$ process in the absence of the denaturant), m (dependence of ΔG_{D} on [denaturant]), and C_{m} (midpoint of denaturation curve). Analysis of the GdnHCl-induced and urea-induced denaturation curves is discussed elsewhere [86–93]. This analysis assumes that the protein denaturation is reversible. The preparation of protein samples for checking the reversibility of the denaturation by urea (or GdnHCl) is described elsewhere [86]. It should be noted that estimates of stability parameters depend on the mechanism of denaturation. However, it has to be validated whether the denaturation is a two-state process [94].

Figure 9 shows the stability studies of human carbonic anhydrase II (HCAII) employing the chemical denaturation method. Here, far-UV CD spectra of HCAII were collected at different urea concentrations (Fig. 9a). It was noted that the α-helical content increased with the addition of low concentrations of urea (0–2 M). Further increase in the urea concentration leads to the peak shift toward 218 nm, indicating the transformation of the α-helix into the β-sheet structure [95]. To obtain the denaturation curve, the molar ellipticity at 222 nm was plotted as a function of [Urea] (Fig. 9b). HCAII undergoes a cooperative triphasic unfolding profile with two distinct intermediate species (X_{I} and X_{II}) populated at around 2 and 4 M [Urea] on the denaturation pathway $N \leftrightarrow X_{\mathrm{I}} \leftrightarrow X_{\mathrm{II}} \leftrightarrow D$. From 0 to 2 M urea, a continuous gain in secondary structure was observed that reduced successively with

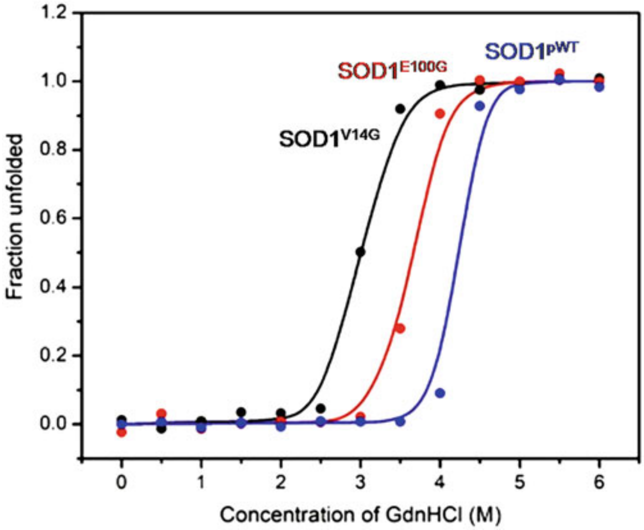

Fig. 10 GdnHCl-induced chemical denaturation curves of SOD1and its mutants monitored by CD spectroscopy. SOD1^{pWT} (blue), SOD1^{E100G} (red), and SOD1^{V14G} (black). Adapted with permission from Tompa et al. [97]

further increase in [Urea] until the protein is completely denatured. Values of the midpoint urea unfolding concentration (C_m) for transitions, $N \leftrightarrow X_I$, $X_I \leftrightarrow X_{II}$, and $X_{II} \leftrightarrow D$, were obtained after analyzing the denaturation curve, assuming that each transition curve follows a two-state mechanism ($C_{mI} = 1.33$ M, $C_{mII} = 3.25$ M, $C_{mIII} = 5.78$ M) [95].

Chemical denaturation studies provide valuable information about the destabilizing mutations that make the native protein either non-functional or prone to aggregation leading to devastating diseases [96–99]. Figure 10 shows the destabilizing effects of two single point mutations on the wild-type SOD1, whose misfolding and aggregation have been implicated in ALS. The GdnHCl-induced unfolding of SOD1^{pWT} and its mutants, SOD1^{E100G} and SOD1^{V14G}, was monitored using the far-UV CD. The denaturation curves of both wild-type and mutant proteins follow a two-state unfolding transition. The C_m values corresponding to the transition midpoints were 4.2, 3.7, and 3.0 M for SOD1^{V14G}, SOD1^{E100G}, and SOD1^{pWT}, respectively. This indicates that both mutations destabilize the wild-type SOD1protein. Although both mutations are positioned far away from the dimer interface and metal-binding site, they somehow perturb the metal loading to the active site. The partially metallated SOD1 was prone to misfolding and aggregation, causing neurodegenerative disorder [97].

8 Time-Resolved CD Measurements

CD spectroscopy has long been known as a reliable technique to determine the structural elements of proteins. It was used only to investigate the static structural properties in the past. However, it can now be employed to study protein dynamics with kinetic measurements using time-resolved CD spectroscopy. CD spectroscopy can be coupled with stopped-flow kinetic techniques to determine time dependence structural changes in the protein. The critical time-resolved CD measurements can detect events occurring on the millisecond resolution at a single wavelength [100]. The information from such studies provides mechanistic details of the protein folding phenomenon. Various excellent examples of time-resolved CD measurements are available in the literature [101–103]. Based on them, it seems that small proteins fold rapidly following a two-state transition without any detectable intermediate state(s). In contrast, larger proteins (more than 100 amino acids) usually fold via multi-state transition pathways. The native-like secondary structure is formed at the early folding stages, followed by acquiring tertiary structure interactions for larger proteins. The early intermediates formed often possess "molten globule" type characteristics.

In many experimental cases, a rapid burst phase in protein folding kinetics is reported using stopped-flow CD spectroscopy. The process occurs during the dead time of the instrument and produces an initial CD signal that differs from that expected for the unfolded protein, referred to as the burst phase. This difference in CD signal indicates that a substantial structural change occurred from D to N states during the initial burst phase [58]. An example of protein folding kinetics investigated using stopped-flow CD spectroscopy is shown in Fig. 11. The C-terminal domain (CTD) of spidroin 1 from the Ma gland (MaSp1) of the nursery web spider *Euprosthenops australis* was chemically denatured and refolded by rapid mixing into the buffer solution. A rapid burst phase of approximately 10 ms was observed within the dead time, which is beyond the detectable time resolution of the instrument. This was followed by a slow, single-exponential relaxation decay phase on the time scale of seconds [104]. To obtain the molecular details of the slow phase, the folding kinetics was measured at different protein concentrations (Fig. 11). After fitting exponential data of various protein concentrations, the time constants are almost identical within the error limits. This shows that the slow phase follows a mono-molecular folding event and is independent of protein concentration. It is also speculated that a dimeric intermediate is formed from the association of unfolded monomers during the rapid, unresolved burst phase. The event of biomolecular dimerization occurs too rapidly to be observed at the protein concentrations needed for sufficient signal in CD spectroscopy.

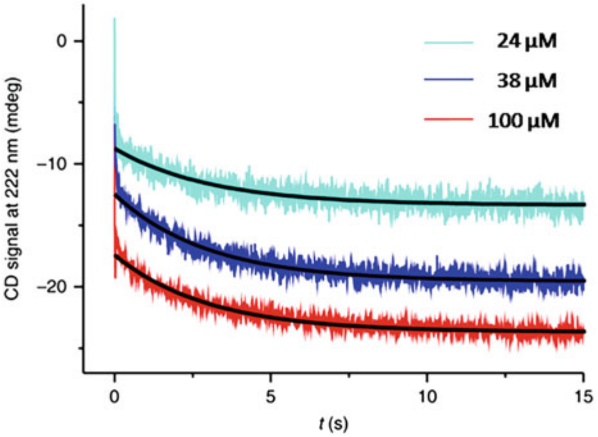

Fig. 11 Folding kinetics of CTD of spidroin 1 from the Ma gland of *Euprosthenops australis*. The far-UV CD stopped-flow spectroscopy measures the kinetic transients. Chemically denatured wild-type CTD samples at different protein concentrations (24, 38, and 100 µM) are refolded by rapid mixing into the buffer solution. The data fits the mono-exponential decay function and is depicted as a black line. Adapted with permission from Rat et al. [104]

9 Concluding Remarks

The protein folding phenomenon is of fundamental and practical importance, making the biophysical studies of protein folding and stability highly crucial. CD spectroscopy is an invaluable tool that monitors structural changes at the secondary and tertiary levels and in millisecond time resolution. It is a fast, reliable, and inexpensive technique for the initial investigation of recombinant proteins or those purified from tissues. Unlike X-ray crystallography and NMR spectroscopy, the two primary powerful structure determination techniques, the ease and the requirement of low protein concentrations, make CD spectroscopy a desirable and demanding method of choice. This chapter provides a comprehensive overview of the CD spectroscopy technique, its principle, and its applications in protein structural biology. Although CD could monitor fine structural details, more advanced and sophisticated instrumentation must be developed to detect events occurring too fast to be observed by currently available stopped-flow CD instruments. With the development of synchrotron radiation circular dichroism (SRCD) that uses high-intensity light sources, the measurement of data at lower wavelengths having more electronic transitions and thus giving more structural details has become feasible. The high signal-to-noise ratio conferred by SRCD enables the CD measurements in the presence of detergents, lipids, and other absorbing buffers.

Acknowledgments This work is supported by grants from the Indian Council of Medical Research, Government of India (Grant No. ISRM/12(22)/2020). P.G. thanks DST for the award of the National Post Doctoral Fellowship (File no. PDF/2017/001084). F.A. is thankful to the Indian National Science Academy for the award of the Senior Scientist Position.

References

1. I.B. Durowoju, K.S. Bhandal, J. Hu, B. Carpick, M. Kirkitadze, Differential scanning calorimetry—a method for assessing the thermal stability and conformation of protein antigen. J. Vis. Exp. **121**, 55262 (2017)
2. P. Gill, T.T. Moghadam, B. Ranjbar, Differential scanning calorimetry techniques: applications in biology and nanoscience. J. Biomol. Tech. **21**, 167 (2010)
3. W. Jiskoot, D. Crommelin, *Methods for Structural Analysis of Protein Pharmaceuticals* (Springer Science & Business Media, New York, 2005)
4. M.R. Eftink, The use of fluorescence methods to monitor unfolding transitions in proteins. Biophys. J. **66**, 482–501 (1994)
5. A.E. Johnson, Fluorescence approaches for determining protein conformations, interactions and mechanisms at membranes. Traffic **6**, 1078–1092 (2005)
6. N.J. Greenfield, Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. **1**, 2876–2890 (2006)
7. S.M. Kelly, T.J. Jess, N.C. Price, How to study proteins by circular dichroism. Biochim. Biophys. Acta **1751**, 119–139 (2005)
8. S.M. Kelly, N.C. Price, The use of circular dichroism in the investigation of protein structure and function. Curr. Protein Pept. Sci. **1**, 349–384 (2000)
9. A.A. Yee, A. Savchenko, A. Ignachenko, J. Lukin, X. Xu, T. Skarina, E. Evdokimova, C.S. Liu, A. Semesi, V. Guido, A.M. Edwards, C.H. Arrowsmith, NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. J. Am. Chem. Soc. **127**, 16512–16517 (2005)
10. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Singapore, Springer Nature, 2020)
11. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods*, 1st edn. (Academic Press, London, 2022)
12. F. van Eerden, *Differential Scanning Calorimetry and Protein Stability*, Faculty of Science and Engineering (2009)
13. H. Chosrowjan, S. Taniguchi, F. Tanaka, Ultrafast fluorescence upconversion technique and its applications to proteins. FEBS J. **282**, 3003–3015 (2015)
14. B. Bhattacharyya, S. Kapoor, D. Panda, Fluorescence spectroscopic methods to analyze drug–tubulin interactions, in *Methods in Cell Biology*, ed. by L. Wilson, J.J. Correia, (Academic Press, Cambridge, MA, 2010), pp. 301–329
15. A.A.T. Naqvi, T. Mohammad, G.M. Hasan, M.I. Hassan, Advancements in docking and molecular dynamics simulations towards ligand-receptor interactions and structure-function relationships. Curr. Top. Med. Chem. **18**, 1755–1768 (2018)
16. D.M. Rogers, S.B. Jasim, N.T. Dyer, F. Auvray, M. Réfrégiers, J.D. Hirst, Electronic circular dichroism spectroscopy of proteins. Chem **5**, 2751–2774 (2019)
17. G. Büyükköroğlu, D.D. Dora, F. Özdemir, C. Hızel, Techniques for protein analysis, in *Omics Technologies and Bio-Engineering*, ed. by D. Barh, V. Azevedo, (Academic Press, London, 2018), pp. 317–351
18. J. Skolnick, H. Zhou, M. Gao, On the possible origin of protein homochirality, structure, and biochemical function. Proc. Natl. Acad. Sci. **116**, 26571–26579 (2019)
19. C.-L. Towse, G. Hopping, I. Vulovic, V. Daggett, Nature versus design: the conformational propensities of D-amino acids and the importance of side chain chirality. Protein Eng. Des. Sel. **27**, 447–455 (2014)
20. Y. Jeong, H.W. Kim, J. Ku, J. Seo, Breakdown of chiral recognition of amino acids in reduced dimensions. Sci. Rep. **10**, 16166 (2020)
21. M. Mulkerrin, *Protein Structure Analysis Using Circular Dichroism* (VCH Publishers, New York, 1996)

22. N.J. Greenfield, Biomacromolecular applications of circular dichroism and ORD, in *Encylopedia of Spectroscopy and Spectrometry*, ed. by J.C. Lindon, G.E. Trantor, J.L. Holmes, (Academic Press, Oxford, 1999), pp. 117–130
23. N. Sreerama, R.W. Woody, Computation and analysis of protein circular dichroism spectra. Methods Enzymol. **383**, 318–351 (2004)
24. D.H.A. Corrêa, C.H.I. Ramos, The use of circular dichroism spectroscopy to study protein folding, form and function. Afr. J. Biochem. Res. **3**, 164–173 (2009)
25. N. Greenfield, Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nat. Protoc. **1**, 2527–2535 (2006)
26. A. Rodger, Near UV protein CD, in *Encyclopedia of Biophysics*, ed. by G.C.K. Roberts, (Springer, Berlin, 2013), pp. 1694–1694
27. A. Naiyer, B. Khan, A. Hussain, A. Islam, M.F. Alajmi, M.I. Hassan, M. Sundd, F. Ahmad, Stability of uniformly labeled (^{13}C and ^{15}N) cytochrome C and its L94G mutant. Sci. Rep. **11**, 6804 (2021)
28. J.M. Antosiewicz, D. Shugar, UV-Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 2: selected applications. Biophys. Rev. **8**, 163–177 (2016)
29. A. Mcauley, J. Jacob, C.G. Kolvenbach, K. Westland, H.J. Lee, S.R. Brych, D. Rehder, G.R. Kleemann, D.N. Brems, M. Matsumura, Contributions of a disulfide bond to the structure, stability, and dimerization of human IgG1 antibody CH3 domain. Protein Sci. **17**, 95–106 (2008)
30. M. Nagai, Y. Nagai, K. Imai, S. Neya, Circular dichroism of hemoglobin and myoglobin. Chirality **26**, 438–442 (2014)
31. L. Whitmore, B.A. Wallace, Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. Biopolymers **89**, 392–400 (2008)
32. A. Micsonai, F. Wien, É. Bulyáki, J. Kun, É. Moussong, Y.H. Lee, Y. Goto, M. Réfrégiers, J. Kardos, BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. Nucleic Acids Res. **46**, W315–W322 (2018)
33. C. Louis-Jeune, M.A. Andrade-Navarro, C. Perez-Iratxeta, Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. Proteins **80**, 374–381 (2012)
34. A.J. Miles, R.W. Janes, B.A. Wallace, Tools and methods for circular dichroism spectroscopy of proteins: a tutorial review. Chem. Soc. Rev. **50**, 8400–8413 (2021)
35. B.A. Wallace, R.W. Janes, *Modern Techniques for Circular Dichroism and Synchrotron Radiation Circular Dichroism Spectroscopy* (IOS Press, Amsterdam, 2009)
36. L. Whitmore, B. Wallace, DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. Nucleic Acids Res. **32**, W668–W673 (2004)
37. J.C. Ezerski, P. Zhang, N.C. Jennings, M.N. Waxham, M.S. Cheung, Molecular dynamics ensemble refinement of intrinsically disordered peptides according to deconvoluted spectra from circular dichroism. Biophys. J. **118**, 1665–1678 (2020)
38. H.A. Lashuel, C.R. Overk, A. Oueslati, E. Masliah, The many faces of α-synuclein: from structure and toxicity to therapeutic target. Nat. Rev. Neurosci. **14**, 38–48 (2013)
39. G.-f. Chen, T.-h. Xu, Y. Yan, Y.-r. Zhou, Y. Jiang, K. Melcher, H.E. Xu, Amyloid beta: structure, biology and structure-based therapeutic development. Acta Pharmacol. Sin. **38**, 1205–1235 (2017)
40. C. Soto, N. Satani, The intricate mechanisms of neurodegeneration in prion diseases. Trends Mol. Med. **17**, 14–24 (2011)
41. S. Negrini, V.G. Gorgoulis, T.D. Halazonetis, Genomic instability—an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. **11**, 220–228 (2010)
42. M. Vihinen, Functional effects of protein variants. Biochimie **180**, 104–120 (2021)

43. M. Li, A. Goncearenco, A.R. Panchenko, Annotating mutational effects on proteins and protein interactions: designing novel and revisiting existing protocols. Methods Mol. Biol. **1550**, 235–260 (2017)
44. M. Li, M. Petukh, E. Alexov, A.R. Panchenko, Predicting the impact of missense mutations on protein–protein binding affinity. J. Chem. Theory Comput. **10**, 1770–1780 (2014)
45. S.L. Clark, A.M. Rodriguez, R.R. Snyder, G.D.V. Hankins, D. Boehning, Structure-function of the tumor suppressor BRCA1. Comput. Struct. Biotechnol. J. **1**, e201204005 (2012)
46. K.L. Bradley, C.A. Stokes, S.J. Marciniak, L.C. Parker, A.M. Condliffe, Role of unfolded proteins in lung disease. Thorax **76**, 92–99 (2021)
47. S.G. Kelsen, The unfolded protein response in chronic obstructive pulmonary disease. Ann. Am. Thorac. Soc. **13**(Suppl 2), S138–S145 (2016)
48. C. Mueller, U. Altenburger, S. Mohl, Challenges for the pharmaceutical technical development of protein coformulations. J. Pharm. Pharmacol. **70**, 666–674 (2018)
49. K.S. Satheeshkumar, R. Jayakumar, Conformational polymorphism of the amyloidogenic peptide homologous to residues 113–127 of the prion protein. Biophys. J. **85**, 473–483 (2003)
50. K. Steinmetzer, A. Hillisch, J. Behlke, S. Brantl, Transcriptional repressor CopR: amino acids involved in forming the dimeric interface. Proteins **39**, 408–416 (2000)
51. V.P.R. Vendra, S. Chandani, D. Balasubramanian, The mutation V42M distorts the compact packing of the human gamma-S-Crystallin molecule, resulting in congenital cataract. PLoS One **7**, e51401 (2012)
52. G. Siligardi, R. Hussain, S.G. Patching, M.K. Phillips-Jones, Ligand- and drug-binding studies of membrane proteins revealed through circular dichroism spectroscopy. Biochim. Biophys. Acta Biomembr. **1838**, 34–42 (2014)
53. F.M. Assadi-Porter, J. Radek, H. Rao, M. Tonelli, Multimodal ligand binding studies of human and mouse G-coupled taste receptors to correlate their species-specific sweetness tasting properties. Molecules **23**, 2531 (2018)
54. R. Hussain, G. Siligardi, Characterisation of conformational and ligand binding properties of membrane proteins using synchrotron radiation circular dichroism (SRCD). Adv. Exp. Med. Biol. **922**, 43–59 (2016)
55. A. Podjarny, A. Dejaegere, B. Kieffer, *Biophysical Approaches Determining Ligand Binding to Biomolecular Targets: Detection, Measurement and Modelling* (Royal Society of Chemistry, Cambridge, 2011)
56. R.S. Mani, F. Karimi-Busheri, C.E. Cass, M. Weinfeld, Physical properties of human polynucleotide kinase: hydrodynamic and spectroscopic studies. Biochemistry **40**, 12967–12973 (2001)
57. S. Anwar, A. Shamsi, T. Mohammad, A. Islam, M.I. Hassan, Targeting pyruvate dehydrogenase kinase signaling in the development of effective cancer therapy. Biochim. Biophys. Acta Rev. Cancer **1876**, 188568 (2021)
58. H. Roder, K. Maki, H. Cheng, Early events in protein folding explored by rapid mixing methods. Chem. Rev. **106**, 1836–1861 (2006)
59. H. Svilenov, U. Markoja, G. Winter, Isothermal chemical denaturation as a complementary tool to overcome limitations of thermal differential scanning fluorimetry in predicting physical stability of protein formulations. Eur. J. Pharm. Biopharm. **125**, 106–113 (2018)
60. W. Messens, J. Van Camp, A. Huyghebaert, The use of high pressure to modify the functionality of food proteins. Trends Food Sci. Technol. **8**, 107–112 (1997)
61. Y. Matsuura, M. Takehira, Y. Joti, K. Ogasahara, T. Tanaka, N. Ono, N. Kunishima, K. Yutani, Thermodynamics of protein denaturation at temperatures over 100 °C: CutA1 mutant proteins substituted with hydrophobic and charged residues. Sci. Rep. **5**, 15545 (2015)
62. N. Dhanapati, M. Ishioroshi, I. Yoshida, K. Samejima, Effects of mechanical agitation, heating and pH on the structure of bovine alpha lactalbumin. Anim. Sci. Technol. **68**, 545–554 (1997)
63. V. De Leo, L. Catucci, A.E. Di Mauro, A. Agostiano, L. Giotta, M. Trotta, F. Milano, Effect of ultrasound on the function and structure of a membrane protein: the case study of

photosynthetic reaction center from Rhodobacter sphaeroides. Ultrason. Sonochem. **35**, 103–111 (2017)

64. J.H. Clark, The denaturation of egg albumin by ultra-violet radiation. J. Gen. Physiol. **19**, 199–210 (1935)
65. R. Dahiya, T. Mohammad, M.F. Alajmi, M.T. Rehman, G.M. Hasan, A. Hussain, M.I. Hassan, Insights into the conserved regulatory mechanisms of human and yeast aging. Biomol. Ther. **10**, 1–27 (2020)
66. S. Singh, C. Tyagi, I.A. Rather, J.S.M. Sabir, M.I. Hassan, A. Singh, I.K. Singh, Molecular modeling of chemosensory protein 3 from Spodoptera litura and its binding property with plant defensive metabolites. Int. J. Mol. Sci. **21**, 1–15 (2020)
67. P. Gupta, F.I. Khan, S. Roy, S. Anwar, R. Dahiya, M.F. Alajmi, A. Hussain, M.T. Rehman, D. Lai, M.I. Hassan, Functional implications of pH-induced conformational changes in the Sphingosine kinase 1. Spectrochim. Acta A Mol. Biomol. Spectrosc. **225**, 117453 (2020)
68. F. Ahmad, Protein stability from denaturation transition curves. Indian J. Biochem. Biophys. **28**, 168–173 (1991)
69. F. Ahmad, *Measuring the Conformational Stability of Enzymes in the Thermostability of Enzymes* (Narosa Publishing House: India, New Delhi, 1993), pp. 95–112
70. D. Matulis, J.K. Kranz, F.R. Salemme, M.J. Todd, Thermodynamic stability of carbonic anhydrase: measurements of binding affinity and stoichiometry using thermofluor. Biochemistry **44**, 5258–5266 (2005)
71. E. Freire, A. Schön, B.M. Hutchins, R.K. Brown, Chemical denaturation as a tool in the formulation optimization of biologics. Drug Discov. Today **18**, 1007–1013 (2013)
72. A. Azuaga, C. Dobson, P. Mateo, F. Conejero-Lara, Unfolding and aggregation during the thermal denaturation of streptokinase. Eur. J. Biochem. **269**, 4121–4133 (2002)
73. A. Schön, B.R. Clarkson, M. Jaime, E. Freire, Temperature stability of proteins: analysis of irreversible denaturation using isothermal calorimetry. Proteins **85**, 2009–2016 (2017)
74. J.M. Sanchez-Ruiz, Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. Biophys. J. **61**, 921–935 (1992)
75. D.C. Rees, A.D. Robertson, Some thermodynamic implications for the thermostability of proteins. Protein Sci. **10**, 1187–1194 (2001)
76. A. Sinha, S. Yadav, R. Ahmad, F. Ahmad, A possible origin of differences between calorimetric and equilibrium estimates of stability parameters of proteins. Biochem. J. **345**, 711–717 (2000)
77. S. Yadav, F. Ahmad, A new method for the determination of stability parameters of proteins from their heat-induced denaturation curves. Anal. Biochem. **283**, 207–213 (2000)
78. F. Ahmad, Protein folding: estimates of stability parameters from heat-induced conformational transition curves of proteins. PINSA **68**, 385–390 (2002)
79. H. Kim, S. Kim, Y. Jung, J. Han, J.-H. Yun, I. Chang, W. Lee, Probing the folding-unfolding transition of a thermophilic protein, MTH1880. PLoS One **11**, e0145853 (2016)
80. J. Ramstein, N. Hervouet, F. Coste, C. Zelwer, J. Oberto, B. Castaing, Evidence of a thermal unfolding dimeric intermediate for the Escherichia coli histone-like HU proteins: thermodynamics and structure. J. Mol. Biol. **331**, 101–121 (2003)
81. E. Freire, A. Schon, B. Hutchins, R. Brown, Chemical denaturation as a tool in the formulation optimization of biologics. Drug Discov. Today **18**, 1007 (2013)
82. F.I. Khan, P. Gupta, S. Roy, N. Azum, K.A. Alamry, A.M. Asiri, D. Lai, M.I. Hassan, Mechanistic insights into the urea-induced denaturation of human sphingosine kinase 1. Int. J. Biol. Macromol. **161**, 1496–1505 (2020)
83. P. Gupta, F.I. Khan, D. Ambreen, D. Lai, M.F. Alajmi, A. Hussain, A. Islam, F. Ahmad, M.I. Hassan, Investigation of guanidinium chloride-induced unfolding pathway of sphingosine kinase 1. Int. J. Biol. Macromol. **147**, 177–186 (2020)
84. R. Singh, M.I. Hassan, A. Islam, F. Ahmad, Cooperative unfolding of residual structure in heat denatured proteins by urea and guanidinium chloride. PLoS One **10**, e0128740 (2015)

85. F. Ahmad, P. McPhie, Thermodynamics of the denaturation of pepsinogen by urea. Biochemistry **17**, 241–246 (1978)
86. F. Ahmad, C.C. Bigelow, Estimation of the free energy of stabilization of ribonuclease a, lysozyme, alpha-lactalbumin, and myoglobin. J. Biol. Chem. **257**, 12935–12938 (1982)
87. F. Ahmad, C.C. Bigelow, Estimation of the stability of globular proteins. Biopolymers **25**, 1623–1633 (1986)
88. F. Ahmad, S. Yadav, S. Taneja, Determining stability of proteins from guanidinium chloride transition curves. Biochem. J. **287**, 481–485 (1992)
89. R. Gupta, S. Yadav, F. Ahmad, Protein stability: urea-induced versus guanidine-induced unfolding of metmyoglobin. Biochemistry **35**, 11925–11930 (1996)
90. T. Tripathi, S. Rahlfs, K. Becker, V. Bhakuni, Structural and stability characteristics of a monothiol glutaredoxin: glutaredoxin-like protein 1 from plasmodium falciparum. Biochim. Biophys. Acta **1784**, 946–952 (2008)
91. T. Tripathi, B.K. Na, W.M. Sohn, K. Becker, V. Bhakuni, Structural, functional and unfolding characteristics of glutathione S-transferase of plasmodium vivax. Arch. Biochem. Biophys. **487**, 115–122 (2009)
92. T. Tripathi, A. Röseler, S. Rahlfs, K. Becker, V. Bhakuni, Conformational stability and energetics of plasmodium falciparum glutaredoxin. Biochimie **92**, 284–291 (2010)
93. T. Tripathi, Calculation of thermodynamic parameters of protein unfolding using far-ultraviolet circular dichroism. J. Protein. Proteomics **4**, 85–91 (2013)
94. H. Rahaman, M.K.A. Khan, M.I. Hassan, A. Islam, A.A. Moosavi-Movahedi, F. Ahmad, Evidence of non-coincidence of normalized sigmoidal curves of two different structural properties for two-state protein folding/unfolding. J. Chem. Thermodyn. **58**, 351–358 (2013)
95. D.M. Wahiduzzaman, M.A. Haque, D. Idrees, M.I. Hassan, A. Islam, F. Ahmad, Characterization of folding intermediates during urea-induced denaturation of human carbonic anhydrase II. Int. J. Biol. Macromol. **95**, 881–887 (2017)
96. P. Gupta, P. Mahlawat, S. Deep, Effect of disease-linked mutations on the structure, function, stability and aggregation of human carbonic anhydrase II. Int. J. Biol. Macromol. **143**, 472–482 (2020)
97. D.R. Tompa, S. Kadhirvel, Far positioned ALS associated mutants of cu/Zn SOD forms partially metallated, destabilized misfolding intermediates. Biochem. Biophys. Res. Commun. **516**, 494–499 (2019)
98. S. Boopathy, T.V. Silvas, M. Tischbein, S. Jansen, S.M. Shandilya, J.A. Zitzewitz, J.E. Landers, B.L. Goode, C.A. Schiffer, D.A. Bosco, Structural basis for mutation-induced destabilization of profilin 1 in ALS. Proc. Natl. Acad. Sci. **112**, 7984–7989 (2015)
99. M. Andreasen, S.B. Nielsen, K. Runager, G. Christiansen, N.C. Nielsen, J.J. Enghild, D.E. Otzen, Polymorphic fibrillation of the destabilized fourth fasciclin-1 domain mutant A546T of the transforming growth factor-β-induced protein (TGFBIp) occurs through multiple pathways with different oligomeric intermediates. J. Biol. Chem. **287**, 34730–34742 (2012)
100. Y.G. Thomas, I. Szundi, J.W. Lewis, D.S. Kliger, Microsecond time-resolved circular dichroism of rhodopsin photointermediates. Biochemistry **48**, 12283–12289 (2009)
101. C.M. Dobson, Protein folding and misfolding. Nature **426**, 884–890 (2003)
102. C.N. Pace, Conformational stability of globular proteins. Trends Biochem. Sci. **15**, 14–17 (1990)
103. R.L. Baldwin, G.D. Rose, How the hydrophobic factor drives protein folding. Proc. Natl. Acad. Sci. **113**, 12462–12466 (2016)
104. C. Rat, J.C. Heiby, J.P. Bunz, H. Neuweiler, Two-step self-assembly of a spider silk molecular clamp. Nat. Commun. **9**, 4779 (2018)

Fluorescence Spectroscopy-Based Methods to Study Protein Folding Dynamics

Ritesh Kumar, Timir Tripathi, and Prakash Saudagar

Abstract The biological function of a protein is characterised by its three-dimensional conformation and encoded by its amino acid sequence. Tremendous effort has been devoted to understanding the mechanism of protein folding and how the amino acid sequences encode the correct functional conformation of a protein. Fluorescence-based methods, such as time-resolved spectroscopy, fluorescence correlation spectroscopy, or labelling of the proteins by external fluorescent dyes, have been employed to understand the protein folding dynamics. Herein, we describe recent fluorescence spectroscopy-based techniques used to study the conformational dynamics of a protein. Furthermore, these techniques are commonly available in most research laboratories and are used to study the protein–protein, protein–DNA, and protein–ligand interactions. We focus on the extrinsic fluorescent dyes to characterise the folding intermediates and detect the amyloid fibril aggregation in Alzheimer's and Parkinson's diseases.

Keywords Fluorescence spectroscopy · Protein folding · Extrinsic fluorescent dyes · Protein aggregates · Intrinsic fluorophore · Fluorescence anisotropy

1 Introduction

Protein folding is a unique mechanism by which the linear sequence of amino acids transforms into a functional three-dimensional (3D) native state [1, 2]. It always excites biophysical researchers how such a linear chain of polypeptides transforms

R. Kumar
Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX, USA

T. Tripathi
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

P. Saudagar (✉)
Department of Biotechnology, National Institute of Technology Warangal, Warangal, India
e-mail: ps@nitw.ac.in

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_2

into a well-folded tertiary structure. It is essential to understand the kinetics of protein folding through which the linear structure acquires the correctly folded native state. Several biophysical methods, including circular dichroism (CD) spectroscopy, optical spectroscopy, fluorescence spectroscopy, single-molecule fluorescence methods, and infra-red assessments, are used to measure the kinetics of the protein folding dynamics [3–7]. These methods measure the real-time changes in protein folding and functional dynamics. Among the spectroscopic techniques, fluorescence-based methods are one of the most powerful tools for analysing protein folding dynamics [8–11].

In an experimental setup, the native state protein can be denatured using chemical denaturants such as urea or guanidium hydrochloride (GdnHCl), the addition of acids or bases, increasing temperature, and eliminating ligands. The ultimate goal of protein denaturation is to understand the frame-by-frame conformational changes occurring in a protein and to measure the kinetics of folding dynamics in real time. Although the kinetic study of protein folding dynamics has been challenging, significant understandings are made, including, but not limited to, the uncovering folding intermediates, the thermodynamics of folding pathways, the extent of folding transition path time, and the examination of secondary structural changes [12, 13]. In this chapter, we discuss the fundamental aspects of fluorescence spectroscopy to study the protein folding dynamics and recent advancements in the uses of fluorescence spectroscopy.

2 Fluorescence Spectroscopy to Study the Kinetics of Protein Folding Dynamics

2.1 Fluorescence Principle

Fluorescence is a simple yet one of the most sensitive tools to study protein folding, denaturation, and aggregation. It is a three-step process which occurs in the molecules called fluorophores. Fluorophores absorb light at a particular wavelength and emit it at different wavelengths upon excitation by light.

The phenomenon of fluorescence takes place when the fluorophore is excited by a photon of energy (hν_{ex}) from the ground state (S_0) to a higher excited state (S_1). The higher excited state lifetime is typically between 1 and 10 ns. During this time, the fluorophore undergoes conformational changes and collides with the solvent. This process causes a loss of energy. The electron from the higher energy state S_1 is partially dissipated, yielding a lower energy state S_2. From this state (S_2), the electron comes back to the ground state (S_0) by quenching due to molecular collision with solvent molecules, fluorescence resonance energy transfer (FRET), or by intersystem crossing. The molecules not excited by the absorption of photons also return to the ground state (S_0). A photon of energy (hν_{em}) is released upon returning the fluorophore to the ground state (S_0) from excited stated (S_2). The difference in energy between $h\nu_{ex}$-$h\nu_{em}$ is termed as stokes shift. In intersystem crossing, the electrons return to the ground state (S_0) by a non-radiative process or emission of light called phosphorescence [14–16].

The entire process of fluorescence is cyclical, in which a single fluorophore molecule generates a large number of detectable photons unless the fluorophore is destroyed in the excited state by the process called photobleaching. In the case of polyatomic molecules, the photons of energy $h\nu_{ex}$ and $h\nu_{em}$ are replaced as broad energy spectra termed as fluorescence excitation spectrum and fluorescence emission spectrum, respectively [17, 18].

2.2 *Fluorescence Instrumentation*

To investigate the kinetics of protein folding dynamics, it is necessary to differentiate the features of the spectrofluorometer used and the available experimental options. The spectrofluorometer consists of four essential components: an excitation monochromator, a sample compartment, an emission monochromator, and a detector. The xenon arc lamp is generally used as a light source. The fluorophore is kept in the sample compartment with the sample. The detector records the photons generated during emission and produces a recordable output electrical signal. The detector is placed at a right angle to the excitation beam to reduce the excitation light on the detector (Fig. 1) [19].

For anisotropy measurements, a set of polarisers are used between the sample compartment and excitation monochromator as well as the emission monochromator [20, 21]. Time-resolved fluorescence spectroscopy consists of a pulsed laser source, a monochromator that selects the excitation wavelength, and a single photon detector connected with a multi-channel photomultiplier for converting the time to amplitude (Fig. 2). This provides an accurate measurement of the time between two events (start and stop signals) and obtains a histogram of photons emitted as a function of time [22–24].

The intensity of fluorescence follows the parameters defined by the Beer-Lambert law of absorbance, such as the optical path length, molar extinction coefficient, and

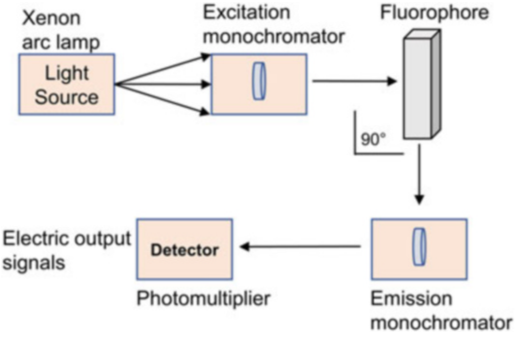

Fig. 1 Schematic representation of the fluorescence spectroscopy instrumentation

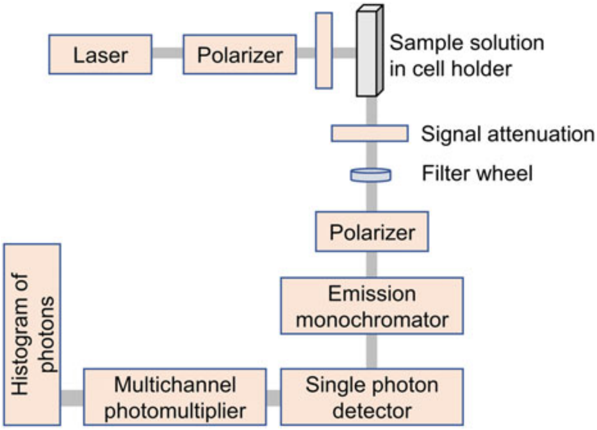

Fig. 2 Schematic representation of the time-resolved fluorescence spectroscopy instrumentation

solute concentration. In the case of turbid solutions, it weakens the excitation beam, and the sample that is facing toward the beam only undergoes fluorescence, termed the inner filter effect. Moreover, in the case of overlapping excitation and emission spectra, the light emitted inside the solution is reabsorbed by neighbouring molecules, producing a weak emission spectrum [16].

2.3 *Fluorescence Measurement Using Intrinsic Fluorophores*

Protein folding dynamics is a highly regulated phenomenon by which a protein adopts its native 3D confirmation. It occurs through the formation of secondary structures followed by tertiary and sometimes quaternary structures in the case of oligomeric proteins [25]. Secondary structures of a protein can be investigated using the methods of CD spectroscopy, whereas the tertiary structure can be monitored using the intrinsic fluorescence property of a protein. There are three aromatic amino acids (tyrosine, tryptophan, and phenylalanine) that contain intrinsic fluorescence properties. Tyrosine and tryptophan provide sufficient quantum yield upon absorption of photons, whereas phenylalanine provides the lowest quantum yield (quantum yield for tyrosine 0.13, tryptophan 0.14, and phenylalanine 0.02). The quantum yield is calculated as the ratio of the number of photons emitted to the number of photons absorbed. The excitation wavelength of tryptophan is 280 nm, tyrosine 285 nm, and phenylalanine 258 nm, while the emission wavelength is 350, 304, and 282 nm, respectively, in water (Table 1) [26–28].

To selectively excite tryptophan only, the excitation wavelength should be at 295 nm. More importantly, the fluorescence properties of these residues are extremely sensitive to the environment in which the folding or unfolding of a protein occurs. In the native confirmation, tryptophan and tyrosine residues are mainly

Table 1 Intrinsic fluorescence spectra of aromatic amino acids in water

Aromatic amino acids	λexcitation (nm)	λemission (nm)	Quantum yield
Tryptophan	280	350	0.14
Tyrosine	285	304	0.13
Phenylalanine	258	282	0.02

buried in the hydrophobic core of a protein, whereas in a partially folded or denatured state, these residues are exposed to the solvents. In a hydrophobic environment, tryptophan and tyrosine give higher quantum yield and hence high fluorescence intensity. In a hydrophilic environment, the tyrosine and tryptophan are exposed to the solvent and have a lower quantum yield, leading to low fluorescence intensity [28].

The tryptophan residue is also highly sensitive to solvent composition. When it is fully buried in the protein core and less exposed to the solvent, the emission spectrum is 325 nm, whereas, upon denaturation, the emission spectrum shifts to 350 nm [29, 30]. The shift in the emission spectra depends on the surrounding water molecules as well the orientation of the indole ring of tryptophan. Thus, tryptophan-containing proteins exhibit an emission spectrum between 330 and 350 nm, depending upon the environment and polarity of the solvent [31]. In tyrosine-containing proteins, the hydroxyl group of tyrosine is ionised at higher pH or in a buffer containing 2.0 M acetate and forms a tyrosinate complex. Instead of the maximum emission spectrum of 304 nm by tyrosine, tyrosinate shows a fluorescence spectrum at 340 nm and overlaps with the emission spectra of tryptophan [32]. The fluorescence signal generated by tryptophan can be easily quenched by electron-rich molecules such as amines, histidine groups, and carboxylic acids [33].

2.4 *Fluorescence Measurement Using Extrinsic Fluorophores*

The intrinsic fluorescence is limited to naturally occurring fluorescent amino acids (tryptophan, tyrosine, and phenylalanine); however, extrinsic fluorescence probes can be chemically attached to a protein, which can provide additional avenues for characterising protein folding dynamics [34]. These extrinsic dyes can bind to a protein through non-covalent interactions, such as electrostatic or hydrophobic interactions. Furthermore, they can bind covalently to proteins via the alpha-amino group at the N-terminus, epsilon-amino group of lysine residues, and thiol group of cysteine residues [34].

The extrinsic fluorophores are generally non-fluorescent in an aqueous environment but become highly fluorescent in non-polar solutions or upon binding to the hydrophobic pockets of a protein. These dyes are widely used in various applications of protein characterisation, such as to study the kinetics of protein folding/unfolding, to investigate the active site of an enzyme, to monitor conformational changes, and to measure the hydrophobic surfaces in a protein. The commonly used extrinsic fluorescent dyes are 1-anilinonaphthalene-8-sulfonate (ANS), the dimeric analogue of ANS 4,4′-bis-1-anilinonaphthalene-8-sulfonate (Bis-ANS), 9-(dicyanovinyl)-julolidine (DCVJ), Thioflavin T (ThT), Nile Red, and Congo Red [35–39]. Stryer et al. described the binding of ANS on the hydrophobic pocket of apohemoglobin and apomyoglobin, leading to an increase in the quantum yield, hence more fluorescence intensity [40]. Rosen and Weber first characterised the dye Bis-ANS, which is the most commonly used dye in protein folding dynamics [35]. Several fluorescent methods widely use extrinsic fluorophores for protein characterisation, such as steady-state fluorescence, anisotropy, time-resolved fluorescence, and fluorescence correlation spectroscopy (FCS).

2.4.1 Steady-State Fluorescence

Steady-state fluorescence is widely used to characterise amyloid fibrils, amyloid precursor proteins, amyloid-like, and colloidal-like protein aggregates. Several dyes such as ThT, Congo Red, curcumin derivatives, thienoquinoxaline-based styryl-quinoxaline, and boron-dipyrromethene-based dyes are used to probe amyloid-β fibrils and aggregates in case of Alzheimer's disease (Table 2) [41–44].

The absorption fluorescence properties of ThT are generally affected by solvent polarity, viscosity, and the rigidity of the microenvironment. ThT, upon interaction with amyloid-β fibrils, shows increased quantum yield and emission maximum at

Table 2 Fluorescence spectra of extrinsic fluorescence probes

Extrinsic fluorescence probes	λexcitation (nm)	λemission (nm)
Thioflavin T (ThioT)	450	482
Congo red	525	625
2-anilinonaphthalene-6-sulfonate (2,6-ANS)	350	417
4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonate (Bis-ANS)	360	493
Nile-red	540–600	640
9-(2,2-Dicyanovinyl) julolidine (DCVJ)	433	498
9-(2-Carboxy-2-cyanovinyl) julolidine (CCVJ)	279	441
2-(p-toluidinyl) naphthalene-6-sulfonate (TNS)	395	440
Thienoquinoxaline	468	530–580
Styryl-quinoxaline	468	530–580
Curcumin	420	500
Curcumin derivatives CRANAD-1	540	640
Curcumin derivatives CRANAD-2	640	715–800
Boron-dipyrromethene (BODIPY)	500	650
Triazole-BODIPY	525	540
5,5′,6,6′-tetracholoro-1,1,3,3′-tetraethylbenzimidazolyl carbocyanine iodide (JC-1)	490	500–600

480 nm. Several studies have reported the interaction of ThT with β-sheet structures and increased fluorescence intensity. However, it should be noted that ThT also induces fluorescence upon binding to some non-β-sheet rich structures, e.g. in transthyretin and acetylcholinesterase [45–47].

In combination with ThT, Congo Red is also used to analyse amyloid-β fibrils and aggregates. This dye interacts with the amyloid via non-ionic bonds in alkaline ethanolic solutions. The UV absorption maxima of Congo Red dye changes from 450 to 540 nm upon binding to amyloids [48, 49]. In addition, amyloid fibrils made in vitro from the bovine insulin showed a CD spectrum between 300 and 600 nm upon binding to Congo Red dye [50]. In an aqueous solution, the dye oligomerises itself and binds amyloid as an oligomeric ligand [39, 51].

Conversely, the colloidal proteins in food industries, such as casein and gliadins, have been characterised using the fluorescent probe Nile Red and curcumin derivatives [52, 53]. The electron transfer from the diethyl amino group of Nile Red to the electron-withdrawing group induces fluorescence. The excitation state of Nile Red is dependent upon the polarity of the medium. It's quantum yield decreases in polar solvents, whereas more fluorescence is observed in non-polar solvents. The fluorescence intensity of Nile Red is reported to increase in the order of methanol, ethanol, and DMSO, in comparison to water [38]. The sensitivity of Nile Red towards the polarity of the environment provides an edge to investigate the conformation of proteins during aggregation and unfolding [54].

2.4.2 Steady-State Fluorescence Anisotropy

Steady-state fluorescence anisotropy measures the differences in the rotational displacement between bound versus free molecules during the lifetime of the excited state of the molecule [21, 55]. This method is widely used to investigate protein–protein, protein–ligand, and protein–DNA interactions. The anisotropy measurements are performed with the illuminating sample with vertically polarised light in such a way that the electric vector of the excitation light is oriented parallel to the z-axis. The fluorescence intensity of the samples is measured with the emission polariser oriented parallel or perpendicular to the excitation polariser. The fluorescence anisotropy using the vertically polarised excitation is obtained using the following equation:

$$A = \frac{(Ivv - Ivh)}{(Ivv + 2Ivh)}$$

Here, *Ivv* is the fluorescence intensity of the sample collected with excitation polariser oriented vertically or parallel to the emission polariser, and *Ivh* is the fluorescence intensity when vertically polarised excitation and horizontally or perpendicular polarised emission [4, 56].

2.4.3 Time-Resolved Fluorescence

Time-resolved fluorescence measurement is a technique to monitor molecular interactions as a function of time. The sample is excited with a pulse of light, and the decay of the emission wavelength is detected over time [57]. The primary application of this technique is in studying protein oligomerisation and amyloid fibril formation [58]. Dyes such as ANS, Bis-ANS, and DCVJ are extensively used in this technique for the detection of the oligomeric state of a protein [59, 60].

ANS and Bis-ANS dyes are sensitive to the surrounding environment, and their fluorescence properties are easily affected by changing viscosity, polarity, and temperature of the medium. In case of decreasing the dielectric constant of the solvent, or changing the solvent from aqueous to organic, the quantum yield or fluorescence intensity of ANS and Bis-ANS increases. ANS and Bis-ANS bind to a protein molecule via hydrophobic and electrostatic interactions. The negatively charged sulfonate group of ANS and Bis-ANS forms ionic interactions with the positively charged arginine, lysine, or histidine moieties of the protein. Bis-ANS primarily interacts with a protein via hydrophobic interactions, and because of its different size than ANS, it binds with more affinity leading to enhanced quantum yield compared to ANS binding [61–63].

In contrast to the other fluorescent dyes, DCVJ is more sensitive toward the viscosity of the environment rather than the polarity [64]. The electron-donating groups of nitrogen in the ring of DCVJ transfer their electron to the nitrile groups. In the case of glycerol as a solvent (highly viscous), the quantum yield of DCVJ and similar dyes such as 9-(2-carboxy-2-cyanovinyl)-julolidine (CCVJ) are found to be increased [65]. These characteristic features of DCVJ and CCVJ are employed to investigate the viscosity of blood plasma [66]. Upon binding the dye to the protein, the viscosity of the microenvironment inhibits the intramolecular rotation and hence increased fluorescence intensity.

2.4.4 Fluorescence Correlation Spectroscopy

Fluorescence correlation spectroscopy (FCS) measures the variability of the fluorescence molecule in a living cell or in a solution to understand molecular events such as conformational changes in the molecule [67]. It is a powerful technique for understanding the dynamics of extremely low-concentrated biomolecules. Unlike other fluorescence techniques where emission spectra are measured, FCS can measure the spontaneous change in fluorescence intensity caused by minute changes in the thermal equilibrium [68, 69].

The biomolecule must be labelled with a fluorescent dye for FCS analysis [70]. Photostability is the most crucial determinant for the dye to withstand the high-power laser. Initially, fluorescein, a very good fluorophore in fluorescence spectroscopy, was tested for the FCS analysis, but it has a high photobleaching effect after laser that leads to artefacts in the analysis of single-molecule applications.

Engineered dyes with a high photostability, such as Alexa488 and other Alexa dyes, demonstrate a wide range of excitation and emission wavelength. Several other dyes, such as cyanine dyes (Cy2, Cy3, Cy5) and rhodamines (rhodamine green, rhodamine B, rhodamine 6G), are found useful in FCS analysis and have a low photobleaching effect [71, 72]. The general disadvantage of using external dyes is the need to label the biomolecule before performing an experiment. Labelling the biomolecule is not suitable in intracellular measurements, where we need to label the biomolecules inside the cell [73]. In this case, an autofluorescent protein such as GFP can be cloned along with the protein of interest and the expression and other biophysical dynamics can be measured using the FCS and fluorescence microscopy [74].

3 Conclusions

In this chapter, we briefly discussed the basics of various fluorescence spectroscopic techniques, including labelling by extrinsic fluorophores and fluorescent dyes such as rhodamine and cyanine to understand the conformational dynamics of a protein. By combining fluorescence spectroscopic techniques with other spectroscopic or simulation-based techniques, the protein folding dynamics and assembly can be deciphered. The main advantage of fluorescence spectroscopic methods to understand conformational dynamics is their availability in most research laboratories and easy data processing.

References

1. M.R. Eftink, M.C. Shastry, Fluorescence methods for studying kinetics of protein-folding reactions. Methods Enzymol. **278**, 258–286 (1997)
2. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Springer Nature, Singapore, 2020)
3. M.A. Hough, Choosing the optimal spectroscopic toolkit to understand protein function. Biosci. Rep. **37**(3), BSR20160378 (2017)
4. M.F. Pignataro, M.G. Herrera, V.I. Dodero, Evaluation of peptide/protein self-assembly and aggregation by spectroscopic methods. Molecules **25**(20), 4854 (2020)
5. S.M. Kelly, T.J. Jess, N.C. Price, How to study proteins by circular dichroism. Biochim. Biophys. Acta **1751**(2), 119–139 (2005)
6. T. Tripathi, Calculation of thermodynamic parameters of protein unfolding using far-ultraviolet circular dichroism. J Protein. Proteomics **4**(2), 85–91 (2013)
7. N. Nag, S. Sasidharan, P. Saudagar, T. Tripathi, Fundamentals of spectroscopy for biomolecular structure and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 1–35
8. S. Basak, K. Chattopadhyay, Studies of protein folding and dynamics using single molecule fluorescence spectroscopy. Phys. Chem. Chem. Phys. **16**(23), 11139–11149 (2014)
9. H.S. Chung, K. McHale, J.M. Louis, W.A. Eaton, Single-molecule fluorescence experiments determine protein folding transition path times. Science **335**(6071), 981–984 (2012)

10. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods*, 1st edn. (Academic Press, Cambridge, MA, 2022)
11. P. Saudagar, T. Tripathi, *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, 1st edn. (Academic Press, San Diego, 2023)
12. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins. Nature **450**(7172), 964–972 (2007)
13. P.C. Whitford, J.N. Onuchic, What protein folding teaches us about biological function and molecular machines. Curr. Opin. Struct. Biol. **30**, 57–62 (2015)
14. F.W.J. Teale, Principles of fluorescence spectroscopy—Lakowicz, Jr. Nature **307**(5950), 486–486 (1984)
15. R.F. Steiner, Principles of fluorescence spectroscopy—Lakowicz, Jr. Anal. Biochem. **137**(2), 539–539 (1984)
16. C.A. Royer, Fluorescence spectroscopy. Methods Mol. Biol. **40**, 65–89 (1995)
17. P.G. Varley, Fluorescence spectroscopy. Methods Mol. Biol. **22**, 203–218 (1994)
18. G. Krishnamoorthy, Fluorescence spectroscopy in molecular description of biological processes. Indian J. Biochem. Biophys. **40**(3), 147–159 (2003)
19. N. Shanker, S.L. Bane, Basic aspects of absorption and fluorescence spectroscopy and resonance energy transfer methods. Methods Cell Biol. **84**, 213–242 (2008)
20. A. Gijsbers, T. Nishigaki, N. Sanchez-Puig, Fluorescence anisotropy as a tool to study protein-protein interactions. J. Vis. Exp. **116**, 54640 (2016)
21. N.G. James, D.M. Jameson, Steady-state fluorescence polarization/anisotropy for the study of protein interactions. Methods Mol. Biol. **1076**, 29–42 (2014)
22. D.P. Millar, Time-resolved fluorescence spectroscopy. Curr. Opin. Struct. Biol. **6**(5), 637–642 (1996)
23. A.R. Holzwarth, Time-resolved fluorescence spectroscopy. Methods Enzymol. **246**, 334–362 (1995)
24. S.R. Anderson, Time-resolved fluorescence spectroscopy. Applications to calmodulin. J. Biol. Chem. **266**(18), 11405–11408 (1991)
25. P.B. Chetri, H. Khan, T. Tripathi, Methods to determine the oligomeric structure of proteins, in *Advances in Protein Molecular and Structural Biology Methods*, ed. by T. Tripathi, V.K. Dubey, (Academic Press, San Diego, 2022), pp. 49–76
26. J. Lee, R.T. Ross, Absorption and fluorescence of tyrosine hydrogen-bonded to amide-like ligands. J. Phys. Chem. B **102**(23), 4612–4618 (1998)
27. J.T. Vivian, P.R. Callis, Mechanisms of tryptophan fluorescence shifts in proteins. Biophys. J. **80**(5), 2093–2109 (2001)
28. C.P. Pan, P.L. Muino, M.D. Barkley, P.R. Callis, Correlation of tryptophan fluorescence spectral shifts and lifetimes arising directly from heterogeneous environment. J. Phys. Chem. B **115**(12), 3245–3253 (2011)
29. N. Hellmann, D. Schneider, Hands on: using tryptophan fluorescence spectroscopy to study protein structure. Methods Mol. Biol. **1958**, 379–401 (2019)
30. A.J. Lopez, L. Martinez, Parametric models to compute tryptophan fluorescence wavelengths from classical protein simulations. J. Comput. Chem. **39**(19), 1249–1258 (2018)
31. E.A. Burstein, S.M. Abornev, Y.K. Reshetnyak, Decomposition of protein tryptophan fluorescence spectra into log-normal components. I. Decomposition algorithms. Biophys. J. **81**(3), 1699–1709 (2001)
32. K.B. Davis, Z. Zhang, E.A. Karpova, J. Zhang, Application of tyrosine-tryptophan fluorescence resonance energy transfer in monitoring protein size changes. Anal. Biochem. **557**, 142–150 (2018)
33. K. Vladislav Victorovich, K. Tatyana Aleksandrovna, P. Victor Vitoldovich, S. Aleksander Nicolaevich, K. Larisa Valentinovna, A. Anastasia Aleksandrovna, Spectra of tryptophan fluorescence are the result of co-existence of certain most abundant stabilized excited state and certain most abundant destabilized excited state. Spectrochim. Acta A Mol. Biomol. Spectrosc. **257**, 119784 (2021)

34. A. Hawe, M. Sutter, W. Jiskoot, Extrinsic fluorescent dyes as tools for protein characterization. Pharm. Res. **25**(7), 1487–1499 (2008)
35. C.G. Rosen, G. Weber, Dimer formation from 1-amino-8-naphthalenesulfonate catalyzed by bovine serum albumin. A new fluorescent molecule with exceptional binding properties. Biochemistry **8**(10), 3915–3920 (1969)
36. C.E. Kung, J.K. Reed, Fluorescent molecular rotors: a new class of probes for tubulin structure and assembly. Biochemistry **28**(16), 6678–6686 (1989)
37. H. Naiki, K. Higuchi, M. Hosokawa, T. Takeda, Fluorometric determination of amyloid fibrils in vitro using the fluorescent dye, thioflavin T1. Anal. Biochem. **177**(2), 244–249 (1989)
38. P. Greenspan, E.P. Mayer, S.D. Fowler, Nile red: a selective fluorescent stain for intracellular lipid droplets. J. Cell Biol. **100**(3), 965–973 (1985)
39. A. Espargaro, S. Llabres, S.J. Saupe, C. Curutchet, F.J. Luque, R. Sabate, On the binding of Congo red to amyloid fibrils. Angew. Chem. Int. Ed. Engl. **59**(21), 8104–8107 (2020)
40. L. Stryer, The interaction of a naphthalene dye with apomyoglobin and apohemoglobin. A fluorescent probe of non-polar binding sites. J. Mol. Biol. **13**(2), 482–495 (1965)
41. H. LeVine 3rd, Thioflavine T interaction with synthetic Alzheimer's disease beta-amyloid peptides: detection of amyloid aggregation in solution. Protein Sci. **2**(3), 404–410 (1993)
42. H. LeVine 3rd, Mechanism of A beta(1-40) fibril-induced fluorescence of (trans,trans)-1-bromo-2,5-bis(4-hydroxystyryl)benzene (K114). Biochemistry **44**(48), 15937–15943 (2005)
43. X.Y. Liu, X.J. Wang, L. Shi, Y.H. Liu, L. Wang, K. Li, Q. Bu, X.B. Cen, X.Q. Yu, Rational design of quinoxalinone-based red-emitting probes for high-affinity and long-term visualizing amyloid-beta in vivo. Anal. Chem. **94**(21), 7665–7673 (2022)
44. A. Ojida, T. Sakamoto, M.A. Inoue, S.H. Fujishima, G. Lippens, I. Hamachi, Fluorescent BODIPY-based Zn(II) complex as a molecular probe for selective detection of neurofibrillary tangles in the brains of Alzheimer's disease patients. J. Am. Chem. Soc. **131**(18), 6543–6548 (2009)
45. A.A. Maskevich, V.I. Stsiapura, V.A. Kuzmitsky, I.M. Kuznetsova, O.I. Povarova, V.N. Uversky, K.K. Turoverov, Spectral properties of thioflavin T in solvents with different dielectric properties and in a fibril-incorporated form. J. Proteome Res. **6**(4), 1392–1401 (2007)
46. A.I. Sulatskaya, A.V. Lavysh, A.A. Maskevich, I.M. Kuznetsova, K.K. Turoverov, Thioflavin T fluoresces as excimer in highly concentrated aqueous solutions and as monomer being incorporated in amyloid fibrils. Sci. Rep. **7**(1), 2146 (2017)
47. M. Groenning, L. Olsen, M. van de Weert, J.M. Flink, S. Frokjaer, F.S. Jorgensen, Study on the binding of thioflavin T to beta-sheet-rich and non-beta-sheet cavities. J. Struct. Biol. **158**(3), 358–369 (2007)
48. A.J. Howie, D.B. Brewer, Optical properties of amyloid stained by Congo red: history and mechanisms. Micron **40**(3), 285–301 (2009)
49. W.E. Klunk, R.F. Jacob, R.P. Mason, Quantifying amyloid by Congo red spectral shift assay. Methods Enzymol. **309**, 285–305 (1999)
50. R. Khurana, V.N. Uversky, L. Nielsen, A.L. Fink, Is Congo red an amyloid-specific dye? J. Biol. Chem. **276**(25), 22715–22721 (2001)
51. H. Inouye, D.A. Kirschner, Alzheimer's beta-amyloid: insights into fibril formation and structure from Congo red binding. Subcell. Biochem. **38**, 203–224 (2005)
52. D.L. Sackett, J. Wolff, Nile red as a polarity-sensitive fluorescent probe of hydrophobic protein surfaces. Anal. Biochem. **167**(2), 228–234 (1987)
53. A. Sahu, N. Kasoju, U. Bora, Fluorescence study of the curcumin-casein micelle complexation and its application as a drug nanocarrier to cancer cells. Biomacromolecules **9**(10), 2905–2912 (2008)
54. A. Cser, K. Nagy, L. Biczok, Fluorescence lifetime of Nile red as a probe for the hydrogen bonding strength with its microenvironment. Chem. Phys. Lett. **360**(5-6), 473–478 (2002)
55. A. Chaudhary, K. Schneitz, Using steady-state fluorescence anisotropy to study protein clustering. Methods Mol. Biol. **2457**, 253–260 (2022)

56. H. Matsuzawa, K. Watanabe, M. Iwahashi, Fluorescence anisotropy and rotational diffusion of two kinds of 4-n-alkyl-4′-cyanobiphenyls in glycerol. J. Oleo Sci. **56**(11), 579–586 (2007)
57. J.A. Steinkamp, Time-resolved fluorescence measurements. Curr Protoc Cytom **Chapter 1**, Unit 1.15 (2001)
58. A. Tiiman, V. Jelic, J. Jarvet, P. Jaremo, N. Bogdanovic, R. Rigler, L. Terenius, A. Graslund, V. Vukojevic, Amyloidogenic nanoplaques in blood serum of patients with Alzheimer's disease revealed by time-resolved thioflavin T fluorescence intensity fluctuation analysis. J. Alzheimers Dis. **68**(2), 571–582 (2019)
59. A. Hawe, T. Rispens, J.N. Herron, W. Jiskoot, Probing bis-ANS binding sites of different affinity on aggregated IgG by steady-state fluorescence, time-resolved fluorescence and isothermal titration calorimetry. J. Pharm. Sci. **100**(4), 1294–1305 (2011)
60. D.M. Togashi, A.G. Ryder, Time-resolved fluorescence studies on bovine serum albumin denaturation process. J. Fluoresc. **16**(2), 153–160 (2006)
61. A. Bothra, A. Bhattacharyya, C. Mukhopadhyay, K. Bhattacharyya, S. Roy, A fluorescence spectroscopic and molecular dynamics study of bis-ANS/protein interaction. J. Biomol. Struct. Dyn. **15**(5), 959–966 (1998)
62. B. Stopa, L. Konieczny, B. Piekarska, I. Roterman, J. Rybarska, M. Skowronek, Effect of self association of bis-ANS and bis-azo dyes on protein binding. Biochimie **79**(1), 23–26 (1997)
63. M.Z. Kamal, J. Ali, N.M. Rao, Binding of bis-ANS to Bacillus subtilis lipase: a combined computational and experimental investigation. Biochim. Biophys. Acta **1834**(8), 1501–1509 (2013)
64. M.A. Haidekker, T.P. Brady, D. Lichlyter, E.A. Theodorakis, Effects of solvent polarity and solvent viscosity on the fluorescent properties of molecular rotors and related probes. Bioorg. Chem. **33**(6), 415–425 (2005)
65. T. Iio, M. Itakura, S. Takahashi, S. Sawada, 9-(Dicyanovinyl)julolidine binding to bovine brain calmodulin. J. Biochem. **109**(4), 499–502 (1991)
66. W.J. Akers, J.M. Cupps, M.A. Haidekker, Interaction of fluorescent molecular rotors with blood plasma proteins. Biorheology **42**(5), 335–344 (2005)
67. J.J. Mittag, J.O. Radler, J.J. McManus, Peptide self-assembly measured using fluorescence correlation spectroscopy. Methods Mol. Biol. **1777**, 159–171 (2018)
68. E.L. Elson, Brief introduction to fluorescence correlation spectroscopy. Methods Enzymol. **518**, 11–41 (2013)
69. J.A. Fitzpatrick, B.F. Lillemeier, Fluorescence correlation spectroscopy: linking molecular dynamics to biological function in vitro and in situ. Curr. Opin. Struct. Biol. **21**(5), 650–660 (2011)
70. E.L. Elson, Introduction to fluorescence correlation spectroscopy-brief and simple. Methods **140–141**, 3–9 (2018)
71. A. Sarkar, V. Namboodiri, J. Enderlein, M. Kumbhakar, Picosecond to second fluorescence correlation spectroscopy for studying solute exchange and quenching dynamics in Micellar media. J. Phys. Chem. Lett. **12**(31), 7641–7649 (2021)
72. M. Shimizu, S. Sasaki, M. Tsuruoka, DNA length evaluation using cyanine dye and fluorescence correlation spectroscopy. Biomacromolecules **6**(5), 2703–2707 (2005)
73. P. Dittrich, F. Malvezzi-Campeggi, M. Jahnz, P. Schwille, Accessing molecular dynamics in cells by fluorescence correlation spectroscopy. Biol. Chem. **382**(3), 491–494 (2001)
74. H. Engelke, D. Heinrich, J.O. Radler, Probing GFP-actin diffusion in living cells using fluorescence correlation spectroscopy. Phys. Biol. **7**(4), 046014 (2010)

Applications of Differential Scanning Calorimetry in Studying Folding and Stability of Proteins

Banesh Sooram, Neharika Gupta, Vihadhar Reddy Chethireddy, Timir Tripathi, and Prakash Saudagar

Abstract Over the last few decades, the analytical method of differential scanning calorimetry (DSC) has been established to investigate the stability, folding, and binding of proteins. The technique typically measures the differential heat between the test and reference samples. The deconvoluted thermogram can provide crucial information on process development and whether there are any stable transition states. Information about all the significant thermodynamical parameters can be extracted from the denaturation curve of a protein. DSC offers several advantages over other spectral methods like fluorescence and circular dichroism (CD) spectroscopy. For instance, CD and fluorescence spectra can only provide information on the secondary structural content of a protein, whereas DSC can be used to study protein stability and different transition states in folding. This chapter summarises the utility of DSC in studying protein stability and folding.

Keywords Calorimetry · Thermogram · Co-operative equilibrium folding · Heat capacity

1 Introduction

Calorimetry is the fundamental technique for measuring the thermal properties of a molecule to establish a relationship between temperature and particular physical attributes of the molecule [1]. Differential scanning calorimetry (DSC) is a thermo-analytical technique for determining the effect of temperature and time on a

B. Sooram · N. Gupta · V. R. Chethireddy · P. Saudagar (✉)
Department of Biotechnology, National Institute of Technology-Warangal, Warangal, Telangana, India
e-mail: ps@nitw.ac.in

T. Tripathi
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

Regional Director's Office, Indira Gandhi National Open University (IGNOU), Regional Centre Kohima, Kohima, India

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_3

molecule's chemical and physical properties [2]. The method offers both qualitative and quantitative information regarding endothermic and exothermic processes and alterations in the heat capacity [3]. This method maintains the reference and sample material at the same temperature while measuring the energy required to maintain the zero temperature difference [4].

Depending on the phase transition process, either more or less heat energy must be provided to the sample to avoid any temperature difference between the sample and the reference chamber [5]. If the process is exothermic (neutralisation, combustion, oxidation, and reduction processes), the sample reaction releases heat, so less heat is required from the source to attain the desired temperature. In contrast, if the reaction is endothermic (melting, boiling, vaporisation, sublimation, and crystallisation processes), more heat must be supplied [6]. In order to assess the thermodynamic properties of biomolecules and nanoscale materials, calorimeters are extensively employed in the fields of chemistry, biology, biotechnology, physics, and pharmacology [7]. DSC measures how a sample's physical characteristics and temperature change over time. In other words, it is a tool for thermal analysis that calculates the temperature and heat flow related to material transitions as a function of temperature and time [8]. DSC measures a heat quantity radiated or absorbed by the sample during a temperature shift based on the intersample temperature difference between the sample and reference material [7].

The molar heat capacity of samples can be determined using the DSC as a temperature function [9]. DSC profiles offer information on the thermal stability of proteins and, to a certain extent, act as a structural signature that can be used to evaluate structural conformation [10]. The recorded thermogram provides information on the melting temperature (T_m) and enthalpy (ΔH) values to ascertain the structural stability and folding of a protein [11]. Comparisons are made between the sample and reference, and variations in the obtained values signify variations in structural conformation and thermal stability. DSC is regularly used to study the stability testing of pharmaceutical formulations, protein stability, protein folding, and ligand binding to a particular protein target [12].

2 Theory and Governing Equations

Based on the mode of operation, differential scanning calorimeters can be of two types: heat-flux DSCs and power-compensated DSCs [13]. In a heat-flux DSC, the sample material in a pan and an empty reference pan are set on a thermoelectric disc encircled by a furnace. Both pans receive heat from the furnace through the thermoelectric disc at a linear heating rate [14]. However, because of the sample's heat capacity (C_p), there would be a temperature difference between it and the reference pan [14]. This temperature difference is monitored by area thermocouples, and the resulting heat flow is calculated using the thermal version of Ohm's equation [7].

$$q = \frac{\Delta T}{R} \tag{1}$$

where q is the sample heat flow, ΔT is the temperature difference between the sample and reference, and R is resistance.

The molar heat capacity obtained through a differential scanning calorimeter is used to calculate the Gibbs free energy (ΔG), entropy (ΔS), and enthalpy (ΔH). The change in heat capacity of the sample can be obtained from the following expression:

$$\Delta C_{\mathrm{p}} = C_{\mathrm{psample}} - C_{\mathrm{preference}} \tag{2}$$

The Gibbs free energy and other thermodynamics values can be obtained from the following equations [7, 15]:

$$\Delta S(T) = \int_{T_0}^{T} \frac{\Delta C_{\mathrm{p}}}{T} dT \tag{3}$$

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) \tag{4}$$

In the above equations, T_0 and T represent the sample's initial transition and final transition temperatures, respectively.

The native (folded) and denatured (unfolded) conformations of a biomolecule are in an equilibrium state when it is in solution [16]. The more stable the molecule is, the higher the thermal transition midpoint (T_{m}) [17]. The DSC measures the enthalpy (ΔH) of unfolding caused by heat-induced denaturation. It is also used to calculate the change in the heat capacity (ΔC_{p}) during denaturation [17]. The mechanisms behind the folding and stability of native biomolecules can be analysed using DSC [18]. Moreover, DSC is used to understand the protein's physiological environment, hydrogen bonds, conformational entropy, and hydrophobic interactions [19]. The accurate and high-quality DSC data provides crucial information on protein stability for the formulation of prospective therapeutic candidates and process development [20]. Proteins and nucleic acids can form macromolecular assemblies (>5000 Da) and undergo thermally induced conformational changes [21]. The distribution of non-covalent bonds during these structural changes causes heat to be absorbed. Using differential scanning calorimeters, this heat intake is quantified [22].

3 Instrumentation

Thermal scanning causes transitions in macromolecules such as protein or nucleic acid polymers [23]. The transitions occur from the native state to partially unfolded states (intermediate) to the fully denatured state. The thermally induced transition can resemble a two-step process between the native and unfolded states, especially regarding small protein unfolding [24]. On the other hand, nucleic acids exhibit

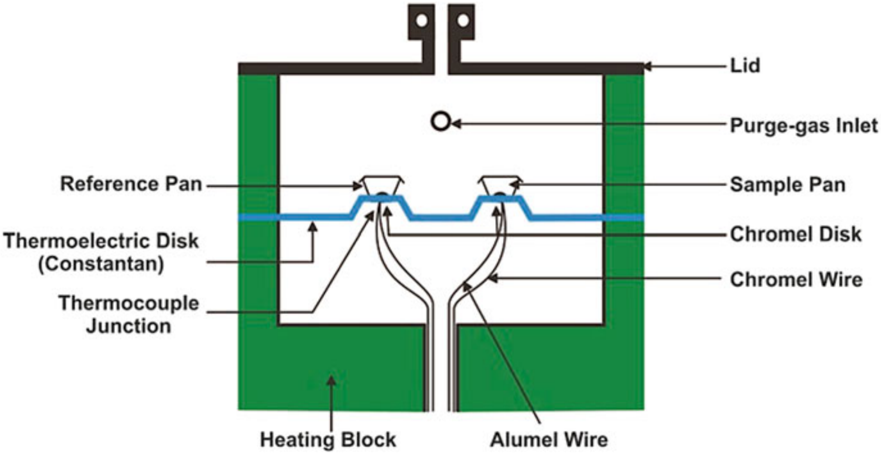

Fig. 1 Typical assembly of DSC instrument. The DSC contains two electrodes, one for each reference and sample. There is a lid and a purge gas inlet. At the base, there is a heating block. A thermos electric disk is placed below the sample and references. (Figure adapted with permission from [25])

strong cooperativity with many domains melting simultaneously or for smaller DNA pieces through several intermediate states [23]. The instrument must have several crucial components to examine these changes. The calorimetric device should be highly sensitive to the minor energy changes associated with the unfolding process, even at low concentrations. A typical instrument setup for DSC is shown in Fig. 1. To accomplish this high sensitivity, a differential power compensation system between a reference and sample cell, and approaches to regulate temperature and scan rate throughout the thermal experiment are used. Depending on the instrument, the cells in these devices have volumes of 0.15 to 0.8 mL of solution and are designed as either a capillary or an inverted lollipop [23].

The sample and reference holder are designed in such a way that they can withstand high pressure and temperatures [26, 27]. In the case of a high-temperature environment, the holders are made of platinum or ceramic, whereas for low temperatures, aluminium holders are used [28]. The shield that surrounds the cells either regulates scanning temperature or serves to maintain the shield and the cells at the same temperature [28]. In the nano DSC series, the temperature program that maintains the scan rate and scan temperature of the shield during the process is controlled by a computer [23]. As there is thermal contact between the cells and the shield, the cells' temperature rises in response to an increase in the shield's temperature [24]. Temperature sensors are placed between the sample and reference to determine the temperature difference. If there is a temperature difference, heaters on the cell surfaces provide the cells with compensating power. The calorimetric output is saved as the power compensation signal [27].

The power compensation signal is adjusted by computer control to compensate for cell mismatch because it is difficult to precisely match the thermal properties of the cells [29]. This adjustment will be visible in the baseline of the calorimetric scan [29]. The experiment is run strictly under adiabatic conditions without any heat transfer to the surroundings to maintain a minimal temperature difference between the cells and the shield. This instrument's scanning is not adiabatic because a temperature gradient is needed to cool the cells [14]. When a thermal event happens within the sample, thermal sensors on the cells detect differences between the sample and reference, and power compensation is used to keep the cells at almost the same temperature throughout the scan. Area thermocouples are used to measure the temperature difference between pans. For high-temperature or corrosive conditions, a platinum-rhodium thermocouple is used, whereas for low-temperature conditions, copper-constantan or chromel-alumel thermocouples are used [25, 29].

3.1 Types of DSC Instruments

DSCs are categorised into two subtypes: heat-compensated DSC (heat-exchanging calorimeter) and power-compensated DSC (heat-compensating calorimeter) [30]. They both differ in design and measurement principles. However, both possess the feature of measuring the signal proportional to the heat flow rate (J.s^{-1}), which facilitates the evaluation of a transition's time dependence using the (t) curve. In addition to these two typical configurations, flash DSC and temperature-modulated DSC are also available nowadays [30].

3.1.1 Heat-Compensated DSC

In a heat-compensated DSC, heat exchanged between the sample and its surroundings can be measured using a well-defined heat conduction path with a specific thermal resistance. Measurement systems to implement the heat exchange path include disk-type, turret-type, and cylinder-type. The disk-type measurement system, which uses a disc as a solid sample support and enables efficient heat exchange, is the most popular. The heat from the furnace is transmitted uniformly to the sample and the reference through the thermoelectric disk. The disc, which supports the sample and the reference crucible, has temperature sensors embedded into it. To minimise measurement errors, the arrangement of the sample and reference crucible and the temperature sensors coupled to them must be the same. According to Fourier's law, the temperature gradient drives the heat flow (i.e. thermodynamic affinity). The difference between the flow into the sample chamber and that into the reference corresponds to the heat flow into the sample itself. The heat flowing into the reference and sample crucible is equal if the sample crucible is empty. This system makes it possible to quickly and precisely measure over a broad temperature range. Depending on the equipment, the heat-flux DSC can often be used in the temperature range of 190 °C to 1600 °C. The DSC measurements can be performed in environments such as nitrogen or argon to prevent sample oxidation [30].

3.1.2 Power-Compensated DSC

In a power-compensated DSC, the sample and reference crucibles are installed in separate furnaces with an independent heating resistor and temperature sensors within each micro-furnace. The same electrical power is applied to both furnaces during the heat-up process. The reference and sample are heated and maintained at the same temperature using separate temperature controllers. A temperature differential between the sample and the reference occurs when any thermal reactions take place in the sample. The temperature difference is both the recorded signal and the input signal of the second controller circuit. By increasing or reducing an additional heating power of the sample furnace, the second circuit adjusts for the reaction heat flow rate of the sample. The compensating heating power, i.e. ΔP, is directly proportionate to the residual temperature difference ΔT. The electrical power necessary to achieve and maintain a state of zero temperature difference is recorded rather than the temperature difference between the two crucibles [30].

3.1.3 Nano-Calorimeter or Flash DSC

For some research, where physical and chemical processes happen significantly more rapidly than the normal scan rate of 10 K/min, a standard DSC's scanning rate is insufficient [31]. The low scan rate of a typical DSC makes it challenging to investigate various phenomena, including metastability, molecular rearrangement, and a variety of kinetic events. Researchers developed ultrafast DSC equipment to address these issues. Such a device was frequently referred to as nano-calorimetry or flash DSC. The scanning rate of these types of DSCs can go up to 750 K/min [32]. The advantage of flash DSC is that it resembles the temperature–time ramp that usually occurs for the cooling rates used in realistic processes. Additionally, the higher sensitivity makes it possible to measure the signals for delicate transitions and small material masses at low heat flow rates. A significantly widened scan range, i.e. from very low scan rates to ultrahigh cooling and heating rates, can be operated by the power compensation twin-type, chip-based fast scanning calorimeter (FSC). The scan rates for cooling and heating can approach 40,000 and 50,000 K/s, respectively. As a result, more than seven orders of magnitude of scan speeds can be covered when using flash DSC in conjunction with conventional DSC [33].

3.1.4 Temperature-Modulated DSC

DSC signals sometimes include a complexity of overlapping dynamic processes, leading to the complicated form of the temperature dependency of the heat capacity. However, using a traditional DSC with the typical linear heating rate, the kinetic and thermodynamic contributions to the heat capacity cannot be deconvoluted. Temperature-modulated DSC (TMDSC) was designed to overcome the drawbacks of conventional DSC methods. After the discovery of frequency-dependent measurements, Reading et al. developed the first commercial TMDSC, which integrates a single frequency oscillation with a typical linear DSC heating rate. Here, the normal temperature scan used in a typical DSC is generally overlaid by a low-frequency sinusoidal perturbation with a range of approximately 0.001 to 0.1 Hz [34].

$$T(t) = T_{\mathrm{o}} + qt + A_{\mathrm{t}} \sin(\omega t) \tag{5}$$

where A_{t} and ω are the amplitude and angular frequency of the sinusoidal oscillation, respectively.

3.2 Method and Sample Preparation for DSC

The crucial part of any DSC experiment and analysis is providing an optimum pressure to suppress the boiling temperature and prevent bubble formation in the sample and reference cells [35]. The pressure should not exceed the cell's tolerance as it may damage the cell. Nitrogen gas is purged through the cells to prevent the abovementioned problems [35]. A proper cleaning detergent is used to remove any remnants after each experiment. Moreover, it is essential to preheat or maintain the sample holder to maintain the experiment's integrity. Sample preparation is the second most important factor for the DSC experiment [36]. The protein samples need to be adequately dialyzed and filtered, ensuring the proper purity and homogenisation of the sample. Further, the protein concentration should be estimated using an appropriate method, as the typical proteins concentrations for a DSC experiment range from 0.5 to 1.0 mg/mL [37]. Finally, the sample is degassed to remove microbubbles that might interfere with the experiment's sensitivity. During the experiment, a blank or control sample is also used, which typically contains the buffer only [11].

Scanning both reference and test samples is done at constant pressure inside the cells because DSC calculates a protein's excess heat capacity at constant pressure in relation to a reference sample as a function of temperature [38]. Usually, a protein's or solution's heat capacity (C_{p}) changes gradually with temperature, but when a thermal event like denaturation occurs, it changes abruptly. The protein unfolding or melting is an endothermic event; thus, it needs energy to raise the temperature of the sample. The excess heat provided to the sample reflects in the specific heat capacity of the sample. The DSC curve is a plot between excess heat capacity (C_{p}) versus temperature. In the DSC curve, a peak appears, and the temperature corresponds to the transition midpoint (T_{m}), where maximum heat capacity can be seen. The area under the peak for the T_{m} curve represents the ΔH, as shown in Fig. 2. The value ΔH represents the secondary structural content of the protein sample [38, 39].

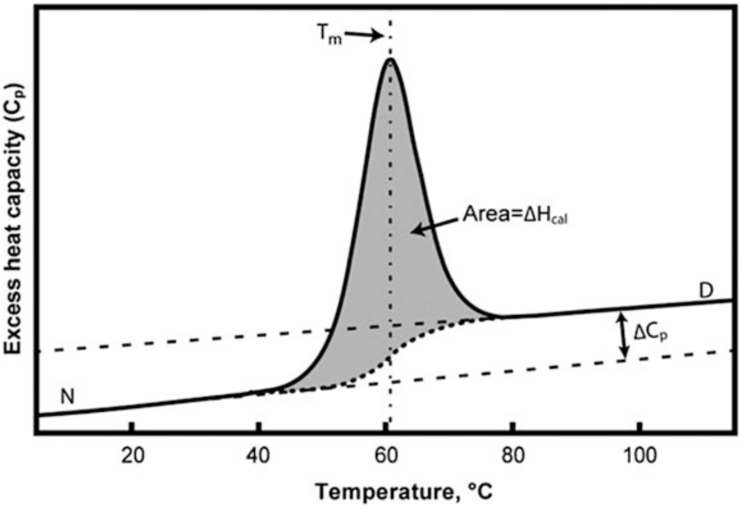

Fig. 2 A typical DSC thermogram. The DSC thermogram represents a plot generated by taking the temperature on X-axis and heat capacity (*C*p) on Y-axis. The sharp peak midpoint is T_m, and the area under the curve can be used to calculate other thermodynamic parameters. (Figure adapted with permission from [39])

4 Current Approaches to Studying Protein Folding and Stability

Today, many methods are available for studying proteins, including fluorescence spectroscopy, circular dichroism (CD) spectroscopy, NMR spectroscopy, and DSC [40–43]. CD uses the absorption difference in polarised light. Because a protein's supramolecular structure affects the CD in addition to its molecular structure, it is feasible to quantify the quantity of secondary and tertiary structures in a protein using the CD spectroscopy [44]. Tryptophan, tyrosine, and phenylalanine are fluorescent amino acids whose excitation and emission spectra depend on their surroundings [45]. For instance, buried tryptophan or tyrosine containing proteins do not show a higher fluorescence intensity [45]. When these residues are exposed to a polar solvent or submerged in the hydrophobic core of a protein, their fluorescences are different. Since the spectra of fluorescent amino acids vary as a protein unfolds, fluorescence can be utilised to monitor changes in the tertiary structure of a protein [46]. By measuring a protein's excess heat capacity as a function of temperature, DSC captures the thermal events that occur in a protein during heating. This is accomplished by comparing the energy requirements for raising the temperature of a protein sample and a reference by the same amount. Thus, DSC can track the unfolding of proteins by measuring the energy released or absorbed by the heat events accompanying protein denaturation. Only DSC allows for the direct determination of the enthalpy of denaturation from experimental data among the methods mentioned above [47]. Protein folding T_m is frequently measured using CD spectroscopy. There is virtually little value for CD spectroscopy in studying complex structures like-hairpin peptides [48]. Measuring CD spectra at various temperatures is essential to estimate the T_{m}. After this, the molar ellipticities recorded at particular wavelengths are plotted against temperature to ascertain the T_{m} [48].

Differential scanning calorimetry is a sophisticated method of determining transition temperatures. We may determine the thermodynamic properties of the folding transition using this method in addition to the folding-transition temperature [49]. Because the experimental output of DSC reflects the energetics of all conformations that become minimally occupied during thermal unfolding, it is a very effective method for studying protein folding and stability [50].

5 DSC as a Tool to Study the Protein Folding

Protein folding can be studied using DSC. The denaturation of a protein is detectable as an endothermic peak in a DSC curve as the device monitors heat capacity (C_{p}), which is maximum at the transition midpoint (T_{m}) [51]. Proteins with a higher T_{m} are more thermostable and are used as an indicator of thermostability [51]. The sharpness of the transition peak can be determined by calculating its breadth at half-peak weight, which is a sign of the cooperativity of the unfolding [52]. A narrow peak in the thermogram indicates multi-state denaturation, whereas a broad peak denotes a co-operative transition [53, 54]. The unfolding and stable intermediate states can be inferred by comparing the experimentally determined enthalpy and calculated enthalpy [54]. The scanning must be thermodynamically reversible, which means that the system must be in equilibrium and that the protein must not aggregate during or after the denaturation to obtain relevant parameters [7]. If aggregation occurs, more processes are going on besides the unfolding that impacts the data.

5.1 Folding of PBX DB (Pre-B-Cell Leukaemia Transcription Factor Homeodomain)

For small proteins, folding and unfolding are highly co-operative processes distinguished by a few short-lived or lack of thermodynamically stable partially folded intermediate states [55]. The structural component of a protein, due to local interactions, demonstrates two-state folding/unfolding activity and is referred to as a co-operative folding unit or co-operative equilibrium folding unit [55]. These proteins fold very quickly, i.e. in a microsecond time scale, and if they have any short-lived intermediate states, it is difficult to distinguish them between transition [56]. Large proteins that fold in a non-cooperative manner do not produce any distinct energy barriers; hence it is difficult to understand the folding pathway. The DSC is an elegant method which can be employed to study both small and large

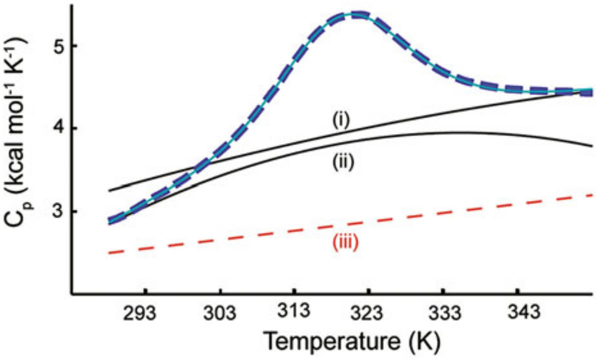

Fig. 3 DSC thermogram of PBX homeodomain fitted to the global folding model. The (i), (ii), and (iii) correspond to denatured- and native-state heat capacity baselines and a theoretical baseline calculated using a set of representative globular proteins, respectively. Experimental PBX-HD DSC data (thick dashed line) fit into a global folding model (thin continuous line). (Figure adapted with permission from [52])

protein folding pathways. A protein, pre-B-cell leukaemia transcription factor homeodomain (PBX DB), when studied DSC produces a broad DSC thermogram, which fits well into a two-state folding and linear heat capacity baselines (Fig. 3) [52]. In contrast to the established laws, the folded state has higher heat capacity than the denatured state of the protein. Upon denaturation, the protein is not compact and has a higher solvent accessible surface area (SASA). The hydrophobic residues exposed to the solvent may increase specific heat but not otherwise. The reason could be linked to the pre-unfolding states of the protein. This behaviour is also observed in other marginally stable proteins and suggests the specific heat capacity emanates from the temperature rather than solvent exposed hydrophobic residues. The DSC thermogram for PBX DB suggests that the folding process in this protein is endothermic pre-unfolding [52].

5.2 *Folding of Tetratricopeptide Repeats*

Tetratricopeptide repeats (TPRs) are 34 amino acid residue alpha-turn-helix peptides that belong to the all alpha-helical class of proteins [19]. They stack together and form a non-globular stable structure. Several studies have observed modular multistate folding as opposed to two-state folding. Two such peptides containing consensus linear repeats are, CTPRa2 to CTPRa10 and CTPR2 to CTPR3, which showed a reversible transition during the thermal unfolding process [57]. The DSC thermogram of these peptides showed a single sharp peak corresponding to the transition, and a reversible unfolding of all CTPRa and CTPR proteins was observed [58]. This was verified by repeating the experiments at various protein concentrations and putting samples through many cycles in the calorimeter cell. The data demonstrated that the T_m did not change noticeably, with a variation of just 0.4 K [57]. The excess heat capacity was calculated using heat capacity adjusted for protein content, buffer reference subtracted, and progress baselines removed. Using previously reported values of C_p presumed to be temperature-invariant, each trace was numerically integrated to produce the area under the heat capacity endotherm and a T_m [57].

5.3 Folding Mechanism of the Bovine Pancreatic Trypsin Inhibitor

DSC can be used to study the equilibrium folding process in proteins, where data is obtained through an analytical thermodynamic mode [59]. For instance, modifications of the 58 amino acid bovine pancreatic trypsin inhibitor (BPTI) with alanine at 21 and 27 positions exhibit co-operative two-state folding, and their thermodynamics were comparable to those of the wild-type variation, which contains 10 alanine residues [60]. The disulphide bonds in the protein were not altered; however, mutations were made at locations not necessary for defining tertiary structure. Intriguingly, a typical co-operative structure can be obtained even with the above-modified sequences. The DSC measurements for BPTI demonstrate that the high-temperature denatured state has a higher heat capacity level than the low-temperature native state [60]. Among other minor contributions, the exposure of hydrophobic groups to solvent in the denatured form is the leading cause of this significant and positive change in heat capacity during the unfolding [60]. Although it may be measured directly from the DSC thermogram, measuring ΔH at multiple T_m values yields a more accurate result (usually by pH variation). The values for the mutants were marginally lower than those for wild-type BPTI. However, this is likely owing to the substantial mutational change in these proteins (11 or 12 residues mutated to alanine) [60].

5.4 Studying Protein Aggregation

DSC calculates the heat capacity compared to the references. Such heat capacity variations are also observed in considerably less specialised condensation processes, such as protein aggregation, proving that they are not solely a characteristic of natural protein unfolding [61]. One such instance is insulin, which in solution, forms amyloid-like fibrils upon thermal denaturation [62]. Typically the DSC thermogram shows a positive heat capacity curve, but the aggregated insulin shows a negative effect (thermogram peak appears in reverse position in thermogram) (Fig. 4) [62]. In a normal folding process similar peak but towards a positive direction was found, and the negative peak in insulin indicates the sample is aggregated. According to this, condensed or closely packed polypeptides, resulting

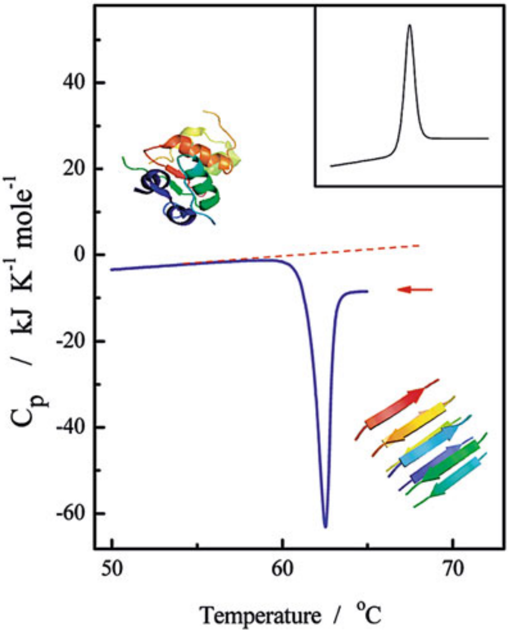

Fig. 4 A typical DSC thermogram of insulin in solution. The DSC thermogram represents a plot generated by taking the temperature on X-axis and heat capacity (C_p) on Y-axis. The properly folded form has a positive peak, whereas a negative peak represents the aggregated form of insulin (Figure adapted with permission from [63])

from specialised folding or non-specific aggregation, have a lower heat capacity than the unwound chain exposed to water. Heat capacity alterations reflect general changes in the polypeptide environment [64].

5.5 *Fast Folding Proteins*

The ultrafast folding of small protein domains that fold and unfold on a microsecond timescale has theoretical speed limitations of the process and on a time regime that is becoming more and more accessible to computer simulations [19]. It might seem odd at first that an equilibrium technique like DSC could be useful for research on systems that fold rapidly, but because of its fundamental properties and close relationship to the protein folding and unfolding kinetics, DSC measurements can theoretically be used to obtain data on the folding free energy surface and energetic barriers [65]. Under conditions of strong native bias, it was argued that the landscape theory would predict downward protein folding over modest or negligible energy barriers [65]. Some proteins may continue to fold downward under all equilibrium conditions, according to a phenomenon known as a one-state or global downhill folding [66]. There is only one ensemble of structures existing in this scenario for each set of parameters, and as parameters change, this ensemble's average properties alternate between native and denatured-like structures [66]. With this behaviour, it is possible to characterise the entire folding process at high resolution using an equilibrium technique like NMR [19]. Furthermore, it has been suggested that global downhill folding provides biological benefits by allowing proteins to act as molecular rheostats, constantly changing their structural composition [67].

The catalytically promiscuous isoform of glutathione S-transferase, GSTA1-1, has a unique low-temperature shoulder, which was revealed from the DSC data using a variable barrier model [68]. Through ligand binding, mutation, and CD investigations, it has been determined that this event is related to repacking of the C-terminal helix rather than any unfolding of the protein. After deconvoluting the DSC scan, analysis of the isolated low-temperature transition supports the idea that the folded C-terminal helix around the active site is sampled conformationally without encountering a significant free energy barrier. In contrast, at higher temperatures, the rest of the protein unfolds with clearly defined large barriers [68]. Comparing this dynamic flexibility to other substrate-specific isoforms that do not have the low-temperature shoulder, the enzyme was shown to be more promiscuous at its functional temperature. It would be fascinating to determine whether the dynamics of these helix motions are consistent with such a barrierless regime and why this flexibility is connected to the significant heat absorption detected by DSC.

The DSC of the GYF domain from human CD2BP2 provides another illustration of the distinct properties of native and denatured state heat capacity levels in a small, barely stable protein [69]. Given that data from DSC and CD thermal denaturation were fitted to a global two-state equilibrium model for this protein, it is fascinating to analyse the time scale of its folding kinetics. It will be interesting to look into the dependability and practical applications of the different methods of removing barriers from DSC data in more detail [69].

5.6 DSC as a Tool to Measure Barrier Heights in Protein Folding

The absence of experimental techniques to precisely determine the free energy barrier's height is a significant challenge to understanding protein folding events [65]. This is particularly disappointing because, as predicted by theory, if folding barriers are low, it might be able to directly resolve folding mechanisms. The DSC is an effective tool for extracting folding barriers from equilibrium DSC thermograms. A study used DSC data to calculate the thermodynamic barrier heights for 15 proteins [65].

6 DSC as a Tool to Determine Protein Stability

It is established that the formation of distinctive three-dimensional (3D) structures in small single-domain proteins is reversible, and the process is governed by thermodynamics [70, 71]. Therefore, it is essential to understand the thermodynamics of these processes. To accomplish this, it is necessary to make direct measurements of the effect of heat using highly sensitive calorimetric techniques like DSC. Thermal protein stability has two components: kinetic and equilibrium, and both can be quantified using DSC [19]. DSC can also evaluate the fundamental thermostability of proteins, or their susceptibility to heat denaturation, as shown by the measured T_m from the thermogram. When there is an irreversible denaturation, this T_m may have an apparent or an equilibrium thermodynamic value. It is crucial to distinguish between recorded thermostability and the degree of equilibrium stability relevant to physiological or experimental situations at low temperatures. If the values of $\Delta C_{\mathrm{pD\text{-}N}}$ in two proteins with an equal T_m value are not the same, their equilibrium stabilities could be significantly different at lower temperatures. The measured $\Delta H_{\mathrm{D\text{-}N}}$ and $\Delta C_{\mathrm{pD\text{-}N}}$ in standard equations are necessary for extrapolating equilibrium stability far from the T_m, where the stabilising free energy ($\Delta G_{\mathrm{D\text{-}N}}$) is zero for a basic reversible system.

$$\begin{aligned}\Delta G_{\mathrm{D-N}} = {} & \left[\Delta H_{\mathrm{D-N}} + \Delta C_{\mathrm{pD-N}}(T - T_{\mathrm{m}})\right] \\ & - \left[\Delta H_{\mathrm{D-N}}/T_{\mathrm{m}} + \Delta C_{\mathrm{pD-N}} \ \ln \ (T/T_{\mathrm{m}})\right]\end{aligned} \tag{6}$$

6.1 Advantages of DSC over Other Techniques in Studying Thermal Denaturation

DSC is a versatile technique to study thermal denaturation. Since it monitors the heat absorption directly rather than relying on changes in the spectroscopic signal (like thermal denaturation based on CD or fluorescence), it can better resolve numerous overlapping processes. To allow scans at higher temperatures, up to 140 °C without the sample boiling, a small amount of extra pressure is typically provided to the sample and reference cell during a DSC measurement [19]. This feature allows studying proteins obtained from thermophilic and hyper-thermophilic organisms with melting temperatures above 100 °C [72]. Biotechnological applications require understanding how this stability is attained (through equilibrium or kinetic mechanisms) and how certain sequences and structural features provide greater thermostability. The preferred approach to figure this out is DSC since it allows for a complete thermodynamic characterisation that can explain how the thermal equilibrium depends on temperature and, consequently, the process by which thermostability is achieved.

The remodelling of thioredoxin enzyme from extinct species using paleogenetics has also shown notable improvements in thermostability [19]. The T_{m} of these ancestral sequences was around 25 °C greater than that of modern *E. coli* or human thioredoxin, consistent with the environment gradually cooling over approximately 5 billion years [19]. The T_{m} of a protein is a known parameter that can reveal valuable characteristics, such as the likelihood that the protein to crystallise successfully for high-resolution structural work and the potential shelf life of the protein in a specific pharmaceutical formulation [73]. These correlations show that populating the native state is necessary for orderly crystal packing and growth and that the denatured state or unfolded intermediates are the most frequent sources of material for irreversible processes, such as aggregation, that result in functional loss. It is known that increasing T_{m} through conditions or stabilising additives is valuable to optimise for crystallisation success. Likewise, one of the primary methods used commercially to aid the structural determination of GPCR proteins is to increase the T_{m} by mutagenesis [74].

The most widely used methodology for high-throughput protein T_{m} measurements includes the use of a fluorescent reporter dye (SYPRO orange) in differential scanning fluorimetry (DSF), which increases fluorescence upon interaction with the hydrophobic groups that are usually buried in the protein core and serves as an indicator of unfolding [75]. It has been demonstrated that the values of T_{m} from the DSF and the DSC closely correlate. The absolute results from DSF are consistently lower, which may be caused by the reporter dye's destabilising effect or issues with the method of fitting to T_{m} [75]. The area and breadth of the enthalpic heat absorption peak may be utilised as additional markers for good crystal growth, although DSC may not have the throughput of fluorescence-based approaches.

Studies on protein stability and denaturation are also crucial in biopharmaceutical formulations due to the rise in protein-based treatments. Protein-based therapeutics are typically injected in small amounts from a highly concentrated solution; therefore, they must be produced in a way that allows the active protein to be stored without degrading [76]. Like other covalent changes, glycosylation and PEGylation of proteins have differential effects on their thermal stabilities [77]. Water can be excluded from proteins in several ways, with freeze-drying being the most popular, which can improve protein storage. DSC can assist in developing these processes and evaluating the protein's viability upon dissolving back into the solution [78]. It is frequently used in conjunction with DSF and other analytical procedures that use temperature, with the apparent T_{m} value serving as a crucial predictor of the desired shelf life [78].

The excipients used in a formulation often have indirect effects on the solvent that impact the protein's ability to self-associate, but in rare cases, they can function to stabilise the system by binding to the protein in its native state [19]. By shifting the equilibrium in a simple mass action effect, ligands will make the binding competent state more stable [30]. If this stabilisation can be quantified, for instance, in an increasing T_{m} shift, then the binding affinity of the ligand can be determined. This method of analysis of DSC data is well known and can determine exceedingly tight binding affinities. The change in T_{m} value can be significantly large on ligand binding [79]. When the metallo-chaperone Sco binds to Cu(II), its T_m rises by 23 °C, corresponding to $K_d = 3.5$ pM. Similarly, the binding of a variety of HIV protease inhibitors from clinical and experimental studies raised the protein's T_m by 6 °C to 22 °C, translating to nM or tighter affinity when extrapolated from the T_m to 25 °C. Two mutants (D25N and D29N) in the protease active site demonstrated noticeably lower inhibitor binding, as shown by the smaller T_m shifts detected in contrast to the wild-type.

6.2 DSC to Determine the Stability of Coacervation: Lysozyme and Heparin

Complex coacervation is the process of interaction of two molecules with opposing charges, usually both of which are macromolecules [80]. This process sometimes leads to the formation of a precipitate known as a complex flocculate. Thus, it is used in protein separation processes, resulting in phase separation between the complex and the bulk solution. It has been found that the formation of complexes often results in elevated thermal stability [81]. Lysozyme, a 14 kDa protein with a pI = 10.5, interacts with heparin glycosaminoglycans (GAGs) between the pH range 2 and 10, forming a coacervate or flocculate. Upon mixing, they form an insoluble, white complex. DSC was used, and thermograms were produced by scanning from 15 °C to 100 °C at a rate of 90 °C per h. Samples were first vacuum degassed for 5 min. The complexes had a lysozyme-heparin ratio of 5:1, with a protein content of about 5 mg/mL. The unfolding enthalpies and melting temperatures were calculated using Origin software [81]. DSC studies proved that this interaction might be potentially detrimental to thermal stability. Lysozyme stability decreased after complexation, which indicates that heparin has a stronger affinity for the unfolded state than the native state. Other proteins may experience similar instability when they come into contact with highly charged polymeric substances or surfaces [81].

6.3 Structural Transitions in Recombinant Human IFNα2a as a Function of pH and Temperature

Interferons are cytokines that have anti-viral, anti-proliferative, and immunoregulatory effects in humans [82]. IFNα2a is one interferon subtype, having a molecular weight of ~19 kDa, 165 amino acids, four cysteines, and two disulphide bonds. The 3D solution structure of IFN2a was determined by NMR, showing that it is an all-helical protein with six α-helices [83]. During unfolding, IFN2a produces a range of partially unfolded states and intermediates, which are sensitive to the pH and temperature of the solution. These partially unfolded conformations significantly influence the aggregation and subsequent long-term stability of IFN2a in solution.

The structural characteristics of IFN2a were investigated using Trp fluorescence emission, fluorescence quenching, near- and far-UV CD spectroscopy, and DSC at pH values (2.0–7.4) and temperatures (5–80 °C). DSC data showed that the beginning of unfolding was about 55 °C for pH 7.4 and 60 °C for pH 5.0, respectively [83]. The changes in the tertiary and secondary structures of IFN2a at moderate temperatures were indicated at lower pH (pH 3.0 and 4.0) by an increase in heat capacity throughout a wide temperature range in the DSC experimental studies. However, at pH 5.0 and 7.4 in the lower temperature range (15–50 °C), DSC was insensitive to the minor modifications shown in the tertiary structure of IFN2a [83].

6.4 Analysing Thermal Stability of Therapeutic Monoclonal Antibodies Using DSC

High selectivity and specificity monoclonal antibodies (mAbs) make up a significant and expanding percentage of the biotherapeutics industry [84]. The IgG class includes the bulk of commercially available mAbs. IgGs have 16 inter- or intramolecular disulphide linkages and are made up of two heavy and light chains. Each heavy chain is disulphide coupled to a light chain, and disulphide bonds connect the two heavy chains. IgGs have crystallisable (Fc) and antigen-binding (Fab) domains; the Fab attaches to the antigen, while the Fc binds to Fc receptors, which control immune responses [85]. Knowledge of therapeutic protein stability has become more crucial as the development of biopharmaceutical products has grown rapidly. More importantly, screening a buffer that retains the stability of such formulations is vital. Several techniques have been established to study the stability of therapeutic antibody formulations, but DSC has been successfully used in high-throughput screening in these experiments.

Post-translational modifications affect the stability of a protein by either increasing or decreasing the stability of mAbs [86]. A study recorded the DSC thermograms to study stability in the presence of glycosylated or deglycosylated human IgG1 antibodies and suggested that the antibodies have three peaks in their DSC thermogram [87]. The first peak with the lowest T_m represents the thermal transition of the CH2 domain in the Fc region. The second peak with the largest peak height represents the T_m of the Fab region. The third peak with the highest T_m is the contribution of the CH3 domain in the Fc region (Fig. 5) [88, 89]. Sometimes, the thermal transitions between the Fab region and the CH3 domain are so near as to combine the last two peaks in the thermogram into a single peak. Similar characteristic peaks were observed when mAbs, mAb1, mAb2, and mAb3 were subjected to DSC. In addition, the deglycosylated forms of the above antibodies have lower transition temperatures ($6 \sim 8$ °C) at CH2 transition [89]. This finding supports earlier findings by Mimura et al. and implies that the oligosaccharide chain stabilises the CH2 domain during temperature-induced unfolding while not affecting other

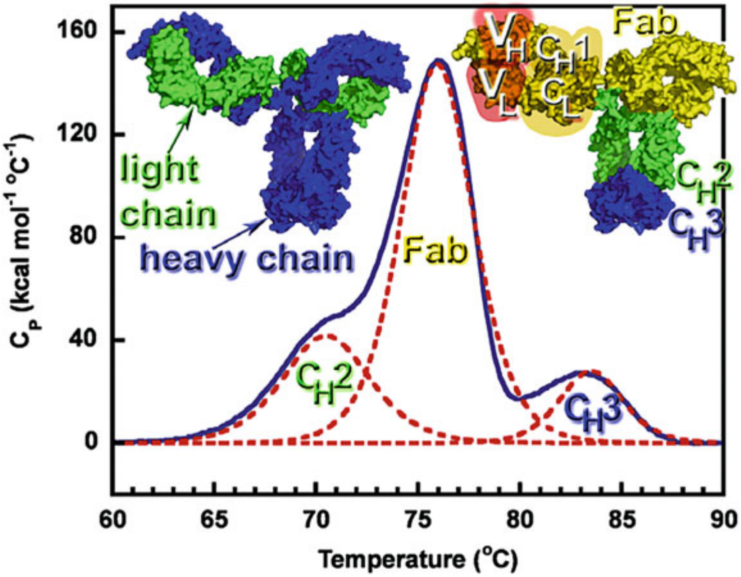

Fig. 5 Typical thermal denaturation thermogram for monoclonal antibodies. CH2 represents the constant heavy chain2, and CH3 represents the constant heavy chain3; both belong to the Fc region of an antibody (Figure adapted with permission from [88])

mAb domains [90]. The interactions between the two oligosaccharide chains or the oligosaccharide chain and the CH2 domain potentially stabilise the compound [89].

Another work studied the role of variable domains on the stability of humanised IgG1 mAbs. The study recorded the DSC thermogram for Fab and Fc fragments of three antibodies [91]. With a few exceptions, the entire antibody's DSC thermogram shows two peaks, and the transition with the higher experimental enthalpy includes the contribution from the Fab fragments. Even for Fab fragments originating from the same human germline, the apparent melting temperatures differed substantially, despite the measured enthalpy for all three investigated Fab fragments being similar [91]. The IgGs containing variable domains generated from complementarity determining regions (CDRs) grafting and humanisation could destabilise the Fab fragment with respect to the CH3 domain. The first transition represents the unfolding of the Fab fragment and the CH2 domain, while the second transition represents the unfolding of the CH3 domain [91]. In other cases, the DSC profile can also show three transitions, with the Fab unfolding at a different temperature than the CH2 and CH3 domains melting. If the model above cannot characterise the DSC profile of a humanised IgG1 monoclonal antibody, it may indicate considerable structural heterogeneity and/or disruption of the Fab co-operative unfolding. Low stability or heterogeneity of the Fab fragment might make long-term storage or manufacturing consistency difficult [91].

6.5 Effects of Electrostatic Repulsions on the Stability and Aggregation of the NIST Monoclonal Antibody

The NIST monoclonal antibody (NISTmAb) is a humanised recombinant IgG1 generated in the suspension culture of murine cells. The stability and aggregation tendency of NISTmAb at four different pHs (5, 6, 7, and 8) were investigated with or without NaCl. To calculate the ΔG and T_m of different domains, HDX-MS and DSC were used, respectively [92]. The T_m and the temperature at which NISTmAb begins to aggregate were determined using nano-DSF. During DSC studies, the temperature differences between the reference and sample cells were continuously recorded and converted to power units. Samples were heated at a rate of 1 °C/min from 25 °C to 110 °C. In the presence of NaCl, NISTmAb was more conformationally stable at a pH closer to its pI than at a pH distant from its pI [92]. The stabilising effects were not localised; they were global. However, the onset of aggregation temperature experiments revealed that NISTmAb is less likely to aggregate at a pH far from its pI, especially when NaCl is absent. This contradictory result, i.e. high conformational stability yet a high aggregation tendency close to its pI value, can be justified by intra- and intermolecular electrostatic repulsions using the Lumry-Eyring model.

7 Conclusions

In the past few decades, the analysis of DSC data using defined thermodynamic models has been essential for improving our understanding of protein stability. Along with other spectroscopic methods, it emerged as a valuable tool for understanding and investigating the unfolding and stability of proteins [42, 43, 93, 94]. This method is used for a single sample and is helpful in high-throughput screening. The applications are not limited to only therapeutic molecules but are also valuable in studying biomolecular interactions and antibody stability in the presence of polysaccharides. Ultrafast folding proteins have made it possible to use DSC to examine the whole folding and unfolding free energy landscape with inherently broad conformational ensembles and minimally co-operative folding, including those conformations near the top of the folding barrier. Studies indicate that the plasma proteome gives a signature thermogram, mainly due to the denaturation of major plasma proteins, and this thermogram varies from healthy to patient plasma samples. The DSC can be employed in the health and diagnostic industry; however, further studies are needed to establish this as a reliable diagnostic tool for studying plasma proteome. Drug screening for a particular target has also been reported using DSC. More importantly, automated DSCs can be employed for large-scale screening of drug molecules against a specified target. Apart from this, the method can also be used in food authenticity testing. Linking DSC with other instruments can provide real-time data. For instance, a combination of FTIR or NMR with DSC can give the structure of a biomolecule along with its thermal stability. Moreover, the combination of SAXS/DSC can be employed to study the crystallisation of polymers. The main limitations of the DSC are reproducibility of data across various labs, as the data varies with operational factors such as the operator. The thermodynamic values may change even in small changes in the system or improper setup. The purity of pharmaceutical products obtained through this method also needs more validation as the method is sensitive to contamination. It is also challenging to use this method to differentiate materials with the same transition temperatures. Identifying the type of molecules (e.g. using FTIR) and separation (e.g. using HPLC) can resolve such limitations. Recent trends suggest that the method could be useful in disease diagnosis and metal alloy industries to check thermal stability. In conclusion, the DSC has emerged as an effective technique that helps analyse material properties such as glass transition temperature, melting, crystallisation, specific heat capacity, purity, oxidation behaviour, thermal stability, etc., for a wide range of materials, including proteins, polymers, plastics, pharmaceuticals, food, organic compounds, chemicals, petroleum, biological samples, and many more.

References

1. C. Demetzos, Differential scanning calorimetry (DSC): a tool to study the thermal behavior of lipid bilayers and liposomal stability. J. Liposome Res. **18**(3), 159–173 (2008)
2. J. Müllerová, Thermal degradation of polymeric material based on cellulose by differential scanning calorimetry (DSC). Int. Multidiscip. Sci. GeoConference **1**, 765–770 (2016)
3. Q. Chen, R. Yang, B. Zhao, Y. Li, S. Wang, H. Wu, Y. Zhuo, C. Chen, Investigation of heat of biomass pyrolysis and secondary reactions by simultaneous thermogravimetry and differential scanning calorimetry. Fuel **134**, 467–476 (2014)
4. S.-D. Clas, C.R. Dalton, B.C. Hancock, Differential scanning calorimetry: applications in drug development. Pharm. Sci. Technol. Today **2**(8), 311–320 (1999)
5. K. Lukas, P.K. LeMair, Differential scanning calorimetry: fundamental overview. Reson. J. Sci. Educ. **14**(8), 807 (2009)
6. A. Baylon, É. Stauffer, O. Delémont, Evaluation of the self-heating tendency of vegetable oils by differential scanning calorimetry. J. Forensic Sci. **53**(6), 1334–1343 (2008)
7. P. Gill, T.T. Moghadam, B. Ranjbar, Differential scanning calorimetry techniques: applications in biology and nanoscience. J. Biomol. Tech. **21**(4), 167 (2010)
8. Y. Dong, Y. Ruan, H. Wang, Y. Zhao, D. Bi, Studies on glass transition temperature of chitosan with four techniques. J. Appl. Polym. Sci. **93**(4), 1553–1558 (2004)
9. C.E. Bernardes, A. Joseph, M.E.M. da Piedade, Some practical aspects of heat capacity determination by differential scanning calorimetry. Thermochim. Acta **687**, 178574 (2020)
10. O. Matsarskaia, L. Bühl, C. Beck, M. Grimaldo, R. Schweins, F. Zhang, T. Seydel, F. Schreiber, F. Roosen-Runge, Evolution of the structure and dynamics of bovine serum albumin induced by thermal denaturation. Phys. Chem. Chem. Phys. **22**(33), 18507–18517 (2020)
11. I.B. Durowoju, K.S. Bhandal, J. Hu, B. Carpick, M. Kirkitadze, Differential scanning calorimetry—a method for assessing the thermal stability and conformation of protein antigen. J. Vis. Exp. **121**, e55262 (2017)
12. T. Krell, Microcalorimetry: a response to challenges in modern biotechnology. Microb. Biotechnol. **1**(2), 126–136 (2008)
13. S.C. Mraw, Mathematical treatment of heat flow in differential scanning calorimetry and differential thermal analysis instruments. Rev. Sci. Instrum. **53**(2), 228–231 (1982)

14. I. Passi, S. Salwan, S.S. Ganti, B. Kumar, Differential scanning calorimetry has emerged as a key analytical tool in the thermal analysis of pharmaceutical formulations. Curr. Pharm. Des. **28**(37), 3082–3084 (2022)
15. S. Mazurenko, A. Kunka, K. Beerens, C.M. Johnson, J. Damborsky, Z. Prokop, Exploration of protein unfolding by modelling calorimetry data from reheating. Sci. Rep. **7**(1), 1–14 (2017)
16. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Springer Nature, Singapore, 2020)
17. A. Michnik, Thermal stability of bovine serum albumin DSC study. J. Therm. Anal. Calorim. **71**(2), 509–519 (2003)
18. G. Valentini, M. Maggi, A.L. Pey, Protein stability, folding and misfolding in human PGK1 deficiency. Biomol. Ther. **3**(4), 1030–1052 (2013)
19. C.M. Johnson, Differential scanning calorimetry as a tool for protein folding and stability. Arch. Biochem. Biophys. **531**(1-2), 100–109 (2013)
20. Y. Yang, Z. Su, G. Ma, S. Zhang, Characterization and stabilization in process development and product formulation for super large proteinaceous particles. Eng. Life Sci. **20**(11), 451–465 (2020)
21. A. Watts, *Protein-Lipid Interactions* (Elsevier, Amsterdam, 1993)
22. L. Burton, R. Gandhi, G. Duke, M. Paborji, Use of microcalorimetry and its correlation with size exclusion chromatography for rapid screening of the physical stability of large pharmaceutical proteins in solution. Pharm. Dev. Technol. **12**(3), 265–273 (2007)
23. C.H. Spink, Differential scanning calorimetry. Methods Cell Biol. **84**, 115–141 (2008)
24. A.W. Vermeer, W. Norde, The thermal stability of immunoglobulin: unfolding and aggregation of a multi-domain protein. Biophys. J. **78**(1), 394–404 (2000)
25. P. Kaur, M. Singh, P. Birwal, Differential Scanning Calorimetry (DSC) for the measurement of food thermal characteristics and its relation to composition and structure, in *Techniques to Measure Food Safety and Quality*, ed. by M.S. Khan, M. Shafiur Rahman, (Springer, Cham, 2021), pp. 283–328
26. P. Gabbott, A practical introduction to differential scanning calorimetry, in *Principles and Applications of Thermal Analysis*, (Wiley, Oxford, 2008), pp. 1–50
27. J.D. Menczel, L. Judovits, R.B. Prime, H.E. Bair, M. Reading, S. Swier, Differential scanning calorimetry (DSC). Ther. Anal. Poly. Fundam. Appl. **7**, 239 (2009)
28. M. Luisi, Characterizing the Measurement Uncertainty of a High-Temperature Heat Flux Differential Scanning Calorimeter, Master of Applied Science Thesis, Graz University of Technology. VITA AUCTORIS (2014)
29. P.G. Laye, *Differential Thermal Analysis and Differential Scanning Calorimetry* (Royal Society of Chemistry, Cambridge, 2002)
30. G. Höhne, J. McNaughton, W. Hemminger, H.-J. Flammersheim, H.-J. Flammersheim, *Differential Scanning Calorimetry* (Springer Science & Business Media, Berlin, 2003)
31. G. Johari, A. Hallbrucker, E. Mayer, Thermal behavior of several hyperquenched organic glasses. J. Phys. Chem. **93**(6), 2648–2652 (1989)
32. S. Wouters, F. Demir, L. Beenaerts, G. Van Assche, Calibration and performance of a fast-scanning DSC—project RHC. Thermochim. Acta **530**, 64–72 (2012)
33. C.R. Quick, P. Dumitraschkewitz, J.E. Schawe, S. Pogatscher, Fast differential scanning calorimetry to mimic additive manufacturing processing: specific heat capacity analysis of aluminium alloys. J. Therm. Anal. Calorim. **148**, 651–662 (2023)
34. J.E. Schawe, T. Hütter, C. Heitz, I. Alig, D. Lellinger, Stochastic temperature modulation: a new technique in temperature-modulated DSC. Thermochim. Acta **446**(1–2), 147–155 (2006)
35. M.H. Chiu, E.J. Prenner, Differential scanning calorimetry: an invaluable tool for a detailed thermodynamic characterization of macromolecules and their interactions. J. Pharm. Bioallied Sci. **3**(1), 39 (2011)
36. V. Plotnikov, A. Rochalski, M. Brandts, J.F. Brandts, S. Williston, V. Frasca, L.-N. Lin, An autosampling differential scanning calorimeter instrument for studying molecular interactions. Assay Drug Dev. Technol. **1**(1), 83–90 (2002)

37. E. Freire, Differential scanning calorimetry. Methods Mol. Biol. **40**, 191–218 (1995)
38. G. Privalov, V. Kavina, E. Freire, P.L. Privalov, Precise scanning calorimeter for studying thermal properties of biological macromolecules in dilute solution. Anal. Biochem. **232**(1), 79–85 (1995)
39. Y. Zhang, M.S. Ardejani, Differential scanning calorimetry to quantify the stability of protein cages. Methods Mol. Biol. **1252**, 101–113 (2015)
40. S.R. Martin, M.J. Schilstra, Circular dichroism and its application to the study of biomolecules. Methods Cell Biol. **84**, 263–293 (2008)
41. T. Tripathi, Calculation of thermodynamic parameters of protein unfolding using far-ultraviolet circular dichroism. J. Protein. Proteomics **4**(2), 85–91 (2013)
42. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods*, 1st edn. (Academic Press, Cambridge, MA, 2022)
43. P. Saudagar, T. Tripathi, *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, 1st edn. (Academic Press, San Diego, 2023)
44. S.M. Kelly, T.J. Jess, N.C. Price, How to study proteins by circular dichroism. Biochim. Biophys. Acta **1751**(2), 119–139 (2005)
45. A.R. Katritzky, T. Narindoshvili, Fluorescent amino acids: advances in protein-extrinsic fluorophores. Org. Biomol. Chem. **7**(4), 627–634 (2009)
46. Y. Chen, M.D. Barkley, Toward understanding tryptophan fluorescence in proteins. Biochemistry **37**(28), 9976–9982 (1998)
47. D.I. Markov, E.O. Zubov, O.P. Nikolaeva, B.I. Kurganov, D.I. Levitsky, Thermal denaturation and aggregation of myosin subfragment 1 isoforms with different essential light chains. Int. J. Mol. Sci. **11**(11), 4194–4226 (2010)
48. M. Kjaergaard, A.B. Nørholm, R. Hendus-Altenburger, S.F. Pedersen, F.M. Poulsen, B.B. Kragelund, Temperature-dependent structural changes in intrinsically disordered proteins: formation of α–helices or loss of polyproline II? Protein Sci. **19**(8), 1555–1564 (2010)
49. C. Nicolini, R. Ravindra, B. Ludolph, R. Winter, Characterization of the temperature-and pressure-induced inverse and reentrant transition of the minimum elastin-like polypeptide GVG (VPGVG) by DSC, PPC, CD, and FT-IR spectroscopy. Biophys. J. **86**(3), 1385–1392 (2004)
50. L.A. Campos, M. Bueno, J. Lopez-Llano, M.Á. Jiménez, J. Sancho, Structure of stable protein folding intermediates by equilibrium ϕ-analysis: the apoflavodoxin thermal intermediate. J. Mol. Biol. **344**(1), 239–255 (2004)
51. J. Boye, Differential scanning calorimetry in the analysis of foods. Food Sci. Technol. **138**(3), 1837 (2004)
52. P. Farber, H. Darmawan, T. Sprules, A. Mittermaier, Analyzing protein folding cooperativity by differential scanning calorimetry and NMR spectroscopy. J. Am. Chem. Soc. **132**(17), 6214–6222 (2010)
53. X.L. Qi, S. Brownlow, C. Holt, P. Sellers, Multi-state thermal unfolding and aggregation of β-lactoglobulin A. Biochem. Soc. Trans. **23**(1), 74S (1995)
54. A. Moosavi-Movahedi, J. Chamani, M. Gharanfoli, G. Hakimelahi, Differential scanning calorimetric study of the molten globule state of cytochrome c induced by sodium n-dodecyl sulfate. Thermochim. Acta **409**(2), 137–144 (2004)
55. K.P. Murphy, E. Freire, Thermodynamics of structural stability and cooperative folding behavior in proteins. Adv. Protein Chem. **43**, 313–361 (1992)
56. A.R. Fersht, On the simulation of protein folding by short time scale molecular dynamics and distributed computing. Proc. Natl. Acad. Sci. **99**(22), 14122–14125 (2002)
57. J. Phillips, Y. Javadi, C. Millership, E. Main, Modulation of the multistate folding of designed TPR proteins through intrinsic and extrinsic factors. Protein Sci. **21**(3), 327–338 (2012)
58. C.S.K. Kim, Recombinant Proteins for Biomedical Applications, Doctoral Dissertations, Virginia Tech, 2020
59. T. Jyothi, S. Sinha, S.A. Singh, A. Surolia, A.A. Rao, Napin from Brassica juncea: thermodynamic and structural analysis of stability. Biochim. Biophys. Acta **1774**(7), 907–919 (2007)

60. D. Krowarsch, J. Otlewski, Amino-acid substitutions at the fully exposed P1 site of bovine pancreatic trypsin inhibitor affect its stability. Protein Sci. **10**(4), 715–724 (2001)
61. W.F. Weiss IV, T.M. Young, C.J. Roberts, Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. J. Pharm. Sci. **98**(4), 1246–1277 (2009)
62. W. Dzwolak, R. Ravindra, R. Winter, Hydration and structure—the two sides of the insulin aggregation process. Phys. Chem. Chem. Phys. **6**(8), 1938–1943 (2004)
63. A. Cooper, Protein heat capacity: an anomaly that maybe never was. J. Phys. Chem. Lett. **1**(22), 3298–3304 (2010)
64. A. Cooper, Heat capacity effects in protein folding and ligand binding: a re-evaluation of the role of water in biomolecular thermodynamics. Biophys. Chem. **115**(2–3), 89–97 (2005)
65. A.N. Naganathan, J.M. Sanchez-Ruiz, V. Munoz, Direct measurement of barrier heights in protein folding. J. Am. Chem. Soc. **127**(51), 17970–17971 (2005)
66. M.M. Garcia-Mira, M. Sadqi, N. Fischer, J.M. Sanchez-Ruiz, V. Munoz, Experimental identification of downhill protein folding. Science **298**(5601), 2191–2195 (2002)
67. A.N. Naganathan, V. Muñoz, Thermodynamics of downhill folding: multi-probe analysis of PDD, a protein that folds over a marginal free energy barrier. J. Phys. Chem. B **118**(30), 8982–8994 (2014)
68. M.T. Honaker, M. Acchione, J.P. Sumida, W.M. Atkins, Ensemble perspective for catalytic promiscuity: calorimetric analysis of the active site conformational landscape of a detoxification enzyme. J. Biol. Chem. **286**(49), 42770–42776 (2011)
69. M. Andujar-Sanchez, E.S. Cobos, I. Luque, J.C. Martinez, Thermodynamic impact of embedded water molecules in the unfolding of human CD2BP2-GYF domain. J. Phys. Chem. B **116**(24), 7168–7175 (2012)
70. V.V. Mozhaev, Mechanism-based strategies for protein thermostabilization. Trends Biotechnol. **11**(3), 88–95 (1993)
71. J. Fitter, The perspectives of studying multi-domain protein folding. Cell. Mol. Life Sci. **66**(10), 1672–1681 (2009)
72. G. Feller, Protein stability and enzyme activity at extreme biological temperatures. J. Phys. Condens. Matter **22**(32), 323101 (2010)
73. J. Liu, Physical characterization of pharmaceutical formulations in frozen and freeze-dried solid states: techniques and applications in freeze-drying development. Pharm. Dev. Technol. **11**(1), 3–28 (2006)
74. J.-P. Renaud, C.-w. Chung, U.H. Danielson, U. Egner, M. Hennig, R.E. Hubbard, H. Nar, Biophysics in drug discovery: impact, challenges and opportunities. Nat. Rev. Drug Discov. **15**(10), 679–698 (2016)
75. T. Wu, J. Yu, Z. Gale-Day, A. Woo, A. Suresh, M. Hornsby, J.E. Gestwicki, Three essential resources to improve differential scanning fluorimetry (DSF) experiments. BioRxiv (2020). https://doi.org/10.1101/2020.03.22.002543
76. Y.-F. Maa, S.J. Prestrelski, Biopharmaceutical powders particle formation and formulation considerations. Curr. Pharm. Biotechnol. **1**(3), 283–302 (2000)
77. R.J. Solá, K. Griebenow, Effects of glycosylation on the stability of protein pharmaceuticals. J. Pharm. Sci. **98**(4), 1223–1245 (2009)
78. W.C. Blocher McTigue, S.L. Perry, Protein encapsulation using complex coacervates: what nature has to teach us. Small **16**(27), 1907671 (2020)
79. D. Witkowska, M. Rowińska-Żyrek, Biophysical approaches for the study of metal-protein interactions. J. Inorg. Biochem. **199**, 110783 (2019)
80. A.B. Kayitmazer, Thermodynamics of complex coacervation. Adv. Colloid Interf. Sci. **239**, 169–177 (2017)
81. M. van de Weert, M.B. Andersen, S. Frokjaer, Complex coacervation of lysozyme and heparin: complex characterization and protein stability. Pharm. Res. **21**(12), 2354–2359 (2004)
82. E. Bartholome, F. Roufosse, Immunoregulatory properties of type-I interferons: relevance to multiple sclerosis and the hypereosinophilic syndrome. Acta Clin. Belg. **52**(6), 350–359 (1997)

83. V.K. Sharma, D.S. Kalonia, Temperature-and pH-induced multiple partially unfolded states of recombinant human interferon-α2a: possible implications in protein stability. Pharm. Res. **20**(11), 1721–1729 (2003)
84. J.V. Rodarte, C. Baehr, D. Hicks, T.L. Liban, C. Weidle, P.B. Rupert, R. Jahan, A. Wall, A.T. McGuire, R.K. Strong, Structures of drug-specific monoclonal antibodies bound to opioids and nicotine reveal a common mode of binding. Structure **31**(1), 20–32.e5 (2023)
85. H. Liu, G.-G. Bulseco, J. Sun, Effect of posttranslational modifications on the thermal stability of a recombinant monoclonal antibody. Immunol. Lett. **106**(2), 144–153 (2006)
86. K. Zheng, M. Yarmarkovich, C. Bantog, R. Bayer, T.W. Patapoff, Influence of glycosylation pattern on the molecular properties of monoclonal antibodies. MAbs **6**, 649–658 (2014)
87. A. Niedziela-Majka, E. Kan, P. Weissburg, U. Mehra, S. Sellers, R. Sakowicz, High-throughput screening of formulations to optimize the thermal stability of a therapeutic monoclonal antibody. J. Biomol. Screen. **20**(4), 552–559 (2015)
88. E. Garber, S.J. Demarest, A broad range of fab stabilities within a host of therapeutic IgGs. Biochem. Biophys. Res. Commun. **355**(3), 751–757 (2007)
89. K. Zheng, C. Bantog, R. Bayer, The impact of glycosylation on monoclonal antibody conformation and stability. MAbs **3**, 568–576 (2011)
90. Y. Mimura, S. Church, R. Ghirlando, P. Ashton, S. Dong, M. Goodall, J. Lund, R. Jefferis, The influence of glycosylation on the thermal stability and effector function expression of human IgG1-fc: properties of a series of truncated glycoforms. Mol. Immunol. **37**(12–13), 697–706 (2000)
91. R.M. Ionescu, J. Vlasak, C. Price, M. Kirchmeier, Contribution of variable domains to the stability of humanized IgG1 monoclonal antibodies. J. Pharm. Sci. **97**(4), 1414–1426 (2008)
92. Y. Hamuro, M.G. Derebe, S. Venkataramani, J.F. Nemeth, The effects of intramolecular and intermolecular electrostatic repulsions on the stability and aggregation of NISTmAb revealed by HDX-MS, DSC, and nanoDSF. Protein Sci. **30**(8), 1686–1700 (2021)
93. S. Sasidharan, N. Nag, T. Tripathi, P. Saudagar, Experimental methods to study the thermodynamics of protein–protein interactions, in *Advances in Protein Molecular and Structural Biology Methods*, ed. by T. Tripathi, V.K. Dubey, (Academic Press, Cambridge, MA, 2022), pp. 103–114
94. N. Nag, S. Sasidharan, P. Saudagar, T. Tripathi, Fundamentals of spectroscopy for biomolecular structure and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 1–35

Nuclear Magnetic Resonance Spectroscopy to Analyse Protein Folding and Dynamics

Nikita V. Saibo, Soumendu Boral, Rituparna Saha, Amit K. Das, and Soumya De

Abstract Proteins are nanoscale machines that perform all the work in living systems. Their function depends on their three-dimensional (3D) structure. These nanomachines are manufactured as linear polymeric chains in the living cell and self-fold into the complex 3D structures that are required for their functions. Mutations in proteins (manufacturing defects) may result in misfolding and aberrant functions, leading to various diseases. Hence, understanding the process of protein folding is very important. Several experimental techniques have been used to study protein folding. In this chapter, we will discuss solution-state NMR spectroscopy as a versatile technique to study the mechanism, thermodynamics, and kinetics of protein folding. We describe the basics and the applications of various NMR methods and discuss the recent developments in this technique for studying protein folding.

Keywords Thermodynamics · Kinetics · Energy landscape · Misfolded proteins · Protein aggregation · Spectroscopy · Real-time · Timescale · Folding-unfolding transition · Equilibrium

1 Introduction

Proteins are nanomachines that carry out almost all the functions in living systems. The activity of a folded protein is defined by its three-dimensional (3D) structure. Understanding how proteins fold into their functional three-dimensional form from a linear polymeric chain is the crux of the protein folding problem [1]. Levinthal's paradox points out that a polypeptide chain (>100 residues) will require a large amount of time (>10^{27} years) to fold if it randomly samples all possible

N. V. Saibo · S. Boral · S. De (✉)
School of Bioscience, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India
e-mail: somde@iitkgp.ac.in

R. Saha · A. K. Das
Department of Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_4

conformations [2]. However, most proteins fold within a few seconds. Thus, it was clear that proteins undergo a biased search and sample a very small number of conformations to reach the final folded state. This was very elegantly depicted by the theoretical constructs of the energy landscape and folding funnel [3, 4]. The energy landscape is described as a partially rugged funnel where the folding is guided down to the low-energy native state by overcoming the entropy of the unfolded states [3–5].

This description of protein folding provided the necessary framework to ask specific questions regarding the folding process that can be answered by experiments, such as What is the kinetics of protein folding? Are there any folding intermediates? Which part of a protein folds first, or which secondary structures form first? And finally, do proteins undergo folding in the same manner in vivo as they do in vitro? The in vivo crowded environment might play a role in changing the energy landscape of protein folding through its influence on protein stability. Protein folding and unfolding are essential events in the cell, thereby understanding their mechanism at an atomic level has fundamental biological relevance. Protein folding is also prone to errors [6]. Misfolding of proteins is involved in diseases, often with mutations that may stabilize the misfolded states or destabilize the correctly folded form [7].

Misfolded proteins with exposed hydrophobic residues may accumulate and can form potentially toxic aggregates [8]. Protein folding in vivo is thereby assisted by chaperone proteins such as Hsp60, Hsp70, and Hsp90 systems that function as molecular chaperones in de novo folding by shielding the exposed hydrophobic residues of the proteins in their non-native conformations [9]. The toxic protein aggregates may sequester components of chaperone networks, as seen in Hsp40 co-chaperones [10, 11]. Certain mutations can lead to dysfunctional metastable proteins that are prone to degradation, e.g. cystic fibrosis, where the mutations in cystic fibrosis transmembrane conductance regulator (CFTR) cause the protein to be misfolded and targeted for degradation [12]. Misfolding can also lead to improper subcellular localization and may result in a loss of function of the protein in its correct location, e.g. liver damage and emphysema due to mutation of α1-antitrypsin, a secreted protease inhibitor [13, 14]. Aggregation of toxic metastable proteins is associated with various neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, Huntington's disease, type II diabetes, and certain forms of heart diseases [7, 8, 15–18]. It is still unclear how these aggregates form or the propensity of the misfolded proteins to form aggregates. In order to identify how protein folding and misfolding events may lead to diseases, it becomes imperative to understand the molecular descriptions of the protein folding pathways.

Experimental techniques such as nuclear magnetic resonance (NMR) spectroscopy, fluorescence resonance energy transfer (FRET), atomic force microscopy (AFM), mass spectrometry, and circular dichroism (CD) have provided important insights into the folding–unfolding mechanisms of proteins [19]. Among these techniques, NMR spectroscopy stands out as it can provide information at atomic resolution. NMR spectroscopy has played a crucial role in the biophysical studies of protein folding. While other spectroscopic methods, such as CD or fluorescence, are instrumental in defining the kinetics or thermodynamics of the folding of a protein, they do not provide detailed structural information on the folding intermediates. In contrast, NMR can present residue-wise structural information on the folding intermediates [20]. Clever use of stopped-flow techniques with NMR, as well as the development of fast acquisition methods, has enabled observing folding events in real time by NMR. Also, hydrogen exchange and other NMR dynamics experiments provide structural and thermodynamic information on higher energy conformations with a very small population. In this chapter, we will discuss the various methods that have been developed in NMR spectroscopy to study protein folding.

2 Studies of Protein Folding and Unfolding at Equilibrium

NMR spectroscopy is a versatile technique that can be used to determine the structures of proteins in a solution [21–25], study protein dynamics in multiple timescales [26–30], study oligomerization and aggregation of proteins [31], as well as study intrinsically disordered proteins [32–34]. The atoms of hydrogen, nitrogen, and carbon are the typical NMR probes, which are distributed throughout a protein, and provide comprehensive structural and dynamic information. NMR-active nuclei (^{1}H, ^{13}C, or ^{15}N) in a protein have distinct chemical shifts in the folded and unfolded states [35]. The NMR timescale of exchange [26, 28] is defined as follows:

$$\begin{array}{rl} \text{Slow exchange} & k_{\text{ex}} < \Delta\omega \\ \text{Intermediate exchange} & k_{\text{ex}} \sim \Delta\omega \\ \text{Fast exchange} & k_{\text{ex}} > \Delta\omega \end{array}$$

where k_{ex} is the rate of exchange, and $\Delta\omega$ is the chemical shift difference between the folded and unfolded states. In partially denaturing conditions, these states of a protein interconvert in the slow NMR timescale. For example, the trypsin inhibitor HPI exchanges (k_{ex}) at 3 s^{-1} between the two states at 59 °C and the $\Delta\omega$ for protons is >36 s^{-1} (>0.1 ppm in a 360 MHz magnet) [36]. Hence, two distinct sets of peaks can be identified for the same nuclei in 1D or 2D NMR experiments (Fig. 1). The peak intensities or volumes (for 2D spectra) provide a direct measure of the relative populations of the two states and the equilibrium constant (K_{U}) for unfolding.

$$K_{\text{U}} = p_{\text{U}}/p_{\text{F}} \text{ and } p_{\text{U}} + p_{\text{F}} = 1 \tag{1}$$

where p_{U} and p_{F} are the populations of the unfolded and folded states, respectively. The free energy of unfolding (ΔG_{U}) is given by

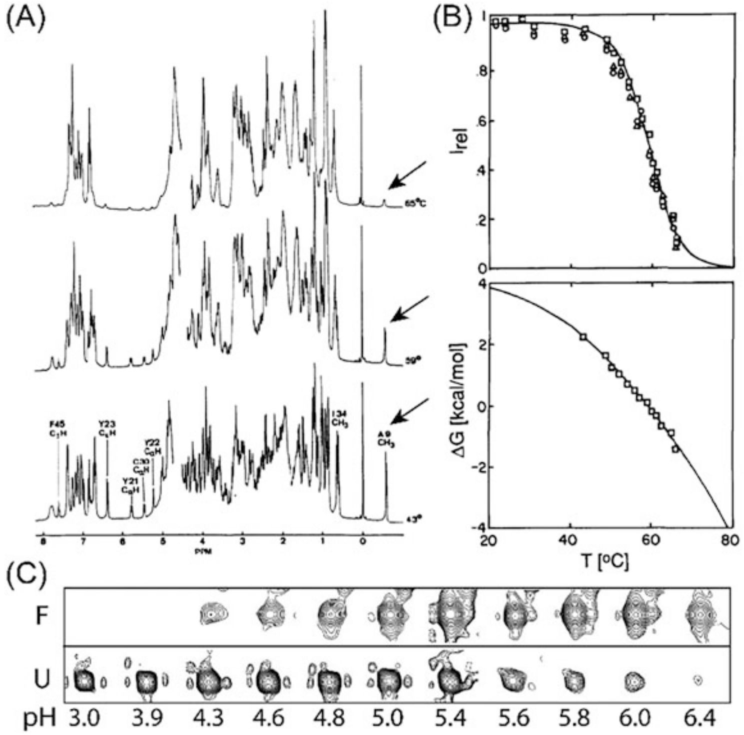

Fig. 1 Folding-unfolding at equilibrium can be monitored by NMR spectroscopy. (**a**) Thermal denaturation of HPI, a trypsin inhibitor, is monitored by 1D NMR [37]. The methyl peak of A9 (indicated by the arrow) corresponds to the folded state. Its intensity decreases with an increase in temperature as the folded population decreases. (**b**) The folded and unfolded populations at each temperature were determined from the assigned peaks of several residues, which resulted in a melting temperature (T_m) of 59 °C of HPI (upper plot). The dependence of Gibb's free energy on temperature allowed the determination of enthalpy and heat capacity of unfolding (lower plot). (**c**) The pH-dependent unfolding of TrAvrPto, an effector protein of the plant pathogen *Pseudomonas syringae*, was monitored by 2D ^{15}N-^{1}H HSQC experiments [35]. F and U represent the folded and unfolded peaks of G95 residue, respectively

$$\Delta G_{\mathrm{U}} = -RT \ln K_{\mathrm{U}}. \tag{2}$$

where R is the universal gas constant, and T is the absolute temperature at which the measurements are done.

An advantage of these NMR measurements is that for most nuclei, the baseline chemical shifts of the folded and unfolded states are independent of the parameter (pH, temperature, or denaturants) used to unfold the protein. Thus, the changing baseline values need not be extrapolated to the folding–unfolding transition zone, which is a major source of systematic error in other methods.

2.1 Folding and Unfolding Studies by 1D NMR

The power of proton 1D NMR for protein folding–unfolding studies can be illustrated by the elegant studies on HPI, a trypsin inhibitor derived from *Helix pomatia*. Thermal unfolding of HPI showed distinct sets of peaks for denatured and native states at equilibrium [36]. This corresponds to the slow exchange on the NMR timescale between the folded and unfolded states. ^{1}H NMR spectrum of HPI was obtained at different temperatures. The spectrum obtained at 43 °C coincided well with the well-resolved assigned spectrum for the native protein in the folded conformation [38]. The spectrum obtained at 65 °C, when the protein was 90% unfolded, resembled a random coil structure. The spectrum at 59 °C, which marked the midpoint of the thermal unfolding transition, showed an overlap between the spectra obtained at 43 °C and 65 °C. This conclusively showed a two-state transition of HPI between the folded and unfolded states with no partially folded intermediate. The magnetization transfer method used to measure the folding and unfolding rate of 3 s^{-1} proved that the conversion rate between the two states was small compared to the difference in frequency of the resonances corresponding to the two states. Transition curves were obtained from ^{1}H resonances of Ala-9 CH_3, Tyr-21 $C_{\alpha}H$, and Tyr-23 $C_{\varepsilon}H$ of HPI and all three curves were found to be identical. Similar results were obtained for BPTI, which has a 50% sequence homology to HPI, and also with a destabilized derivative of BPTI, known as RCOM-BPTI, which showed minor deviation from the two-state transition seen in HPI [39]. The populations of molecules with folded and unfolded conformations were measured from the normalized intensity of the resolved peaks in the spectrum. The thermodynamic parameters were determined using Eqs. (1) and (2).

Folding–unfolding of proteins as a function of the concentration of chemical denaturant has also been studied by NMR spectroscopy. Refolding of unfolded apoplastocyanin was triggered by manual dilution of the denaturant guanidine hydrochloride and followed by a series of 1D NMR spectra. The folding takes place in several hours due to the slow trans-to-cis isomerization of two proline residues. Measurement of the kinetics of the folding reaction by 1H NMR identified a folding intermediate, which had one of the prolines in the incorrect trans configuration [40]. A similar 1D NMR experiment on the unfolding of ribonuclease A also revealed the presence of a folding intermediate. The intermediate was characterized as a ‘dry molten globule’ that had side chains free to rotate, but the hydrophobic core was still devoid of water molecules [41].

2.2 Folding and Unfolding Studies by 2D NMR

Two-dimensional NMR spectroscopy provides higher spectral resolution and more structural information about the folded and unfolded states. 2D NMR experiments were used to investigate the pH-dependent folding–unfolding of AvrPto, a

Pseudomonas syringae effector protein [35]. A set of two distinct peaks for each residue was observed in the 2D ^{15}N–^{1}H HSQC spectrum of TrAvrPto, in which the disordered N and C-terminal tails of AvrPto are removed. These two discrete populations of peaks correspond to the folded and unfolded states in slow exchange. The well-resolved peaks of the folded and unfolded states of 13 backbone amides distributed throughout the protein were used to show the pH dependence of the folded and unfolded populations [35]. The protein denatured under acidic conditions as the pH was lowered from 7 to 4. Determination of the sidechain pK_a of all histidines by NMR revealed anomalously low pKa of His87. Solvent-exposed histidine sidechain has a pK_a of 6.2 [42]; His87 in TrAvrPto has a pK_a of 4.8, as it is buried in the core of the protein. His87 was shown to be a pH-sensitive folding switch that facilitates the transport of AvrPto through the narrow type III secretion system [35].

2.3 Measurement of Residue-Wise Stability by Hydrogen Exchange (HX) Experiments

One shortcoming of the protein folding–unfolding studies at equilibrium is the limited range over which these studies can be performed. NMR-based hydrogen exchange (HX) experiments allow the estimation of unfolding equilibrium parameters under conditions far removed from the folding–unfolding transition zone.

Labile hydrogens (NH or OH) in proteins readily exchange with the bulk water. This exchange is catalysed by acid or base, and the exchange rate for amides is minimum at pH 3 [43]. The exchange rate is significantly slowed for hydrogens protected by hydrogen bonds or hydrogens rendered inaccessible to the bulk water due to the folded structure of the protein. The model for HX is given as follows:

$$F(H) \underset{k_U}{\overset{k_F}{\leftrightharpoons}} U(H) \underset{D_2O}{\overset{k_C}{\rightarrow}} U(D) \tag{3}$$

where the folded (F) and unfolded (U) states exchange with rates of k_U and k_F, and the protons (H) in the unfolded state exchange with the bulk D_2O with a rate k_C. The exchange rate, k_C, depends upon a variety of conditions (pH, temperature, neighbouring amino acid side chains, and isotope effects), which have been calibrated in several unfolded models [43–45]. Under steady-state conditions, the overall exchange rate, k_{EX}, is given by

$$k_{EX} = (k_U k_C)/(k_U + k_F + k_C) \tag{4}$$

In the case of stable structures $k_U \ll k_F$

$$k_{\mathrm{EX}} = (k_{\mathrm{U}} k_{\mathrm{C}})/(k_{\mathrm{F}} + k_{\mathrm{C}}) \tag{5}$$

Depending on the ratio of the refolding rate (k_{F}) to the chemical exchange rate (k_{C}), two limiting situations exist, i.e. bimolecular exchange EX2 and unimolecular exchange EX1. Under the EX2 limit, where folding is faster than the chemical exchange, i.e. $k_{\mathrm{F}} \gg k_{\mathrm{C}}$

$$k_{\mathrm{EX2}} = (k_{\mathrm{U}} k_{\mathrm{C}})/k_{\mathrm{F}} = K_{\mathrm{U}} k_{\mathrm{C}}, \qquad K_{\mathrm{U}} = k_{\mathrm{U}}/k_{\mathrm{F}} \tag{6}$$

where K_{U} is the equilibrium constant for the rate-determining structural opening reaction [Eq. (3). The HX protection factor (P) is given by $P = k_{\mathrm{C}}/k_{\mathrm{EX2}} = 1/K_{\mathrm{U}}$. The protection factor (P) provides residue-wise thermodynamic stability of a protein and insights into the local fluctuations within a protein [23, 24, 46, 47].

Under the EX1 limit, where the chemical exchange is faster than refolding ($k_{\mathrm{F}} < k_{\mathrm{C}}$)

$$k_{\mathrm{EX1}} = k_{\mathrm{U}} \tag{7}$$

The measured exchange rate k_{EX1} directly gives the rate of unfolding. While EX2 exchange provides accurate results regarding the thermodynamics of protein stability, EX1 exchange follows the kinetics of protein unfolding and folding. Exchanges in the EX2 limit usually occur under native conditions, whereas exchanges in the EX1 limit are typically observed under unfolding conditions or at extremes of temperature or pH. However, small increases in temperature, pH, or mutations can induce a change in mechanism [48, 49].

2.4 Equilibrium HX Experiments

Decades of protein folding–unfolding studies have firmly established that (1) exchange of the core protons in a stable protein requires major unfolding, comparable to the conformational changes associated with denaturation; (2) the structural unfolding model [Eq. (3)] provides a firm basis for the quantitation of free energy changes and protein stability, and (3) exchange rates measured in the fully unfolded protein or derived from model peptide data are good approximations of the actual chemical exchange rates (k_{C}) in transiently unfolded states.

Hydrogen exchange experiments on various trypsin inhibitors have shown that the exchange rates for the slowly exchanging amide protons in the core of the protein were correlated with thermal stability. Thus, the global folding–unfolding events are responsible for the exchange of these protons [50–52]. Interestingly, the roles of loops in protein folding have been highlighted by hydrogen exchange experiments [53]. In cytochrome c, an omega loop (residues 40–57) acted as a cooperative unfolding/refolding unit under native conditions [54].

2.5 Relaxation Dispersion Experiments

Proteins in their native state sample higher energy conformations, which have low populations according to Boltzmann distribution ($p_U/p_F = \exp.(-\Delta G/RT)$, where p_U and p_F are the populations of the higher energy and native states, ΔG is the free energy difference between the two states, R is the gas constant, and T is the absolute temperature). This high energy conformation results in different chemical shifts compared to the native state for several nuclei on several residues of the protein. The relaxation dispersion (RD) experiment is sensitive to the rate of exchange as well as the chemical shift difference between these two states, i.e. the lower energy native state and the higher energy state [55]. This experiment can detect populations as low as 0.5%. RD measurements have been used to identify on-pathway folding intermediates, characterize the partial unfolding of protein segments, and the process of folding upon binding of intrinsically disordered proteins. RD experiment provides detailed information on the kinetics and thermodynamics of an exchange process and, in favourable conditions, allows the structure determination of the high energy conformation [56].

Folding of the FF domain, derived from the human protein HYPA/FBP11, was shown to have an on-pathway intermediate state. RD NMR studies enabled the structural characterization of this intermediate state and showed that in this state, three out of four helices were partially formed, and the fourth helix was disordered [57]. RD experiments have been successfully used to characterize the spontaneous folding and unfolding of a helix appended to the DNA-binding homeodomain of PBX [58]. Using RD experiments on multiple nuclei ($^1H^N$, ^{13}C, and ^{15}N), the folding upon binding of the disordered domain of Sendai virus nucleoprotein (NT) was characterized. It folds into a helix upon binding to the C-terminal domain of the phosphoprotein (PX). It was shown that NT samples several helical sub-states, which form encounter complex with PX and finally bind a helical grove on PX, resulting in a stable complex [59].

3 Studies of Protein Folding–Unfolding Kinetics

Detection of folding or unfolding events directly by NMR spectroscopy is difficult for most proteins. However, proteins with sufficiently slow folding have been studied by NMR. The experimental methods for direct detection vary from simple manual mixing to temperature-jump and stopped-flow NMR. Several devices have been designed to monitor protein folding in real time by NMR [60, 61]. Typically, 50 μL of concentrated protein in denaturing buffer is injected into 450 μL of refolding buffer, which is present in the NMR tube inside the magnet (Fig. 2). Several proteins have been studied by this method, such as α-lactalbumin [62, 63], RNase T1 mutant (S54G/P55N) [64], and amyloidogenic protein β2-microglobulin (B2M) [65].

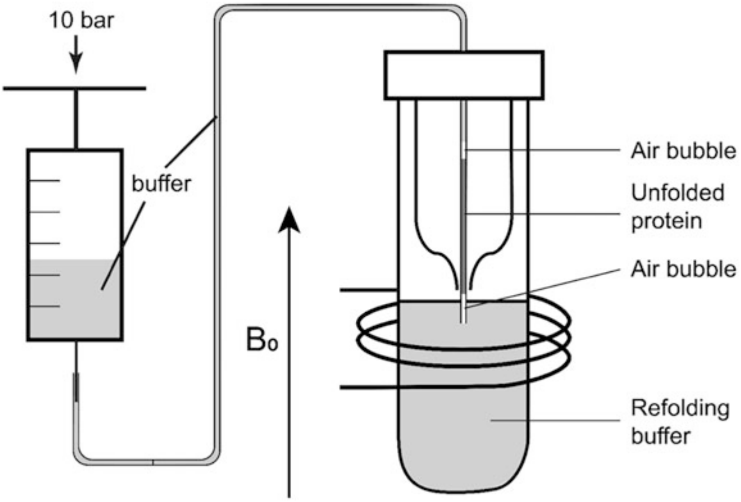

Fig. 2 Rapid mixing device used for the study of protein folding or unfolding by real-time NMR experiments. The NMR tube containing the refolding buffer is inserted into the probe of the spectrometer. A small air bubble separates the unfolded protein present in the transfer line. Another air bubble separates the protein from the injection buffer in the remainder of the transfer line. Protein injection into the NMR tube is triggered by a piston outside the magnet. Figure adapted from [61]

3.1 Protein Folding Studies by Fast 2D NMR Experiments

One-dimensional NMR (1D NMR) has been instrumental in studying slow conformational changes and kinetics of protein folding and unfolding [66]. These processes are studied by recording the 1D NMR spectra after inducing the folding or unfolding reaction, often through the addition of denaturants such as urea or guanidine hydrochloride. Despite it being a very fast and sensitive technique, its major drawback is the low spectral resolution seen for biological macromolecules. The lack of dispersion in 1D spectra of denatured proteins results in its spectrum resembling that of mixtures of free amino acids making sequence-specific resonance assignments challenging. The limited spectral dispersion and resonance line width information only allow the confirmation of the unstructured or globular conformation of a protein. Multidimensional NMR (2D, 3D, or more) helps in investigating the local structural and dynamic processes of macromolecules. Homonuclear 2D experiments such as correlation spectroscopy (COSY) [67] allow for resolving the 1D ^{1}H spectrum in a 2D plane and thereby help reduce the 1D signal overlap. One major advancement in the characterization of unfolded protein states came from the use of uniform isotope labelling using ^{13}C and ^{15}N. The application of multidimensional heteronuclear experiments allowed better discrimination of resonances. Commonly used heteronuclear 2D NMR experiments include single quantum correlation (HSQC) [68], multiple quantum correlation (HMQC) [69], and multiple bond correlation (HMBC) [70]. Despite its high potential, the use of 2D

NMR experiments has limitations and suffers from intrinsic drawbacks. A 2D spectrum is collected as a series of 1D spectra where the second dimension, often the chemical shift corresponding to the ^{13}C or ^{15}N nuclei, is recorded indirectly by progressively increasing the time during t1 evolution. This results in a longer duration of the 2D experiments, typically several minutes. Since folding–unfolding reactions happen in a faster timescale (milliseconds to seconds), these standard 2D NMR experiments are not very helpful in studying the kinetics of the process. However, the endpoints, i.e. the completely folded state under native conditions and the completely unfolded state under denaturing conditions, can be studied in great detail.

Several approaches have been developed to shorten the duration of 2D NMR experiments, such as reducing the interscan delay in the band-selective optimized-flip-angle short-transient (SOFAST) spectroscopy [71], along with the similar band-selective excitation short-transient (BEST) spectroscopy [72], acceleration by sharing adjacent polarization (ASAP) [73], and small recovery times (SMART) [74]. Some approaches reduce experiment time by collecting sparse data, such as nonuniform sampling (NUS) [75], Hadamard, and projection reconstruction sampling [75]. An alternative approach named ultrafast (UF) 2D NMR uses multiplexing instead of sequential sampling in the indirect dimension by spatial encoding [76].

In BEST and SOFAST experiments, only the protons of interest are excited by band-selective radio-frequency (RF) pulses. This results in dipolar interactions of the excited protons with a large number of unexcited protons and enhances the longitudinal relaxation rate, thereby significantly reducing the delay time between scans of an NMR experiment. In SOFAST-HMQC experiments, the ^{1}H steady-state polarization is further enhanced by Ernst-angle excitation. A SOFAST-HMQC spectrum can be recorded in a few seconds. Sparse nonuniform data sampling can further reduce this time. The ultra SOFAST technique, based on the gradient-assisted spatial encoding of the NMR frequencies, can record 10 spectra per second.

3.2 Protein Folding by Real-Time NMR Spectroscopy

The 2D experiment ^{1}H-^{15}N SOFAST-HMQC has been used to observe the real-time folding of α-lactalbumin from its molten globular state to its native state [77]. Under acidic conditions (pH ~ 2), α-lactalbumin forms a molten globular state and, at neutral pH, is properly folded. The folding in the NMR tube, from the molten globular state to the native state, was triggered by a pH jump using fast mixing (Fig. 2) and monitored by collecting SOFAST spectra. It followed the first-order kinetics with a folding rate of 10^{-2} s^{-1} [77]. A 3D experiment named BEST-HSQC-HNCA has been used to study the real-time folding of RNase T1 mutant (S54G/P55N), an 11 kDa protein [64]. Refolding was triggered by the fast mixing of denatured protein in 6 M guanidine hydrochloride into excess refolding buffer in the NMR tube. This protein has two proline residues in the *cis* configuration in the folded state. The trans-to-cis conversion is a slow process, resulting in the slow folding of this protein. Previous studies had reported a folding intermediate state with one of the prolines trapped in the trans configuration [78]. The 3D BEST-HSQC-HNCA experiment enabled the backbone assignment of the transient intermedia state and provided structural details of this state [64].

Recently, it has been demonstrated that the two powerful methods of NMR spectroscopy, i.e. relaxation dispersion (which detects high energy low population states) and real-time NMR (which allows kinetic measurements), can be combined to study the conformational exchange dynamics of the short-lived excited protein states that are transiently formed during protein folding [65]. BEST-TROSY CPMG relaxation dispersion experiment was used to measure the conformational exchange dynamics of the major folding intermediate of the amyloidogenic protein β2-microglobulin (B2M). This study showed that the transient intermediate also forms a dimer similar to the folded native state of the protein; however, the dimer has a higher population and is formed at a faster rate for the intermediate [65].

3.3 Determination of Folding Pathways by HX Labelling Experiments

For relatively faster folding proteins, the stopped-flow method of labelling labile protons followed by NMR has been used [79–81] (Fig. 3). Initially, the protein is dissolved in D_2O in the presence of a suitable denaturing agent (such as guanidine hydrochloride) and kept in a syringe (S1). The amide protons (NH) in this completely unfolded protein are exchanged with the solvent deuterium (ND). A rapid mixing apparatus is employed where the solutions are passed by applying stopped-flow, and the flow rate can be controlled manually [36]. Another syringe (S2) contains a refolding buffer. Refolding is initiated when the solutions of the two syringes are mixed together in a mixer (M1) and flowed at a rate for a certain amount of time, termed refolding time (t_f). After some refolding time ($t_f \sim 50$ ms), a proton pulse (t_p) is applied by mixing with a high pH buffer in H_2O kept in a syringe (S3). Solvent-exposed amides are exchanged to NH, while the amides, which are part of the already formed secondary structure, are protected from the exchange and remain deuterated (ND). Finally, the exchange is quenched by adding the protein to a low-pH buffer in H_2O (Fig. 3). The low-pH refolding buffer not only terminates any further exchange of protons but also allows complete refolding of the protein. The native protein is studied by 1D and 2D NMR experiments. Separate spectra are collected by changing the refolding time ($t_f \sim 100$, 150, 200 ms, and so on). After collecting several spectra, a snapshot of the refolding pattern can be obtained where the time sequence of various secondary structure formations can be determined. The amount of protected amide proton in refolded native protein can be analysed by measuring the intensities of resolved NH resonances from ^{1}H NMR or 2D NMR. The relative proton occupancies, P, at each site can be calculated by normalizing the measured signal intensities, I_m, as follows:

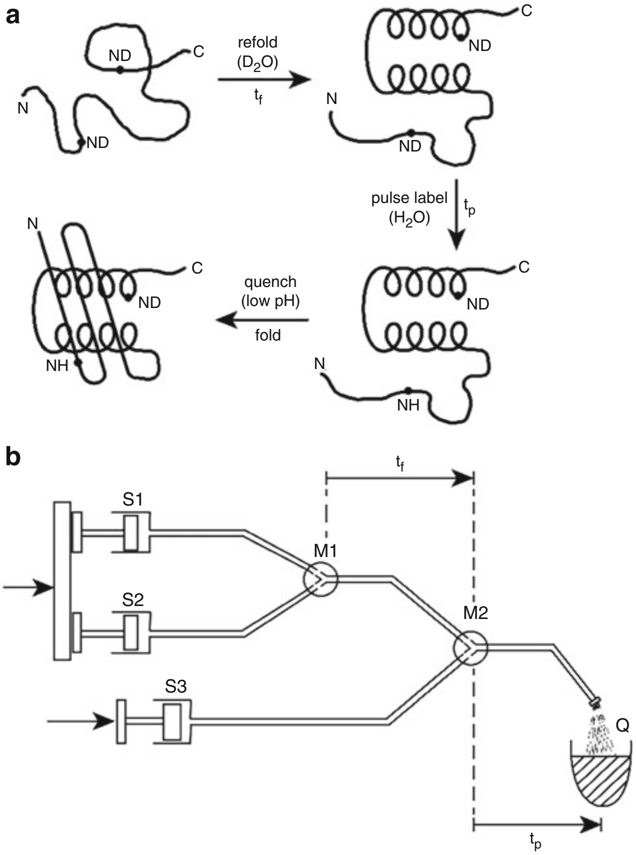

Fig. 3 Stopped-flow analysis of protein folding. (**a**) The protein is unfolded in a deuterated (D_2O) denaturing buffer. Refolding is triggered by mixing with deuterated refolding buffer for a time period of t_f. The partially folded protein is added to a protonated buffer (H_2O) to exchange the unprotected amide ND with the bulk solvent (H_2O) to form NH. The protium-deuterium exchange is quenched by adding the protein to a low-pH buffer. This sample is collected and analysed by 1D or 2D NMR experiments. (**b**) A schematic diagram of the stopped-flow device is shown. Syringe S1 contains the unfolded protein, and S2 contains the refolding buffer. They are simultaneously injected and mixed in the mixture M1 for a time t_f. Syringe S3 contains a protonated buffer which is mixed with the partially folded protein to pulse label the exposed amides in the mixture M2. The exchange reaction is quenched by putting the protein in the low-pH solution (Q). Figure adapted from [36, 82]

$$P = [(I_m/I_0) - f_h]/(1 - f_h) \tag{8}$$

where I_0 is the signal intensity of the fully protonated group, and f_h is the residual fraction of water present in the reaction mixture. A change in the pH or t_p provides information about the stability of the protein when the exchange follows the EX2 mechanism [83]. It can also calculate the k_U and k_F when the exchange follows the EX1 mechanism [54, 84].

4 Monitoring Protein Folding in Live Cells

Cells have mechanisms to prevent misfolding of proteins. One such mechanism is through the aid of molecular chaperones along with various cofactors that assist in protein folding. These molecular chaperones and cofactors collaborate with the protein degradation machinery to maintain protein homeostasis within the cells. In vitro studies cannot recapitulate this aspect of in vivo protein folding [85]. Also, the dense cellular environment can impact protein stability and the folding process by changing the energy landscape of protein folding [86, 87]. Understanding protein folding in the cellular environment is still a challenging problem. NMR spectroscopy has contributed significantly to the studies of in vivo protein folding.

The low abundance of NMR-active nuclei N^{15} (0.4%) and C^{13} (1.1%) is exploited for in vivo studies of protein folding. Proteins with N^{15} and C^{13} labels are routinely expressed and purified from bacterial systems for NMR studies [21, 23, 46]. The labelled protein is incorporated into target cells by microinjection [88, 89], attachment of cell-penetrating peptide [90], diffusion through pore-forming toxins [91], or electroporation [92]. Since the naturally expressed proteins within the cells have mostly the NMR inactive N^{14} and C^{12} isotopes, they are not observed by NMR spectroscopy. Using the ^{15}N and ^{13}C labelled in vitro samples, the HSQC, HMQC, and CON fingerprint spectra are assigned [34].

Despite its advantages, in-cell NMR suffers from various drawbacks. The first is the line broadening of NMR resonances which is mostly observed in the soluble globular proteins compared to disordered proteins [93]. In comparison to protein folding studied in buffers, the tumbling of a protein is seen to be slower in the cytoplasm, more so in globular proteins, which makes them more difficult to be observed by in-cell NMR. The local internal motions of the disordered proteins are independent of their global motions and are not affected to the same extent as the global motions of globular proteins. The hindered rotational diffusion of globular proteins gives rise to low-resolution NMR spectra [94]. The second limitation arises due to the low sensitivity and longer experimental times of higher-dimensional heteronuclear NMR experiments required to derive long-range distance restraints considering the limited lifetimes of in-cell NMR samples. In order to counter these limitations, various advances have been made in NMR methods. Fast acquisition routines such as the fast pulsing method (SOFAST) [95] shorten the duration of each scan and decrease the interscan delay to allow the acquisition of more scans and reduce the time required for the acquisition of multidimensional NMR experiments. Nonuniform sampling procedures reduce the number of scans and shorten the time required for the data acquisition of multidimensional NMR experiments [96, 97].

Using in-cell NMR, the folding and maturation events of a homodimeric human Cu,Zn-SOD1 metalloprotein were investigated [98]. This protein is involved in defending the cells against oxidative stress. hSOD1 attains its mature form by the incorporation of one Zn^{2+} and one Cu^{2+} ion per subunit. It also forms an intramolecular disulphide bridge through two conserved cysteine residues. The folding and formation of intermediate maturation states of hSOD1 were characterized from cell samples in minimal media, overexpressing the protein by varying the amounts of metal cofactors. A monomeric unfolded apo form was also detected. Apo-hSOD1 is an immature form of hSOD1 with a misfolded structure. Apo-hSOD1 is implicated in ALS pathology. ^{1}H,^{15}N-SOFAST-HMQC spectra were recorded first on cell samples that overexpressed hSOD1 in a metal-free medium, following which a second spectrum was obtained from the cell lysates after cell lysis. Figure 4a, b shows the ^{1}H,^{15}N-SOFAST-HMQC spectra of apo-hSOD1 and cell lysate, respectively. Both spectra show the presence of unfolded regions, with most peaks occurring within the 8.0–8.3 ppm ^{1}H region with few dispersed peaks in Fig. 4a, indicating some structured regions in the protein. Figure 4b also shows a few other dispersed peaks, and this spectrum compares well with the monomeric apo form with reduced cysteine E,E-hSOD1SH-SH, which is seen through the overlay of the spectrum from the cell lysate of hSOD1 without the addition of Zn^{2+} and that of the ^{1}H-^{15}N HSQC of an in vitro sample of E,E-hSOD1SH-SH. This indicated that the newly formed protein hSOD1 remained in a metal-free state in the cytoplasm in the absence of metal ions. To identify if the apo-protein remains completely unfolded in the cytoplasm, the in-cell NMR spectra were compared to the in vitro 2D ^{1}H-^{15}N HSQC NMR spectra of E,E-hSOD1SH-SH denatured with increasing amounts of guanidinium chloride up to 0.5 M. With the exception of the signals typical of unfolded regions of hSOD1, the latter spectrum differed from the in-cell spectrum to some extent. This indicated that the apo-hSOD1 is not completely unfolded in the cellular environment. In-cell NMR also helped to identify the Zn-bound monomer and dimer as well as the Cu-Zn form of hSOD1.

5 Conclusions

Over the years, NMR spectroscopy has evolved to incorporate increasingly sophisticated experiments to enhance our understanding of protein folding. Almost all aspects of protein folding, such as thermodynamics, kinetics, formation of intermediates as well as in-cell folding, have been studied by NMR spectroscopy. This technique also has the added advantage of providing information at an atomic resolution on the folding intermediates. The elegant coupling of the stopped-flow technique with NMR spectroscopy and the development of fast NMR methods have made NMR spectroscopy, a truly unique technique for studying and understanding protein folding.

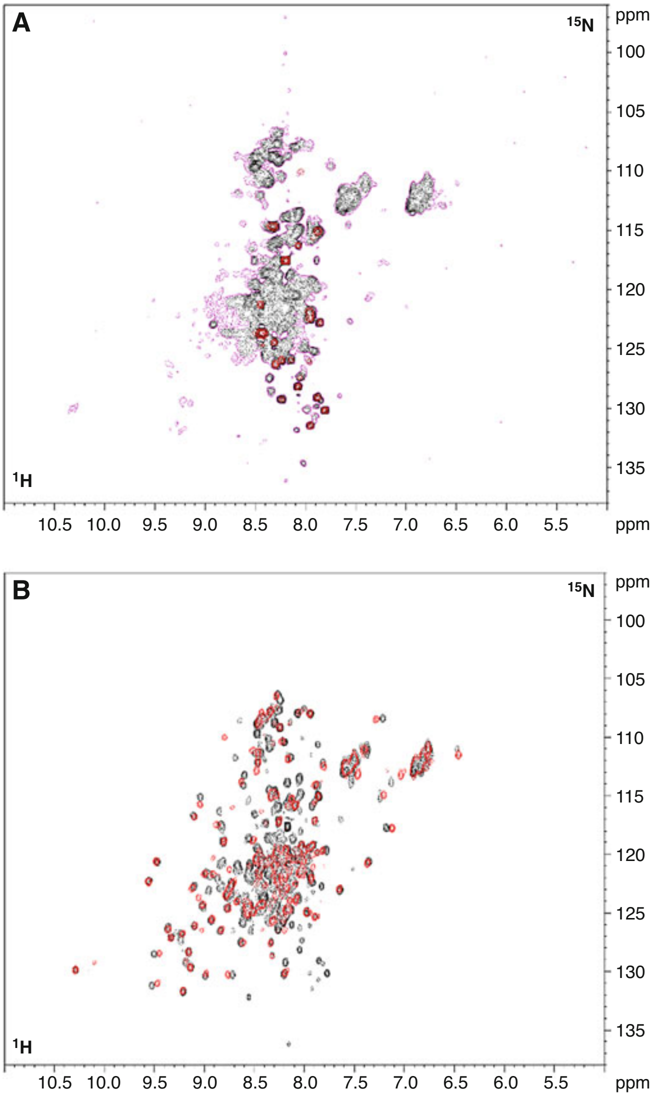

Fig. 4 In-cell NMR of protein folding. (**a**) ^{1}H-^{15}N SOFAST-HSQC spectrum is shown for hSOD1 overexpressed in *E. coli* cells in a metal-free medium [98]. Peaks (red) are visible within the 8.0 to 8.3 ^{1}H ppm range, indicating mostly unfolded protein. The lower contour level (black) shows additional broad peaks. (**b**) ^{1}H-^{15}N SOFAST-HSQC spectrum of the lysed cells in anaerobic condition without the addition of Zn(II) (black) is overlaid with the spectrum of an in vitro sample of E,E-hSOD1SH-SH (red)

Acknowledgements This work has been supported by funding from the Science and Engineering Research Board (SERB), India (CRG/2020/004984). NSV acknowledges a NET-JRF fellowship (Joint CSIR-UGC NET JUNE 2016). S.B. acknowledges a NET-JRF fellowship [918/(CSIR-UGC NET DEC. 2016)] provided by the University Grants Commission (UGC), Government of India.

References

1. K.A. Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, The protein folding problem. Annu. Rev. Biophys. **37**, 289–316 (2008). https://doi.org/10.1146/annurev.biophys.37.092707.153558
2. R. Zwanzig, A. Szabo, B. Bagchi, Levinthal's paradox. Proc. Natl. Acad. Sci. U. S. A. **89**, 20–22 (1992). https://doi.org/10.1073/pnas.89.1.20
3. J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins **21**, 167–195 (1995). https://doi.org/10.1002/prot.340210302
4. J.N. Onuchic, Z. Luthey-Schulten, P.G. Wolynes, Theory of protein folding: the energy landscape perspective. Annu. Rev. Phys. Chem. **48**, 545–600 (1997). https://doi.org/10.1146/annurev.physchem.48.1.545
5. P.E. Leopold, M. Montal, J.N. Onuchic, Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc. Natl. Acad. Sci. U. S. A. **89**, 8721–8725 (1992). https://doi.org/10.1073/pnas.89.18.8721
6. K.A. Dill, J.L. Maccallum, P. Folding, The protein-folding problem, 50 years on. Science **338**, 1042–1047 (2012)
7. M.S. Hipp, S.H. Park, U.U. Hartl, Proteostasis impairment in protein-misfolding and -aggregation diseases. Trends Cell Biol. **24**, 506–514 (2014). https://doi.org/10.1016/j.tcb.2014.05.003
8. F. Chiti, C.M. Dobson, Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. **75**, 333–366 (2006). https://doi.org/10.1146/annurev.biochem.75.101304.123901
9. M. Brehme, C. Voisine, T. Rolland, S. Wachi, J.H. Soper, Y. Zhu, K. Orton, A. Villella, D. Garza, M. Vidal, et al., A chaperome subnetwork safeguards proteostasis in aging and neurodegenerative disease. Cell Rep. **9**, 1135–1150 (2014). https://doi.org/10.1016/j.celrep.2014.09.042
10. Y.J. Choe, S.H. Park, T. Hassemer, R. Körner, L. Vincenz-Donnelly, M. Hayer-Hartl, F.U. Hartl, Failure of RQC machinery causes protein aggregation and proteotoxic stress. Nature **531**, 191–195 (2016). https://doi.org/10.1038/nature16973
11. S.H. Park, Y. Kukushkin, R. Gupta, T. Chen, A. Konagai, M.S. Hipp, M. Hayer-Hartl, F.U. Hartl, PolyQ proteins interfere with nuclear degradation of cytosolic proteins by sequestering the Sis1p chaperone. Cell **154**, 134–145 (2013). https://doi.org/10.1016/j.cell.2013.06.003
12. B.H. Qu, E.H. Strickland, P.J. Thomas, Localization and suppression of a kinetic defect in cystic fibrosis transmembrane conductance regulator folding. J. Biol. Chem. **272**, 15739–15744 (1997). https://doi.org/10.1074/jbc.272.25.15739
13. T. Hidvegi, B.Z. Schmidt, P. Hale, D.H. Perlmutter, Accumulation of mutant α1-antitrypsin Z in the endoplasmic reticulum activities caspases-4 and -12, NFκB, and BAP31 but not the unfolded protein response. J. Biol. Chem. **280**, 39002–39015 (2005). https://doi.org/10.1074/jbc.M508652200
14. D.A. Lomas, D. Li-Evans, J.T. Finch, R.W. Carrell, The mechanism of Z α1-antitrypsin accumulation in the liver. Nature **357**, 605–607 (1992). https://doi.org/10.1038/357605a0
15. B. Caughey, P.T. Lansbury, Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. Annu. Rev. Neurosci. **26**, 267–298 (2003). https://doi.org/10.1146/annurev.neuro.26.010302.081142

16. A. Fisher, I. Bezprozvanny, L. Wu, D.A. Ryskamp, N. Bar-Ner, N. Natan, R. Brandeis, H. Elkon, V. Nahum, E. Gershonov, et al., AF710B, a novel M1/σ1 agonist with therapeutic efficacy in animal models of Alzheimer's disease. Neurodegener Dis **16**, 95–110 (2016a). https://doi.org/10.1159/000440864
17. C.L. Fisher, R.J. Resnick, S. De, L.A. Acevedo, K.P. Lu, F.C. Schroeder, L.K. Nicholson, Cyclic cis-locked phospho-dipeptides reduce entry of AβPP into amyloidogenic processing pathway. J. Alzheimers Dis. **55**, 391–410 (2016b). https://doi.org/10.3233/JAD-160051
18. J. Labbadia, R.I. Morimoto, The biology of proteostasis in aging and disease. Annu. Rev. Biochem. **84**, 435–464 (2015). https://doi.org/10.1146/annurev-biochem-060614-033955
19. A.R. Fersht, V. Daggett, Review protein folding and unfolding at atomic resolution. Cell **108**, 573–582 (2002)
20. M.S. Lee, B. Cao, Nuclear magnetic resonance chemical shift: comparison of estimated secondary structures in peptides by nuclear magnetic resonance and circular dichroism. Protein Eng. **9**, 15–25 (1996). https://doi.org/10.1093/protein/9.1.15
21. A.J. Basak, S. Maiti, A. Hansda, D. Mahata, K. Duraivelan, S.V. Kundapura, W. Lee, G. Mukherjee, S. De, D. Samanta, Structural insights into N-terminal IgV domain of BTNL2, a T cell inhibitory molecule, suggests a non-canonical binding Interface for its putative receptors. J. Mol. Biol. **432**, 5938–5950 (2020). https://doi.org/10.1016/j.jmb.2020.09.013
22. J.R.C. Bergeron, L.J. Worrall, S. De, N.G. Sgourakis, A.H. Cheung, E. Lameignere, M. Okon, G.A. Wasney, D. Baker, L.P. McIntosh, et al., The modular structure of the inner-membrane ring component PrgK facilitates assembly of the type III secretion system basal body. Structure **23**, 161–172 (2015). https://doi.org/10.1016/j.str.2014.10.021
23. S. Boral, S. Maiti, A.J. Basak, W. Lee, S. De, Structural, dynamic, and functional characterization of a DnaX mini-intein derived from Spirulina platensis provides important insights into Intein-mediated catalysis of protein splicing. Biochemistry **59**, 4711–4724 (2020). https://doi.org/10.1021/acs.biochem.0c00828
24. H.J. Coyne, S. De, M. Okon, S.M. Green, N. Bhachech, B.J. Graves, L.P. McIntosh, Autoinhibition of ETV6 (TEL) DNA binding: appended helices sterically block the ETS domain. J. Mol. Biol. **421**, 67–84 (2012). https://doi.org/10.1016/j.jmb.2012.05.010
25. S. De, A.C.K. Chan, H.J. Coyne, N. Bhachech, U. Hermsdorf, M. Okon, M.E.P. Murphy, B.J. Graves, L.P. McIntosh, Steric mechanism of auto-inhibitory regulation of specific and non-specific dna binding by the ETS transcriptional repressor ETV6. J. Mol. Biol. **426**, 1390–1406 (2014). https://doi.org/10.1016/j.jmb.2013.11.031
26. S. De, A.I. Greenwood, M.J. Rogals, E.L. Kovrigin, K.P. Lu, L.K. Nicholson, Complete thermodynamic and kinetic characterization of the isomer-specific interaction between Pin1-WW domain and the amyloid precursor protein cytoplasmic tail phosphorylated at Thr668. Biochemistry **51**, 8583–8596 (2012). https://doi.org/10.1021/bi3008214
27. S. De, A.I. Greenwood, A.I. Acevado, N.E. Korson, L.K. Nicholson, Lineshape analysis as a tool for probing functional motions at biological interfaces, in *NMR Spectroscopy for Probing Functional Dynamics at Biological Interfaces*, ed. by A. Bhunia, H.S. Atreya, N. Sinha, (Royal Society of Chemistry, London, 2022), pp. 82–121
28. A.I. Greenwood, M.J. Rogals, S. De, K.P. Lu, E.L. Kovrigin, L.K. Nicholson, Complete determination of the Pin1 catalytic domain thermodynamic cycle by NMR lineshape analysis. J. Biomol. NMR **51**, 21–34 (2011). https://doi.org/10.1007/s10858-011-9538-9
29. W.F. Hawse, S. De, A.I. Greenwood, L.K. Nicholson, J. Zajicek, E.L. Kovrigin, D.M. Kranz, K.C. Garcia, B.M. Baker, TCR scanning of peptide/MHC through complementary matching of receptor and ligand molecular flexibility. J. Immunol. **192**, 2885–2891 (2014). https://doi.org/10.4049/jimmunol.1302953
30. A. Sekhar, L.E. Kay, An NMR view of protein dynamics in health and disease. Annu. Rev. Biophys. **48**, 297–319 (2019)
31. S. Roy, S. Boral, S. Maiti, T. Kushwaha, A.J. Basak, W. Lee, A. Basak, S.L. Gholap, K.K. Inampudi, S. De, Structural and dynamic studies of the human RNA binding protein

RBM3 reveals the molecular basis of its oligomerization and RNA recognition. FEBS J. **289**, 2847–2864 (2021). https://doi.org/10.1111/febs.16301

32. S. Maiti, S. De, Identification of potential short linear motifs (SLiMs) in intrinsically disordered sequences of proteins by fast time-scale backbone dynamics. J. Magn. Reson. Open **10–11**, 100029 (2022). https://doi.org/10.1016/j.jmro.2021.100029
33. S. Maiti, B. Acharya, V.S. Boorla, B. Manna, A. Ghosh, S. De, Dynamic studies on intrinsically disordered regions of two paralogous transcription factors reveal rigid segments with important biological functions. J. Mol. Biol. **431**, 1353–1369 (2019). https://doi.org/10.1016/j.jmb.2019.02.021
34. N.V. Saibo, S. Maiti, B. Acharya, S. De, Analysis of structure and dynamics of intrinsically disordered regions in proteins using solution NMR methods, in *Advances in Protein Molecular and Structural Biology Methods*, ed. by T. Tripathi, V.K. Dubey, (Academic Press, London, 2022), pp. 535–550. https://doi.org/10.1016/b978-0-323-90264-9.00032-5
35. J.E. Dawson, J. Šečkute, S. De, S.A. Schueler, A.B. Oswald, L.K. Nicholson, Elucidation of a pH-folding switch in the pseudomonas syringae effector protein AvrPto. Proc. Natl. Acad. Sci. U. S. A. **106**, 8543–8548 (2009). https://doi.org/10.1073/pnas.0809138106
36. H. Roder, G.A. Elöve, S.W. Englander, Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. Nature **335**, 700–704 (1988). https://doi.org/10.1038/335700a0
37. H. Roder, [22] structural characterization of protein folding intermediates by proton magnetic resonance and hydrogen exchange. Enzym. Dyn. **176**, 446–473 (1989)
38. K. Wüthrich, G. Wagner, Nuclear magnetic resonance of labile protons in the basic pancreatic trypsin inhibitor. J. Mol. Biol. **130**, 1–18 (1979). https://doi.org/10.1016/0022-2836(79)90548-5
39. K. Wüthrich, G. Wagner, R. Richarz, W. Braun, Correlations between internal mobility and stability of globular proteins. Biophys. J. **32**, 549–560 (1980). https://doi.org/10.1016/S0006-3495(80)84989-7
40. S. Koide, H.J. Dyson, P.E. Wright, Characterization of a folding intermediate of apoplastocyanin trapped by proline isomerization. Biochemistry **32**, 12299–12310 (1993). https://doi.org/10.1021/bi00097a005
41. T. Kiefhaber, A.M. Labhardt, R.L. Baldwin, Direct NMR evidence for an intermediate preceding the rate-limiting step in the unfolding of ribonuclease a. Nature **375**, 513–515 (1995). https://doi.org/10.1038/375513a0
42. G. Platzer, M. Okon, L.P. McIntosh, PH-dependent random coil 1H, 13C, and 15N chemical shifts of the ionizable amino acids: a guide for protein pK a measurements. J. Biomol. NMR **60**, 109–129 (2014). https://doi.org/10.1007/s10858-014-9862-y
43. R.S. Molday, S.W. Englander, R.G. Kallen, Primary structure effects on peptide group hydrogen exchange. Biochemistry **11**, 150–158 (1972)
44. Y. Bai, J.S. Milne, L. Mayne, S.W. Englander, Primary structure effects on peptide group hydrogen exchange. Proteins **17**, 75–86 (1993). https://doi.org/10.1002/prot.340170110
45. G.P. Connelly, Y. Bai, M.-F. Jeng, S.W. Englander, Isotope effects in peptide group hydrogen exchange. Proteins **17**, 87–92 (1993). https://doi.org/10.1002/prot.340170111
46. S. De, M. Okon, B.J. Graves, L.P. McIntosh, Autoinhibition of ETV6 DNA binding is established by the stability of its inhibitory helix. J. Mol. Biol. **428**, 1515–1530 (2016). https://doi.org/10.1016/j.jmb.2016.02.020
47. V.J. Hilser, E. Freire, Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. J. Mol. Biol. **262**, 756–772 (1996). https://doi.org/10.1006/jmbi.1996.0550
48. S.N. Loh, C.A. Rohl, T. Kiefhaber, R.L. Baldwin, A general two-process model describes the hydrogen exchange behavior of RNase A in unfolding conditions. Proc. Natl. Acad. Sci. U. S. A. **93**, 1982–1987 (1996). https://doi.org/10.1073/pnas.93.5.1982

49. S. Perrett, J. Clarke, A.M. Hounslow, A.R. Fersht, Relationship between equilibrium amide proton exchange behavior and the folding pathway of bamase. Biochemistry **34**, 9288–9298 (1995). https://doi.org/10.1021/bi00029a003
50. K.S. Kim, C. Woodward, Protein internal flexibility and global stability: effect of urea on hydrogen exchange rates of bovine pancreatic trypsin inhibitor. Biochemistry **32**, 9609–9613 (1993). https://doi.org/10.1021/bi00088a013
51. E. Moses, H.J. Hinz, Basic pancreatic trypsin inhibitor has unusual thermodynamic stability parameters. J. Mol. Biol. **170**, 765–776 (1983). https://doi.org/10.1016/S0022-2836(83)80130-2
52. H. Roder, G. Wagner, K. Wüthrich, Individual amide proton exchange rates in thermally unfolded basic pancreatic trypsin inhibitor. Biochemistry **24**, 7407–7411 (1985)
53. R. Li, C. Woodward, The hydrogen exchange core and protein folding. Protein Sci. **8**, 1571–1590 (1999). https://doi.org/10.1110/ps.8.8.1571
54. M.M.G. Krishna, Y. Lin, L. Mayne, S. Walter Englander, Intimate view of a kinetic protein folding intermediate: residue-resolved structure, interactions, stability, folding and unfolding rates, homogeneity. J. Mol. Biol. **334**, 501–513 (2003). https://doi.org/10.1016/j.jmb.2003.09.070
55. P. Neudecker, P. Lundström, L.E. Kay, Relaxation dispersion NMR spectroscopy as a tool for detailed studies of protein folding. Biophys. J. **96**, 2045–2054 (2009). https://doi.org/10.1016/j.bpj.2008.12.3907
56. G. Bouvignies, P. Vallurupalli, D.F. Hansen, B.E. Correia, O. Lange, A. Bah, R.M. Vernon, F.W. Dahlquist, D. Baker, L.E. Kay, Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. Nature **477**, 111–117 (2011). https://doi.org/10.1038/nature10349
57. D.M. Korzhnev, T.L. Religa, P. Lundström, A.R. Fersht, L.E. Kay, The folding pathway of an FF domain: characterization of an on-pathway intermediate state under folding conditions by 15N, 13Cα and 13C-methyl relaxation dispersion and 1H/2H-exchange NMR spectroscopy. J. Mol. Biol. **372**, 497–512 (2007). https://doi.org/10.1016/j.jmb.2007.06.012
58. P.J. Farber, J. Slager, A.K. Mittermaier, Local folding and misfolding in the PBX homeodomain from a three-state analysis of CPMG relaxation dispersion NMR data. J. Phys. Chem. B **116**, 10317–10329 (2012). https://doi.org/10.1021/jp306127m
59. R. Schneider, D. Maurin, G. Communie, J. Kragelj, D.F. Hansen, R.W.H. Ruigrok, M.R. Jensen, M. Blackledge, Visualizing the molecular recognition trajectory of an intrinsically disordered protein using multinuclear relaxation dispersion NMR. J. Am. Chem. Soc. **137**, 1220–1229 (2015). https://doi.org/10.1021/ja511066q
60. J. Grimaldi, J. Baldo, C. Mcmurray, B.D. Sykes, Stopped-flow nuclear magnetic resonance spectroscopy NMR. J. Am. Chem. Soc. **94**, 7641–7645 (1972)
61. M. Zeeb, J. Balbach, Protein folding studied by real-time NMR spectroscopy. Methods **34**, 65–74 (2004). https://doi.org/10.1016/j.ymeth.2004.03.014
62. J. Balbach, V. Forge, W.S. Lau, N.A.J. Van Nuland, K. Brew, C.M. Dobson, Protein folding monitored at individual residues during a two-dimensional NMR experiment. Adv. Sci. **274**, 1161–1163 (1996)
63. N.A.J. Van Nuland, C.M. Dobson, L. Regan, Characterization of folding the four-helix bundle protein Rop by real-time NMR. Protein Eng. Des. Sel. **21**, 165–170 (2008). https://doi.org/10.1093/protein/gzm081
64. C. Haupt, R. Patzschke, U. Weininger, S. Gröger, M. Kovermann, J. Balbach, Transient enzyme—substrate recognition monitored by real-time NMR. J. Am. Chem. Soc. **133**, 11154–11162 (2011). https://doi.org/10.1021/ja2010048
65. R. Franco, S. Gil-Caballero, I. Ayala, A. Favier, B. Brutscher, Probing conformational exchange dynamics in a short-lived protein folding intermediate by real-time relaxation–dispersion NMR. J. Am. Chem. Soc. **139**, 1065–1068 (2017). https://doi.org/10.1021/jacs.6b12089
66. N.A.J. Van Nuland, V. Forge, J. Balbach, C.M. Dobson, Real-time NMR studies of protein folding. Acc. Chem. Res. **31**, 773–780 (1998). https://doi.org/10.1021/ar970079l

67. A. Bax, R. Freeman, Investigation of complex networks of spin-spin coupling by two-dimensional NMR. J. Magn. Reson. **44**, 542–561 (1981). https://doi.org/10.1016/0022-2364(81)90287-0
68. G. Bodenhausen, D.J. Ruben, Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem. Phys. Lett. **69**, 185–189 (1980). https://doi.org/10.1016/0009-2614(80)80041-8
69. A. Bax, R.H. Griffey, B.L. Hawkins, Correlation of proton and nitrogen-15 chemical shifts by multiple quantum NMR. J. Magn. Reson. **55**, 301–315 (1983). https://doi.org/10.1016/0022-2364(83)90241-X
70. A. Bax, M.F. Summers, 1H and 13C assignments from sensitivity-enhanced detection of heteronuclear multiple-bond connectivity by 2D multiple quantum NMR. J. Am. Chem. Soc. **108**, 2093–2094 (1986). https://doi.org/10.1021/ja00268a061
71. B. Brutscher, P. Schanda, Rapid multidimensional NMR: fast-pulsing techniques and their applications to proteins, in *Encyclopedia of Magnetic Resonance*, (Wiley, New York, 2009), pp. 1–10. https://doi.org/10.1002/9780470034590.emrstm1154
72. P. Schanda, H. Van Melckebeke, B. Brutscher, Speeding up three-dimensional protein NMR experiments to a few minutes. J. Am. Chem. Soc. **128**, 9042–9043 (2006). https://doi.org/10.1021/ja062025p
73. K. Eriks, F. Ray, Fast multidimensional NMR by polarization sharing. Magn. Reson. Chem. **45**, 2–4 (2007). https://doi.org/10.1002/mrc
74. B. Vitorge, G. Bodenhausen, P. Pelupessy, Speeding up nuclear magnetic resonance spectroscopy by the use of SMAll recovery times—SMART NMR. J. Magn. Reson. **207**, 149–152 (2010). https://doi.org/10.1016/j.jmr.2010.07.017
75. R. Freeman, E. Kupče, New methods for fast multidimensional NMR. J. Biomol. NMR **27**, 101–114 (2003). https://doi.org/10.1023/a:1024960302926
76. A. Tal, L. Frydman, Single-scan multidimensional magnetic resonance. Prog. Nucl. Magn. Reson. Spectrosc. **57**, 241–292 (2010). https://doi.org/10.1016/j.pnmrs.2010.04.001
77. P. Schanda, V. Forge, B. Brutscher, Protein folding and unfolding studied at atomic resolution by fast two-dimensional NMR spectroscopy. Proc. Natl. Acad. Sci. U. S. A. **104**, 11257–11262 (2007). https://doi.org/10.1073/pnas.0702069104
78. T. Kiefhaber, H.P. Grunert, U. Hahn, F.X. Schmid, Replacement of a Cis Proline simplifies the mechanism of ribonuclease T1 folding. Biochemistry **29**, 6475–6480 (1990). https://doi.org/10.1021/bi00479a020
79. D.N. Brems, R.L. Baldwin, Amide proton exchange used to monitor the formation of a stable α-helix by residues 3 to 13 during folding of ribonuclease S. J. Mol. Biol. **180**, 1141–1156 (1984). https://doi.org/10.1016/0022-2836(84)90274-2
80. P.S. Kim, R.L. Baldwin, Structural intermediates trapped during the folding of ribonuclease a by amide proton exchange. Biochemistry **19**, 6124–6129 (1980). https://doi.org/10.1021/bi00567a027
81. F.X. Schmid, R.L. Baldwin, Detection of an early intermediate in the folding of ribonuclease A by protection of amide protons against exchange. J. Mol. Biol. **135**, 199–215 (1979). https://doi.org/10.1016/0022-2836(79)90347-4
82. H.J. Dyson, P.E. Wright, Insights into protein folding from NMR. Annu. Rev. Phys. Chem. **47**, 369–395 (1996). https://doi.org/10.1146/annurev.physchem.47.1.369
83. S.W. Englander, L. Mayne, Using hydrogen-exchange labeling and two-dimensional NMR. Annu. Rev. Biophys. Biomol. Struct. **21**, 243–265 (1992)
84. J.B. Udgaonkar, R.L. Baldwin, Early folding intermediate of ribonuclease a. Proc. Natl. Acad. Sci. U. S. A. **87**, 8197–8201 (1990). https://doi.org/10.1073/pnas.87.21.8197
85. J. Frydman, F.U. Hartl, Principles of chaperone-assisted protein folding: differences between in vitro and in vivo mechanisms. Science **272**, 1497–1502 (1996). https://doi.org/10.1126/science.272.5267.1497

86. A.P. Minton, The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. J. Biol. Chem. **276**, 10577–10580 (2001). https://doi.org/10.1074/jbc.R100005200
87. A.P. Minton, Influence of macromolecular crowding upon the stability and state of association of proteins: predictions and observations. J. Pharm. Sci. **94**, 1668–1675 (2005). https://doi.org/10.1002/jps.20417
88. J.F. Bodart, J.M. Wieruszeski, L. Amniai, A. Leroy, I. Landrieu, A. Rousseau-Lescuyer, J.P. Vilain, G. Lippens, NMR observation of tau in Xenopus oocytes. J. Magn. Reson. **192**, 252–257 (2008). https://doi.org/10.1016/j.jmr.2008.03.006
89. P. Selenko, Z. Serber, B. Gadea, J. Ruderman, G. Wagner, Quantitative NMR analysis of the protein G B1 domain in Xenopus laevis egg extracts and intact oocytes. Proc. Natl. Acad. Sci. U. S. A. **103**, 11904–11909 (2006). https://doi.org/10.1073/pnas.0604667103
90. K. Inomata, A. Ohno, H. Tochio, S. Isogai, T. Tenno, I. Nakase, T. Takeuchi, S. Futaki, Y. Ito, H. Hiroaki, et al., High-resolution multi-dimensional NMR spectroscopy of proteins in human cells. Nature **458**, 106–109 (2009). https://doi.org/10.1038/nature07839
91. S. Ogino, S. Kubo, R. Umemoto, S. Huang, N. Nishida, I. Shimada, Observation of NMR signals from proteins introduced into living mammalian cells by reversible membrane permeabilization using a pore-forming toxin, streptolysin O. J. Am. Chem. Soc. **131**, 10834–10835 (2009). https://doi.org/10.1021/ja904407w
92. F.X. Theillet, A. Binolfi, B. Bekei, A. Martorana, H.M. Rose, M. Stuiver, S. Verzini, D. Lorenz, M. Van Rossum, D. Goldfarb, et al., Structural disorder of monomeric α-synuclein persists in mammalian cells. Nature **530**, 45–50 (2016). https://doi.org/10.1038/nature16531
93. C. Li, L.M. Charlton, A. Lakkavaram, C. Seagle, G. Wang, G.B. Young, J.M. Macdonald, G.J. Pielak, Differential dynamical effects of macromolecular crowding on an intrinsically disordered protein and a globular protein: implications for in-cell NMR spectroscopy. J. Am. Chem. Soc. **130**, 6310–6311 (2008). https://doi.org/10.1021/ja801020z
94. F.X. Theillet, A. Binolfi, T. Frembgen-Kesner, K. Hingorani, M. Sarkar, C. Kyne, C. Li, P.B. Crowley, L. Gierasch, G.J. Pielak, et al., Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs). Chem. Rev. **114**, 6661–6714 (2014). https://doi.org/10.1021/cr400695p
95. P. Schanda, E. Kupĉe, B. Brutscher, SOFAST-HMQC experiments for recording two-dimensional deteronuclear correlation spectra of proteins within a few seconds. J. Biomol. NMR **33**, 199–211 (2005). https://doi.org/10.1007/s10858-005-4425-x
96. J.C. Hoch, M.W. Maciejewski, M. Mobli, A.D. Schuyler, A.S. Stern, Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. Acc. Chem. Res. **47**, 708–717 (2014). https://doi.org/10.1021/ar400244v
97. C.A. Waudby, J. Christodoulou, An analysis of NMR sensitivity enhancements obtained using non-uniform weighted sampling, and the application to protein NMR. J. Magn. Reson. **219**, 46–52 (2012). https://doi.org/10.1016/j.jmr.2012.04.013
98. L. Banci, L. Barbieri, I. Bertini, F. Cantini, E. Luchinat, In-cell NMR in E. coli to monitor maturation steps of hSOD1. PLoS One **6**, e23561 (2011). https://doi.org/10.1371/journal.pone.0023561

Molecular Dynamics Simulation Methods to Study Structural Dynamics of Proteins

Anil Kumar and Krishna Kumar Ojha

Abstract Molecular dynamics (MD) simulation is a computational technique for understanding the physical motions of atomic and molecular particles. In this approach, atoms and molecules interact for a defined time period, revealing information on the dynamic evolution of the system. Newton's equations of motion are used to determine the trajectories of atoms and molecules. The forces and potential energy between atoms and molecules are calculated using molecular mechanics force fields or interatomic potentials. The approach was originally created for applications in the field of theoretical physics; however, it is now used in other areas, including materials science, theoretical chemistry, computational biology, etc. This technique determines the time-dependent behaviour of a molecular system. MD simulation has been widely used to study the conformational changes of biomacromolecules to explore the structure and dynamics of proteins, nucleic acids, and their complexes. It has also been used to study the protein–ligand interactions, which are essential for various processes inside the cell, such as signal transduction, immune reaction, and gene regulation. The data help explore the regulatory mechanisms of various biological processes. MD studies also provide a theoretical background for drug design and discovery. Therefore, MD simulation has been extensively used by researchers in combination with biochemical and biophysical methods to obtain a dynamic understanding of biomolecular behaviour. This chapter discusses various MD simulation methods and how they are used to study the structural dynamics of proteins.

Keywords MD simulation · Force fields · Statistical mechanics · Classical mechanics · Periodic boundary condition · Energy minimization · Metadynamics · Umbrella sampling · Protein folding

A. Kumar (✉) · K. K. Ojha
Department of Bioinformatics, Central University of South Bihar, Gaya, India
e-mail: kumaranil@cub.ac.in

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_5

1 Introduction

In living beings, proteins carry out various cellular functions, including transport, cell signalling, and metabolic processes, such as catalysis. All of these processes rely heavily on the structural dynamics of the protein. The protein sequence folds in a specific manner to get a 3D conformation stabilized by various chemical interactions, including covalent and non-covalent interactions. The general approach to understanding the folding process of a protein is studying its unfolding behaviour. Spectroscopic techniques, such as circular dichroism (CD) and fluorescence spectroscopy, are most commonly used for understanding the forces and interactions involved in protein unfolding dynamics [1]. Understanding the protein folding/ unfolding processes necessitates detailed atomic-level data, which could not be obtained using conventional wet-lab spectroscopic techniques. Recent years have seen the development of molecular dynamics (MD) simulation as a tool for understanding protein dynamics at the atomic level [2]. This method provides information about each atom as a function of time to characterize a molecule's dynamic behaviour. MD simulation has the merit of delivering time-dependent information regarding the folding and unfolding processes and inter-residue interactions [3].

In the late 1950s, Alder and Wainwright employed the MD method to investigate the interactions of hard spheres for the first time [4, 5]. Their discoveries shed light on the behaviour of simple liquids in various ways. Rahman made the following significant achievement in 1964 when he simulated liquid argon for the first time using a realistic potential [6]. Rahman and Stillinger's simulation of liquid water in 1971 was the first MD simulation study of a realistic system [7]. In 1977, the first protein simulation was performed [8]. MD simulation of solvated proteins, protein–DNA complexes, and lipids are very common in today's published reports, dealing with various challenges such as ligand binding thermodynamics and protein folding [9]. The number of simulation methodologies has exploded, and there is now a plethora of techniques for specific problems, such as mixed quantum-classical simulations for studying enzyme activities in the context of the entire protein. MD simulation approaches are also extensively utilized in experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy to provide dynamic information about the proteins [10].

The field of MD simulation is rapidly expanding. The improvement of numerous approaches, such as force field advancement, sampling techniques, and superior processing power, has enabled us to do simulations in the microsecond to millisecond range with femtosecond coordinates [11]. MD simulation has the potential to shed light on a variety of biological problems. The use of MD simulation, on the other hand, necessitates the development of optimum models that closely resemble the cellular environment. As a result, the MD simulation will be more effective if more robust algorithms for modelling, docking, scoring, and energy calculations are developed [12].

In the last few decades, various approaches, including X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM), have been used to

Table 1 RCSB PDB statistics as of December 25, 2022

RCSB PDB total entries (199,507)					
S. no.	Methods	Entries	S. no.	Types of molecules	Entries
1	X-ray	171,640	1	Protein only	173,443
2	Nuclear magnetic resonance (NMR)	13,878	2	Protein -NA	11,157
3	Electron microscopy (EM)	13,665	3	DNA only	2247
4	Multi-method	213	4	RNA only	1681
5	Others	111	5	Others	10,979

produce structures of a large number of biomacromolecules. However, there is still a significant disparity between the number of available protein sequences and the available protein structures. UniProtKB/TrEMBL's recent release comprises 22,95,80,745 sequence entries, whereas the protein data bank (PDB) only has 1,99,507 structures as of December 25, 2022. This shows that only a tiny part of all sequences have known structures. The PDB data as of December 25, 2022, are displayed in Table 1. As a result, protein structure prediction is critical for closing this huge gap. The recent development of high-end computing, such as DeepMind's AlphaFold, has been used to create models for half of the understudied (dark) human proteins [13]. It has also determined around 200 million protein structures from almost 1 million species, now available to scientists in DeepMind's database [14].

Though biomolecules are highly dynamic, most of the above approaches provide the structural information of a biomolecule in a static manner [15], such as a protein–small molecule complex presented as a static pose through molecular docking. MD simulation can aid large-scale computing to predict dynamic behaviour [5, 16]. Protein conformational dynamics serve a variety of functions, including transport, signalling molecules, sensors, and mechanical effectors, as well as interacting with the various substrates [17, 18].

In MD simulation, the protein and water molecules are used to create and mimic in vivo environments. Protein and water atoms move in femtosecond (fs) time scale. The forces on each atom are calculated using a force field. The force field includes bonded and non-bonded potential terms in potential energy functions. Newton's law of motion is used to update the velocity and coordinates of the systems, which are also updated in the trajectory with time. The first MD simulation of biological macromolecules was performed by McCammon et al. in 1977 for the bovine pancreatic trypsin inhibitor [8]. Later, researchers explored the role of thermal factor (β) in the internal movements of protein [19–21]. Aspects of mean square variations versus residue number were investigated in these studies. Subsequent advancements in MD simulation revealed a broad spectrum of nucleic acid and protein motions. Since the trajectories can store all the coordinates, they can provide an ensemble of conformations of any structure. From MD simulation data, principal component analysis (PCA) can also be performed [22, 23]. MD simulation provides information on macromolecular structural flexibility and abet in comprehending experimental results, such as NMR parameter dynamics and the effect of solvent and temperature on the stability of a protein [24, 25]. For X-ray structure refinement and NMR structure determination, the simulated annealing method is commonly used [26].

MD simulation can be used for computing the temporal evolution of atomic degrees of freedom by solving Newtonian equations of motion [27]. It allows researchers to observe atomic processes, such as chemical reactions and atomic diffusion, at the atomic time and length scales in large or complex systems. Analysis of repeated simulations, each run under various conditions, allows for the development of a model for a dynamic process [28]. It is among the essential tools for understanding biomolecules theoretically. This approach determines the time-dependent behaviour of a system. It provides precise information on the conformational and structural changes of proteins and nucleic acids [29]. Biomolecular processes occur over a wide range of time scales: side chain and loop motions are classified as local motions (0.01 to 5 Å range) that take 10^{-15} to 10^{-1} s to complete; rigid body motions (1 to 10 Å range) include helix, domain, and subunit motions that typically take 10^{-9} to 1 s; and helix-coil transitions and protein folding are examples of large-scale motions (>5 Å range) that take 10^{-7} to 10^{4} s. Changes that occur in a short period are difficult to view using macroscopic experiments, but by simulating under physiological conditions computationally, the majority of the changes that happen in a short period of time can be visualized. MD simulation enables the investigation of complex biological systems, such as protein stability, protein folding, molecular recognition, ion transport, etc. It also allows researchers to investigate computer-aided drug design using structural information of biomolecules obtained through X-ray and NMR.

2 Statistical Mechanics

MD simulation generates microscopic data such as atomic positions and velocities. Using statistical mechanics, this data can be converted into macroscopic observables such as pressure, energy, and heat capacity [30]. Statistical mechanics is essential for MD simulation of biological systems. MD simulation is commonly used to investigate a system's macroscopic properties using microscopic simulations, such as the calculation of changes in the binding free energy of a candidate drug or to investigate the energetics and processes of conformational changes [31]. The mathematical formulae that correlate macroscopic properties with the motion of the atoms and molecules are provided by statistical mechanics. MD simulation, on the other hand, provides methods for solving particle equations of motion and evaluating these formulae [32]. MD simulation can also be used to investigate thermodynamic features as well as time-dependent (kinetic) processes.

Statistical mechanics is a discipline of physics that looks at macroscopic systems from a molecular perspective to deduce macroscopic phenomena from the properties of the molecules that make up the system and to forecast them [33]. Time-independent statistical averages are frequently used to connect the macroscopic

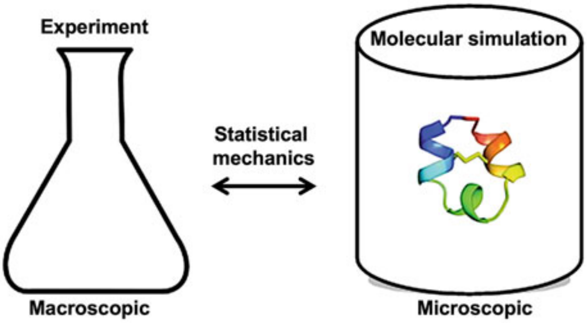

Fig. 1 MD simulation is commonly used to understand the macroscopic properties of a system using microscopic simulations via statistical mechanics

system to the microscopic system. In the following paragraphs, we will try to explain a few definitions of statistical mechanics to represent a physical system.

A thermodynamic state of a system is characterized by a set of parameters, such as temperature, pressure, and the number of particles, *N*. The equations of state and other fundamental thermodynamic equations can be used to calculate various thermodynamic properties. The atomic locations, *q*, and momenta, *p*, constitute the mechanical or microscopic state of a system, which can alternatively be considered coordinates in a multi-dimensional space (phase space). This space has 6*N* dimensions for a system of *N* particles. The state of the system is represented by *G*, a single point in phase space. A group of locations in a phase space that satisfies the criteria of a specific thermodynamic state is termed an ensemble. As a function of time, MD simulation generates a series of points in phase space that are part of the same ensemble and correspond to the various conformations and momenta of the system. There are descriptions of several different ensembles. An ensemble is a collection of all feasible systems with distinct microscopic states but the same macroscopic or thermodynamic state. There are several ensembles available for studying physical systems, such as microcanonical ensemble (NVE), canonical ensemble (NVT), isobaric-isothermal ensemble (NPT), and grand canonical ensemble (μVT). A given number of atoms, a fixed volume, and a fixed energy characterize the thermodynamic state of NVE. This is similar to an isolated system. In NVT, the number of atoms, volume, and temperature are considered fixed. In NPT, the number of atoms, pressure, and temperature remains fixed. However, in μVT, volume and temperature are fixed for a given chemical potential.

An experiment is frequently performed on a macroscopic sample containing a large number of atoms that sample a vast number of different conformations. Averages for experimental observables are defined using ensemble averages in statistical mechanics [34]. An ensemble average is a calculation that takes into account a large number of system copies at the same time (Fig. 1).

The ensemble average is computed as follows:

$$(A)\text{Ensemble} = \iint dp^N dr^N A\left(p^N, r^N\right) \rho\left(p^N, r^N\right)$$

where A (p^N, r^N) is observable, defined as a function of the system's momenta (p) and locations (r). Integration is performed on all possible variables of r and p. The ensemble's probability density is given by

$$\rho\left(p^N, r^N\right) = \frac{1}{Q} \exp\left[-H\left(p^N, r^N\right)/k_{\mathrm{B}}T\right]$$

where H represents the Hamiltonian, T is the temperature, k_{B} is Boltzmann's constant, and Q is the partition function.

$$Q = \iint dp^N dr^N \ \exp\left[-H\left(p^N, r^N\right)/k_{\mathrm{B}}T\right]$$

This integral is highly difficult to calculate since it necessitates calculating all possible system states [35]. Because the points in an ensemble are generated sequentially in time in an MD simulation, the simulations must traverse through all conceivable states that match the specific thermodynamic constraints to calculate an ensemble average. Another method, which is used in MD simulations, is to calculate a temporal average of A, which is written as:

$$(A)\text{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int\limits_{t=0}^{\tau} A\left(p^N(t), r^N(t)\right) dt \approx \frac{1}{M} \sum_{t=1}^{M} A\left(p^N, r^N\right)$$

where t is the time of simulation, M represents time steps, and *A(pN,rN)* is the value of A at a particular instant.

MD simulation can compute only temporal averages, while the experimental observables are considered ensemble averages. The ergodic hypothesis, which is one of the fundamental principles of statistical mechanics, states that the temporal average equals the ensemble average [35]. The central premise is that if a system is allowed to grow indefinitely, it will pass through all possible states. As a result, one of the goals of an MD simulation is to create enough sample conformations to satisfy this equality [36]. Experimentally relevant data on structural and thermodynamic properties can be calculated with a reasonable resource of computing power. As the simulations have a defined time, it is important to sample enough phase space [37]. The average potential energy of the system is represented as

$$V = (V) = \frac{1}{M} \sum_{i=1}^{M} V^i$$

where M represents the trajectory configurations and V^i is the potential energy of a particular configuration.

The average kinetic energy is expressed with the following equation:

$$K = (K) = \frac{1}{M}\sum_{j=1}^{M}\left\{\sum_{i=1}^{N}\frac{M_{\mathrm{i}}}{2}v_{\mathrm{i}}.v_{\mathrm{i}}\right\}j$$

where M represents the number of configurations, N is the atoms number, m_{i} and v_{i} are the mass and velocity of the particle i, respectively. An MD simulation must last long enough to sample a large number of relevant conformations.

3 Classical Mechanics

Newton's second law, $F = ma$ (where F is the force applied on the particle, m is the particle's mass, and a is the particle's acceleration), is the sole foundation of the traditional MD simulation. It is possible to determine each atom's acceleration in a system using the force acting on each atom [38]. A trajectory that depicts the locations, velocities, and accelerations of the particles over time is created after integrating the equations of motion [37]. This method can be used to calculate the average values of the particle properties. Since it is a deterministic method, the system's state can be calculated at any point in time, past or future, if the positions and motions of each atom are known. MD simulation can be time-consuming and expensive; computers, on the other hand, are becoming robust and cheaper. Up to the nanosecond time scale, simulations of solvated proteins can be calculated; nonetheless, simulations of the millisecond time scale have also been recorded using high performance computing. Newton's equation of motion is expressed as:

$$F_{\mathrm{i}} = m_{\mathrm{i}}a_{\mathrm{i}}$$

where F_{i} represents the force acting on particle i, while m_{i} and a_{i} are the mass and acceleration, respectively. The force can also be described as a potential energy gradient.

$$F_{\mathrm{i}} = -\nabla_{\mathrm{i}}V$$

Combining these two equations result to:

$$-\frac{dV}{dr_{\mathrm{i}}} = m_{\mathrm{i}}\frac{d^2r_{\mathrm{i}}}{dt^2}$$

where V represents the system's potential energy. This equation can be used to relate the derivative of potential energy to changes in position as a function of time.

3.1 Newton's Second Law of Motion

$$F = m \cdot a = m \cdot \frac{dv}{dt} = m \cdot \frac{d^2x}{dt^2}$$

Considering the acceleration as constant

$$a = \frac{dv}{dt}$$

After integration, the expression for the velocity can be written as

$$v = at + v_0$$

since

$$v = \frac{dx}{dt}$$

after further integration

$$x = v \cdot t + x_0$$

Combining the above equation with the velocity, we get the below relation that gives the value of x at time t as a function of the initial position (x_0), the acceleration (a), and the initial velocity (v_0).

$$x = \frac{1}{1} a * t^2 + v_0 * t + x_0$$

The acceleration is calculated using the derivative of potential energy with respect to the position (r).

$$a = -\frac{1}{m} \frac{dE}{dr}$$

Therefore, the initial positions of the atoms, an initial velocity distribution, as well as the acceleration determined by the gradient of the potential energy function are all required to construct a trajectory [39]. The positions and velocities at time zero determine the positions and velocities at every other time (t), as the motion equations are deterministic. The initial positions can be taken from experimental structures, such as the protein's X-ray crystal structure or NMR structure. The initial velocity distribution is commonly derived from a random distribution with magnitudes that conform to the requisite temperature and that are corrected to ensure that there is no overall momentum, which is represented by

$$p = \sum_{i=1}^{N} m_{\mathrm{i}} v_{\mathrm{i}} = 0$$

The probabilities of an atom having a velocity v_{x} in the x direction at a temperature T are determined by selecting velocities, v_{i}, at random from a Maxwell-Boltzmann or Gaussian distribution.

$$p(v_{\mathrm{ix}}) = \left(\frac{m_{\mathrm{i}}}{2\pi k_{\mathrm{b}} T} \right) \frac{1}{2} \exp\left[- \frac{1}{2} \frac{m_{\mathrm{i}} v_{\mathrm{ix}}^2}{k_{\mathrm{b}} T} \right]$$

The temperature can be obtained as follows:

$$T = \frac{1}{(3N)} \sum_{i=1}^{N} \frac{|\, p_{\mathrm{i}} \,|}{2m_{\mathrm{i}}}$$

where N represents the number of atoms in the system.

3.2 Integration Algorithms

The potential energy is a function of all atoms in a system's atomic locations ($3N$). The equations of motion have no analytic solution due to the intricate nature of this function; they must be solved numerically [40]. Several numerical techniques have been devised to integrate the equations of motion, such as the Verlet algorithm, Leap-frog algorithm, Velocity Verlet ,and Beeman's algorithm. While choosing an algorithm, one should consider that the algorithm should conserve energy and momentum. It should be computationally efficient and allow a long-time step for integration. The locations, velocities, and accelerations of all the integration techniques are assumed to be approximated by a Taylor series expansion:

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2} a(t)\delta t^2 + \ldots$$

$$r(t + \delta t) = v(t) + a(t)\delta t + \frac{1}{2} b(t)\delta t^2 + \ldots$$

$$a(t + \delta t) = a(t) + b(t)\delta t + \ldots$$

where r represents the position, v is the velocity (the first derivative with respect to time), and a is the acceleration (the second derivative with respect to time), etc.

3.2.1 Verlet Algorithm

To derive the **Verlet** algorithm, one can write:

$$r(t+\delta t)=r(t)+v(t)\delta t+\frac{1}{2}a(t)\delta t^2$$

$$r(t-\delta t)=r(t)-v(t)\delta t+\frac{1}{2}a(t)\delta t^2$$

After summing the above two equations:

$$r(t+\delta t)=2r(t)-r(t-\delta t)+\frac{1}{2}a(t)\delta t^2$$

The Verlet method calculates new positions at time *t* + *dt* by combining locations and accelerations at time *t* with positions from time *t-dt*. There are no stated velocities in the Verlet algorithm. The Verlet algorithm is simple with minimal storage needs, but the disadvantage is that the algorithm is of moderate precision [41].

3.2.2 The Leap-Frog Algorithm

In this method, the velocities are calculated at time *t* + 1/2*dt*. Further, these are used to find the positions (*r*) at time *t* + *dt*. In this way, the velocities *leap* over the positions, and then the positions *leap* over the velocities [42].

$$r(t+\delta t)=r(t)+v\left(t+\frac{1}{2}\delta t\right)\delta t$$

$$v\left(t+\frac{1}{2}\delta t\right)=v\left(t+\frac{1}{2}\delta t\right)+a(t)\delta t$$

This approach has the advantage of explicitly calculating velocities; however, it has the disadvantage of not doing so simultaneously with the positions. The relationship can be used to approximate the velocities at time *t*.

$$v(t)=\frac{1}{2}\left[v\left(t-\frac{1}{2}\delta t\right)+v\left(t+\frac{1}{2}\delta t\right)\right]$$

3.2.3 The Velocity Verlet Algorithm

This algorithm returns positions, velocities, and accelerations at time t. Precision is not uncompromised.

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2$$

$$r(t+\delta t) = v(t) + \frac{1}{2}[a(t) + a(t+\delta t)]\delta t$$

3.2.4 Beeman's Algorithm

This algorithm is very similar to the Verlet algorithm.

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{2}{3}a(t)\delta t^2 - \frac{1}{6}a(t-\delta t)\delta t^2$$

$$v(t+\delta t) = v(t) + v(t)\delta t + \frac{1}{3}a(t)\delta t + \frac{5}{6}a(t)\delta t - \frac{1}{6}a(t-\delta t)\delta t$$

This algorithm has the advantage of providing a more accurate expression for velocities and better energy conservation [43]. The disadvantage is that more complex expressions increase the cost of the calculation.

4 Principle of MD Simulation

The Born-Oppenheimer approximation, which separates the slow atomic degrees of freedom from the fast motion of light electrons, is the core of MD simulation [43]. Binding in solids and molecules is due to the interaction of electrons with the atomic core, which is seen almost at rest by the electrons. This interaction also provides interatomic forces when the atom cores are treated as classical particles. While it was required to approximate these forces with appropriate interatomic potentials initially, the introduction of fast electronic computers and the Car–Parrinello method enabled the interaction to be treated on a first-principles basis, allowing predictive quantitative simulations [43].

4.1 Periodic Boundary Condition

Periodic boundary conditions (PBCs) are a group of boundary conditions that are used to approximate a large (infinite) system with a small component called a unit cell. Simulations and mathematical modelling frequently employ PBCs. PBCs are used in MD simulation to eliminate finite-size boundary effects and to make the system similar to an infinite one at the expense of potential periodicity effects [43]. The existence of PBC ensures that every atom that exits a simulation box via the right-hand face must re-enter via the left-hand face. If we look at the face of the simulation box opposite the one from where the protein is protruding in the case of a large protein, we will notice a hole in the solvent. The molecule(s) shift from where they were initially situated within the box since they are free to diffuse around in most simulations [44]. During the simulation, the box is not centred on anything. Molecules are not automatically made complete. Using PBCs to solve the surface-effects problem is an alternate and preferred method. PBCs can be approached in various ways, but we will stick to the minimum-image convention. We must first understand the concept of a unit cell before discussing PBCs. A unit cell is the simplest representation of a system. If we are imitating a crystal, we might pick a tiny cell with a few hundred atoms that match the desired crystal form. When simulating a gas, our unit cell could be a small volume containing several hundred gas molecules. We can start with a tiny unit cell volume, even if the purpose of the simulation is to obtain insight into bulk crystal or gas properties (easily $>10^{10}$ molecules). Then we can make neighbouring copies (images) of the unit cell that duplicate the contents of the unit cell in adjacent volumes. The images are a duplicate of the original simulation region and are used to lessen or eliminate border effects by providing an equivalent surrounding environment of atoms to every atom in the unit cell, independent of position in the unit cell. We can update the original unit cell's positions, forces, and velocity. Mirror replicas of the unit cell will be updated in the surrounding image cells. As a result, the atoms in the image cells have no physical significance on their own and are only constructs for PBCs.

The seeming "never-ending" aspect of the unit cell, that is, when an atom exits through a wall in the unit cell, it subsequently re-enters on the opposite side of the unit cell with the same velocity, is one of the simple and attractive effects of PBCs [45]. The layout of the image cells supports this continuity because as an atom departs the unit cell, the same atom's image may be seen entering the unit cell from an image.

4.2 Ewald Summation

Ewald summation is a technique for calculating long-range interactions in periodic systems, such as electrostatic interactions [46]. The total electrostatic energy of NN particles and their periodic images can be calculated using the following:

$$V = \frac{f}{2} \sum_{n_\mathrm{x}} \sum_{n_\mathrm{y}} \sum_{n_\mathrm{z}^*} \sum_{i}^{N} \sum_{j}^{N} \frac{q_\mathrm{i} q_\mathrm{j}}{r_{\mathrm{ij,n}}}$$

The box index vector is $(n_\mathrm{x}, n_\mathrm{y}, n_\mathrm{z}) = n$, and the asterisk mark designates that terms with $i = j$ should be omitted in case $(n_\mathrm{x}, n_\mathrm{y}, n_\mathrm{z}) = (0,0,0)$. The distance $r_{\mathrm{ij,n}}$, as opposed to the minimum image, represents the actual distance between the charges. Although incredibly slow, this sum is conditionally convergent. Ewald summation was initially developed to determine the long-range interactions of the periodic images in crystals. The goal is to split the single, slowly convergent sum into two components that swiftly converge and a constant term.

4.3 Particle Mesh Ewald (PME) Method

The Particle Mesh Ewald (PME) method is given by Tom Darden to enhance the reciprocal sum performance. The charges are interpolated to a grid instead of just adding wave vectors. Cardinal B-spline interpolation or smooth PME (SPME), is used in GROMACS [47]. Using a 3D FFT technique, the grid is then Fourier transformed, and the reciprocal energy term is calculated by summing the grid in k-space. The inverse transformation is used to calculate the potential at grid points, and interpolation factors are used to determine the forces working on each atom. In a medium to large systems, the PME technique is noticeably faster than standard Ewald summation. Ewald may still be preferable on relatively small systems to save time setting up grids and transforms. The PME direct space potential is moved by a constant in the Verlet cut-off scheme so that the potential is zero at the cut-off. In contrast to the Lennard-Jones potential, where all shifts add up, this shift is minor, and because the net system charge is almost zero, the total shift is also minimal. We nonetheless apply the shift to make the potential precisely equal to the integral of the force.

4.4 Thermostat in MD

By altering the system’s temperature in some way, thermostats are intended to assist a simulation sample from the appropriate ensemble (i.e. NVT or NPT). We must first define what is meant by temperature. The “instantaneous (kinetic) temperature” in simulations is typically calculated from the system’s kinetic energy using the equipartition theorem. In other words, the system’s total kinetic energy is used to calculate the temperature [48]. The purpose of a thermostat is not to maintain a constant temperature because doing so would mean fixing the total kinetic energy, which is wrong and not what NVT or NPT are intended to do. Instead, it guarantees that a system’s average temperature is correct.

Consider a glass of water placed in a space to understand this case. Consider estimating the kinetic energy of a few molecules in a small area of the glass by looking at them extremely closely [49]. Because there are so few particles, you would not anticipate the kinetic energy to be perfectly constant; instead, you would anticipate fluctuations in the kinetic energy. The fluctuations in the average decrease as you average across more and more particles, and when you ultimately consider the entire glass, you can conclude that it has a “constant temperature”. Compared to a glass of water, MD simulations are quite small, which causes larger fluctuations [50]. Therefore, it would be fair to consider the role of the thermostat in this situation to ensure that we have the proper average temperature and fluctuations of the correct size.

4.5 Solvent Models

A solvent model is a computer technique used in computational chemistry to predict the behaviour of solvated condensed phases. Simulations and thermodynamic calculations for reactions and processes that occur in solutions are made possible by solvent models [51]. Environmental, chemical, and biological processes are among them. Such computations can result in new predictions about the physical processes due to greater understanding. Generally, there are two groups of models: explicit and implicit models, each of which has advantages and disadvantages of its own [52]. Implicit models often have good computing efficiency and can provide a good description of the behaviour of the solvent, but they are unable to consider the local variations in solvent density near a solute molecule. When water is used as a solvent, the density fluctuation behaviour caused by solvent ordering around a solute is more common. Explicit models can provide a physical, spatially detailed description of the solvent but are frequently inefficient in terms of computational efficiency [53]. Although many of these explicit models may fail to replicate specific experimental results, this is often due to differences in fitting methods and parameterization.

4.6 Energy Minimization

Energy minimization is the process of arranging a group of atoms in space in such a way so that the net interatomic force acting on each atom is as close to zero as possible while it is stationary on the potential energy surface (PES) [53]. The atoms could combine to form a single molecule, an ion, a condensed phase, a transition state, or a combination of these.

5 Current Tools for Molecular Dynamics

Various tools are available for performing MD simulations both in proprietary and open-source domains. Some of them are discussed below:

5.1 Gromacs

Gromacs is an MD simulation software package designed primarily for simulating proteins, nucleic acids, and lipids. It was created in the University of Groningen's Biophysical Chemistry department, and it is now maintained by various contributors from research institutions all over the world. It is one of the widely used software available that can run on a computer with basic configuration as well as on high-end workstations. It is a freely available open-source software distributed under the General Public License (GNU) [54, 55].

5.2 Amber

A set of bimolecular simulation tools are included in Amber. It was started in the late 1970s, and a vibrant development community continues to maintain it. Two objects are being referred to by the term "Amber". First, it is a collection of molecular mechanical force fields for simulating biomolecules available in the public domain. Second, it is a collection of molecular simulation programs that also includes demonstrations. AmberTools21 and Amber20 are the two components of Amber. AmberTools21 can be used without Amber20, but not the other way around [56].

5.3 CHARMM

CHARMM is a molecular simulation tool with extensive applicability to many-particle systems that supports multi-scale methods, including quantum mechanics/molecular mechanics (QM/MM), molecular mechanics/coarse-grained (MM/CG), a variety of implicit solvent models, and a large collection of energy functions. It targets biomolecules such as proteins, small molecules, nucleic acids, lipids, and carbohydrates found in solution, crystals, and membrane environments. CHARMM also have a wide range of applications for inorganic materials. CHARMM includes a comprehensive set of tools for analysis and model construction. It performs well on a variety of systems, such as GPUs and parallel clusters [57].

5.4 NAMD

NAMD is a parallel MD programme designed for the high-performance modelling of large biomolecular systems. It won the Gordon Bell Award in 2002, the Sidney Fernbach Award in 2012, and the Gordon Bell Prize in 2020. As NAMD is based on Charm++ parallel objects, it can scale from hundreds to over half a million cores for larger simulations. NAMD uses the well-known molecular graphics application VMD to set up the simulation and analyse the trajectory, in addition to being file-compatible with AMBER, CHARMM, and X-PLOR. The source code for NAMD is freely available. NAMD is available as a self-built project or as downloadable binaries for many different platforms [58].

5.5 HyperChem

HyperChem is a powerful molecular modelling system with a reputation for excellence, adaptability, and user-friendliness. HyperChem provides access to several molecular modelling tools more than any other programme by fusing 3D visualization and animation with different computational approaches. It incorporates several computational techniques, including density functional theory, semi-empirical and ab-initio molecular orbital methods, MD, and MM. HyperChem includes HyperChem Data and HyperNMR. Among the frequently added features are CHARMM protein simulations, molecules in magnetic fields, compatibility with third-party applications, etc. Both large and small molecules can be used with HyperChem. It also supports scripting [59].

6 GUI-Based Software for MD Trajectories Analysis

The MD simulation output trajectories can be visualized using GUI-based software. The following are some of the most popular software:

6.1 Visual Molecular Dynamics (VMD)

A research group of theoretical biophysics at the University of Illinois created visual molecular dynamics (VMD) [60–62]. It is a highly effective tool for observing and investigating various biological systems, including nucleic acids, proteins, lipids, and carbohydrates. It supports a wide range of formats, such as PDB and GROMOS for biomolecules. It can handle a massive amount of data processing to display changes in a trajectory [63]. The molecules can be seen as animated, and the input trajectory can also be used to make a movie. It works with any operating system that has a basic computer configuration. It is included with NAMD as well. The additional features of VMDs include [64]:

1. It can be used to visualize macromolecules.
2. An amino acid and atoms can be chosen.
3. Structure alignment can be performed.

4. Support for user activity logs is present.
5. Support for the Raster3D format is available.
6. Ramachandran plots can be generated.
7. It supports various molecular image types.
8. It supports command lines.
9. It utilizes vectors and arrays.
10. It has JavaScript support.

6.2 PyMOL

Structural biologists extensively use the PyMOL software [65]. PyMOL can accept various file formats, including SDF, Mol2, PDB, etc. The trajectory can be imported, and the simulation results can be analysed on PyMOL. A surface view model can be generated. To further study the MD simulation results, several additional plug-ins are available. The user can use this tool to create high-quality figures as well as animated movies.

6.3 Chimera

UCSF Chimera is a sophisticated tool for molecular modelling systems that is free for academic usage [66]. Advanced UCSF ChimeraX is also freely available for academic use. The Gromacs and Amber trajectory formats are supported by Chimera 1.13.1 and later versions. Following the import of these trajectories, the user can create a movie with a time frame and produce attractive images. Aligning two or more structures is possible. The surface cavity analysis during the trajectory run can also be generated. It supports the command line option and has a variety of functions.

7 Other Advanced MD Simulation Methods

7.1 Metadynamics

An improved sampling technique known as metadynamics uses a set of collective variables (CVs) that specify transitions along a reaction coordinate to explain the system. The system's position in this CV space is established during the simulation, and then positive biasing Gaussian functions are added, modifying the system's Hamiltonian [67].

$$H = T + V + \sum \mathrm{V}_{\mathrm{GAUSSH}} = T + V + + \sum \mathrm{V}_{\mathrm{GAUSSH}}$$

As a result of the accumulation of these Gaussian functions in properly sampled regions of the CV space, the system can more easily navigate through regions of CV space that correspond to free energy maxima while simulating the unmodified Hamiltonian. The simulation can now examine the entire energy landscape. Knowing the sampling of the modified Hamiltonian and the deposited Gaussian functions allows one to retrieve the free energy surface of the unmodified Hamiltonian.

7.2 Umbrella Sampling

The purpose of any MD simulation is to sample all possible states in which a molecule may exist. Based on this method, the probability (free energy) for the molecule to be in any state can be determined. Often, some protein states are separated from others by very high energy barriers. Sometimes, it would take years of conventional MD simulation to go through all molecular states. Umbrella sampling allows us to accelerate the sampling by flattening those hills and ridges, which prevents MD simulation from accessing certain states. In umbrella sampling, the energy landscape is flattened by adding artificial umbrella potentials that are supposed to mirror and thus annihilate the real barriers. However, making an umbrella potential account for all degrees of freedom in the system would be difficult. Hence, the umbrella potential involves only a few (one to three) degrees of freedom, often called CVs or reaction coordinates. The sampling of a system is considered complete when it has visited all values of CVs for an accurate and unbiased calculation of state probabilities [68].

8 Structural Parameters to Analyse MD Simulation Data

8.1 Root Mean Square Deviation (RMSD)

The root mean square deviation (RMSD) is the Euclidean distance between the structure and a reference structure that measures the relaxation between the structures [69]. It is a common way to quantify the distance of structural coordinates. It determines how far apart, on average, a group of atoms, such as the protein's backbone atoms, are from one another [70]. Calculating the RMSD between two sets of atomic coordinates, such as two points in time from the trajectory, measures how much the protein structure has changed. It is possible to compute the RMSD for each residue, the backbone, the side chains, and C-alpha. It is calculated with reference to the simulation time [71]. A lower RMSD value indicates a very stable structure over the course of a simulation.

8.2 Root Mean Square Fluctuation (RMSF)

The average variation of a particle over time from a reference position is measured by the root mean square fluctuation (RMSF) [69]. As a result, RMSF examines the structural elements that deviate the most from their mean structure. The variability around each atom's average position is captured by the RMSF. This reveals information on the flexibility of the protein's various regions and relates to the crystallographic B-factors. It can be used to check whether the simulation findings are consistent with the crystal structure because one would typically anticipate similar profiles for the RMSF and the B-factors. Atoms in bends and coils fluctuate more than in helices and sheets; hence they have lower RMSF values, whereas bends and coils have higher RMSF values.

8.3 Radius of Gyration (Rg)

The measurement of the radius of gyration (Rg) indicates the shape and compactness of a molecule at a particular time. The gyrating radius is compared to the hydrodynamic radius that can be measured empirically [69]. Additionally, this provides the individual components that are equivalent to the eigenvalues of the inertia matrix. This means that the first component corresponds to the molecule's longest axis and the last to its shortest. The three axes effectively provide a global indication of the shape of the molecule [72].

8.4 Solvent Accessible Surface Area (SASA)

The area of the protein that is accessible to solvent is known as solvent accessible surface area (SASA), which can be further divided into a hydrophilic and hydrophobic SASA [73, 74]. The SASA of the expanded form of the protein is higher than that of the folded globular protein. It is known that when a temperature of a system rises, proteins begin to unfold and expose their hydrophobic interiors to the solvent. SASA consequently increases upon unfolding. Additionally, the SASA can be used to estimate the free energy of solvation together with a few empirical parameters [75].

8.5 Hydrogen Bonds

The number of internal hydrogen bonds with a protein or external hydrogen bonds between a protein and its surrounding solvent is another characteristic that can be informative [76, 77]. The distance between a donor-H acceptor pair and the donor-H acceptor angle can be used to determine the presence or absence of a hydrogen bond. Hydrogen bonds are vital for maintaining protein secondary structures; therefore, simulations of protein folding must adequately represent the hydrogen bond interactions. Hydrogen bonding is treated as a non-bonded interaction in modern classical force fields where electrostatic interactions are predominant. Atomic charges, on the other hand, are fixed and are established in a mean-field fashion in the frequently utilized non-polarizable force fields. The native structure cannot be appropriately populated when the non-polarizable AMBER force field is used in the folding simulations of small peptides. When the polarization effect is added to the simulation using either the on-the-fly charge fitting or the polarizable hydrogen bond model, the native structure becomes more prominent in the free energy landscape. These results emphasize how crucial the electrostatic polarization effect is for simulating proteins [78].

9 Summary

MD simulation has been a popular method for understanding the dynamic representation of any biological system for the last few decades [79–81]. In recent years, GPU-based high computational capability systems have significantly reduced the time required in the MD simulation of biological macromolecules. It is a handy technique for understanding molecular interactions such as protein–protein and protein–ligand interactions, as well as protein folding. It creates the cell-like environment around the macromolecules by considering pH and the surrounding molecules such as water, lipids, ions, as well as co-enzymes. It provides atomic-level interaction details that offer insights into how molecules function. Tools such as MM-PBSA can be used to predict the free energy of binding, various energy constituents, and contribution to binding with a small molecule at the residue level. The implementation of the QM and MM method in the MD simulation has improved the accuracy of these predictions. MD simulation can thus be used to investigate the dynamics of a biological system by selecting an appropriate model and physical conditions.

References

1. S.M. Kelly, N.C. Price, The application of circular dichroism to studies of protein folding and unfolding. BBA-Protein Struct. M. **1338**, 161–185 (1997)
2. U. Mayor, C.M. Johnson, V. Daggett, A.R. Fersht, Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proc. Natl. Acad. Sci. **97**, 13518–13522 (2000)
3. H.A. Scheraga, M. Khalili, A. Liwo, Protein-folding dynamics: overview of molecular simulation techniques. Annu. Rev. Phys. Chem. **58**, 57–83 (2007)

4. B.J. Alder, T.E. Wainwright, Phase transition for a hard sphere system. J. Chem. Phys. **27**, 1208–1209 (1957)
5. B.J. Alder, T.E. Wainwright, Studies in molecular dynamics. I. General method. J. Chem. Phys. **31**, 459–466 (1959)
6. A. Rahman, Correlations in the motion of atoms in liquid argon. Phys. Rev. **136**, A405 (1964)
7. A. Rahman, F.H. Stillinger, Molecular dynamics study of liquid water. J. Chem. Phys. **55**, 3336–3359 (1971)
8. J.A. McCammon, B.R. Gelin, M. Karplus, Dynamics of folded proteins. Nature **267**, 585–590 (1977)
9. A. Khandelwal, V. Lukacova, D. Comez, D.M. Kroll, S. Raha, S. Balaz, A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. J. Med. Chem. **48**, 5437–5447 (2005)
10. M. Billeter, G. Wagner, K. Wüthrich, Solution NMR structure determination of proteins revisited. J. Biomol. NMR **42**, 155–158 (2008)
11. C. Park, J. Jung, G.J. Yun, Multiscale micromorphic theory compatible with MD simulations in both time-scale and length-scale. Int. J. Plast. **129**, 102680 (2020)
12. M.C. Zwier, L.T. Chong, Reaching biological timescales with all-atom molecular dynamics simulations. Curr. Opin. Pharmacol. **10**, 745–752 (2010)
13. J.L. Binder, J. Berendzen, A.O. Stevens, Y. He, J. Wang, N.V. Dokholyan, T.I. Oprea, AlphaFold illuminates half of the dark human proteins. Curr. Opin. Struct. Biol. **74**, 102372 (2022)
14. E. Callaway, “The entire protein universe”: aI predicts shape of nearly every known protein. Nature **608**, 15–16 (2022)
15. R.O. Dror, R.M. Dirks, J.P. Grossman, H. Xu, D.E. Shaw, Biomolecular simulation: a computational microscope for molecular biology. Annu. Rev. Biophys. **41**, 429–452 (2012)
16. V. Rajendran, R. Shukla, H. Shukla, T. Tripathi, Structure-function studies of the asparaginyl-tRNA synthetase from Fasciola gigantica: understanding the role of catalytic and non-catalytic domains. Biochem. J. **475**, 3377–3391 (2018)
17. T. Pandey, R. Shukla, H. Shukla, A. Sonkar, T. Tripathi, A.K. Singh, A combined biochemical and computational studies of the rho-class glutathione s-transferase sll1545 of synechocystis PCC 6803. Int. J. Biol. Macromol. **94**, 378–385 (2017)
18. A. Sonkar, H. Shukla, R. Shukla, J. Kalita, T. Pandey, T. Tripathi, UDP-N-acetylglucosamine enolpyruvyl transferase (MurA) of acinetobacter baumannii (AbMurA): structural and functional properties. Int. J. Biol. Macromol. **97**, 106–114 (2017)
19. B. Brooks, M. Karplus, Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. **80**, 6571–6575 (1983)
20. A.T. Brünger, C.L. Brooks 3rd, M. Karplus, Active site dynamics of ribonuclease. Proc. Natl. Acad. Sci. **82**, 8458–8462 (1985)
21. J. Smith, S. Cusack, U. Pezzeca, B. Brooks, M. Karplus, Inelastic neutron scattering analysis of low frequency motion in proteins: a normal mode study of the bovine pancreatic trypsin inhibitor. J. Chem. Phys. **85**, 3636–3654 (1986). https://doi.org/10.1063/1.450935
22. C.C. David, D.J. Jacobs, Principal component analysis: a method for determining the essential dynamics of proteins, in *Protein Dynamics*, ed. by D. Livesay, (Springer, Heidelberg, 2014), pp. 193–226
23. A. Wolf, K.N. Kirschner, Principal component and clustering analysis on molecular dynamics data of the ribosomal L11 23S subdomain. J. Mol. Model. **19**, 539–549 (2013)
24. F. Colonna-Cesari, D. Perahia, M. Karplus, H. Eklund, C.I. Brädén, O. Tapia, Interdomain motion in liver alcohol dehydrogenase. Structural and energetic analysis of the hinge bending mode. J. Biol. Chem. **261**, 15273–15280 (1986)
25. L. Nilsson, G.M. Clore, A.M. Gronenborn, A.T. Brünger, M. Karplus, Structure refinement of oligonucleotides by molecular dynamics with nuclear overhauser effect interproton distance restraints: application to 5′ d (CGTACG) 2. J. Mol. Biol. **188**, 455–475 (1986)

26. D.A. Case, M. Karplus, Dynamics of ligand binding to heme proteins. J. Mol. Biol. **132**, 343–368 (1979)
27. P. Banáš, P. Jurečka, N.G. Walter, J. Šponer, M. Otyepka, Theoretical studies of RNA catalysis: hybrid QM/MM methods and their comparison with MD and QM. Methods **49**, 202–216 (2009)
28. S. Ahmadi, L. Barrios Herrera, M. Chehelamirani, J. Hostaš, S. Jalife, D.R. Salahub, Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: a tutorial review. Int. J. Quantum Chem. **118**, e25558 (2018)
29. A.W. Götz, M.A. Clark, R.C. Walker, An extensible interface for QM/MM molecular dynamics simulations with AMBER. J. Comput. Chem. **35**, 95–108 (2014)
30. R.C. Tolman, *The Principles of Statistical Mechanics* (Courier Corporation, New York, 1979)
31. R.H. Fowler, *Statistical Mechanics* (Cambridge University Press, Cambridge, 1967)
32. Y. Levin, R. Pakter, F.B. Rizzato, T.N. Teles, F.P. Benetti, Nonequilibrium statistical mechanics of systems with long-range interactions. Phys. Rep. **535**, 1–60 (2014)
33. M. Takano, K. Nagayama, A. Suyama, Investigating a link between all-atom model simulation and the Ising-based theory on the helix–coil transition: equilibrium statistical mechanics. J. Chem. Phys. **116**, 2219–2228 (2002)
34. M.J. Uline, D.W. Siderius, D.S. Corti, On the generalized equipartition theorem in molecular dynamics ensembles and the microcanonical thermodynamics of small systems. J. Chem. Phys. **128**, 124301 (2008)
35. J.R. Ray, H. Zhang, Correct microcanonical ensemble in molecular dynamics. Phys. Rev. E **59**, 4781 (1999)
36. A. Johnson, T. Johnson, A. Khan, Thermostats in molecular dynamics simulations. UMass. **1**, 29 (2012)
37. P.V. Coveney, S. Wan, On the calculation of equilibrium thermodynamic properties from molecular dynamics. Phys. Chem. Chem. Phys. **18**, 30236–30240 (2016)
38. R.D. Gregory, *Classical Mechanics* (Cambridge University Press, Cambridge, 2006)
39. T. Kibble, F.H. Berkshire, *Classical Mechanics* (World Scientific Publishing Company, Singapore, 2004)
40. H. Grubmüller, H. Heller, A. Windemuth, K. Schulten, Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. Mol. Simul. **6**, 121–142 (1991)
41. Q. Spreiter, M. Walter, Classical molecular dynamics simulation with the velocity Verlet algorithm at strong external magnetic fields. J. Comput. Phys. **152**, 102–119 (1999)
42. W.F. Van Gunsteren, H.J. Berendsen, A leap-frog algorithm for stochastic dynamics. Mol. Simul. **1**, 173–185 (1988)
43. R.W. Pastor, B.R. Brooks, A. Szabo, An analysis of the accuracy of Langevin and molecular dynamics algorithms. Mol. Phys. **65**, 1409–1419 (1988)
44. J. Shimada, H. Kaneko, T. Takada, Efficient calculations of Coulombic interactions in biomolecular simulations with periodic boundary conditions. J. Comput. Chem. **14**, 867–878 (1993)
45. W. Tian, L. Qi, X. Chao, J. Liang, M. Fu, Periodic boundary condition and its numerical implementation algorithm for the evaluation of effective mechanical properties of the composites with complicated micro-structures. Compos. Part B **162**, 1–10 (2019)
46. A.Y. Toukmaji, J.A. Board Jr., Ewald summation techniques in perspective: a survey. Comput. Phys. Commun. **95**, 73–92 (1996)
47. J. Kolafa, J.W. Perram, Cutoff errors in the Ewald summation formulae for point charge systems. Mol. Simul. **9**, 351–368 (1992)
48. E.A. Koopman, C.P. Lowe, Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations. J. Chem. Phys. **124**, 204103 (2006)
49. E. Rosta, N.-V. Buchete, G. Hummer, Thermostat artifacts in replica exchange molecular dynamics simulations. J. Chem. Theory Comput. **5**, 1393–1399 (2009)
50. R. Kutteh, R.B. Jones, Rigid body molecular dynamics with nonholonomic constraints: molecular thermostat algorithms. Phys. Rev. E **61**, 3186–3198 (2000). https://doi.org/10.1103/PhysRevE.61.3186

51. P. Ferrara, J. Apostolakis, A. Caflisch, Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins **46**, 24–33 (2002). https://doi.org/10.1002/prot.10001
52. H. Nguyen, D.R. Roe, C. Simmerling, Improved generalized born solvent model parameters for protein simulations. J. Chem. Theory Comput. **9**, 2020–2034 (2013). https://doi.org/10.1021/ct3010485
53. A. Malevanets, R. Kapral, Mesoscopic model for solvent dynamics. J. Chem. Phys. **110**, 8605–8613 (1999). https://doi.org/10.1063/1.478857
54. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, GROMACS: fast, flexible, and free. J. Comput. Chem. **26**, 1701–1718 (2005). https://doi.org/10.1002/jcc.20291
55. S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics **29**, 845–854 (2013). https://doi.org/10.1093/bioinformatics/btt055
56. D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, R.J. Woods, The Amber biomolecular simulation programs. J. Comput. Chem. **26**, 1668–1688 (2005). https://doi.org/10.1002/jcc.20290
57. B.R. Brooks, C.L. Brooks, A.D. MacKerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, M. Karplus, CHARMM: the biomolecular simulation program. J Comput Chem **30**, 1545–1614 (2009). https://doi.org/10.1002/jcc.21287
58. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD. J. Comput. Chem. **26**, 1781–1802 (2005). https://doi.org/10.1002/jcc.20289
59. M. Froimowitz, HyperChem: a software package for computational chemistry and molecular modeling. BioTechniques **14**, 1010–1013 (1993)
60. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. J. Mol. Graph. **14**, 33–38., 27–28 (1996). https://doi.org/10.1016/0263-7855(96)00018-5
61. J. Hsin, A. Arkhipov, Y. Yin, J.E. Stone, K. Schulten, Using VMD: an introductory tutorial. Curr Protoc Bioinformatics **Chapter 5**, Unit 5.7 (2008). https://doi.org/10.1002/0471250953.bi0507s24
62. B. Knapp, N. Lederer, U. Omasits, W. Schreiner, vmdICE: a plug-in for rapid evaluation of molecular dynamics simulations using VMD. J Comput Chem **31**, 2868–2873 (2010). https://doi.org/10.1002/jcc.21581
63. S. Falsafi-Zadeh, Z. Karimi, H. Galehdari, VMD DisRg: new user-friendly implement for calculation distance and radius of gyration in VMD program. Bioinformation **8**, 341–343 (2012). https://doi.org/10.6026/97320630008341
64. I.V. Likhachev, N.K. Balabaev, O.V. Galzitskaya, Available instruments for analyzing molecular dynamics trajectories. Open Biochem. J. **10**, 1–11 (2016). https://doi.org/10.2174/1874091X01610010001
65. D. Seeliger, B.L. de Groot, Ligand docking and binding site analysis with PyMOL and autodock/Vina. J. Comput. Aided Mol. Des. **24**, 417–422 (2010). https://doi.org/10.1007/s10822-010-9352-6
66. E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. **25**, 1605–1612 (2004). https://doi.org/10.1002/jcc.20084
67. D. Hollas, L. Šištík, E.G. Hohenstein, T.J. Martínez, P. Slavíček, Nonadiabatic Ab initio molecular dynamics with the floating occupation molecular orbital-complete active space configuration interaction method. J. Chem. Theory Comput. **14**, 339–350 (2018). https://doi.org/10.1021/acs.jctc.7b00958

68. G.M. Torrie, J.P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J. Comput. Phys. **23**, 187–199 (1977). https://doi.org/10.1016/0021-9991(77)90121-8
69. R. Brüschweiler, R.M.S.D. Efficient, Measures for the comparison of two molecular ensembles. Root-mean-square deviation. Proteins **50**, 26–34 (2003). https://doi.org/10.1002/prot.10250
70. O. Carugo, How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. J. Appl. Cryst. **36**, 125–128 (2003). https://doi.org/10.1107/S0021889802020502
71. R.K. Pathak, M. Baunthiyal, R. Shukla, D. Pandey, G. Taj, A. Kumar, In silico identification of mimicking molecules as defense inducers triggering jasmonic acid mediated immunity against alternaria blight disease in brassica species. Front. Plant Sci. **8**, 609 (2017). https://doi.org/10.3389/fpls.2017.00609
72. M.I. Lobanov, N.S. Bogatyreva, O.V. Galzitskaia, Radius of gyration is indicator of compactness of protein structure. Mol Biol (Mosk) **42**, 701–706 (2008)
73. E. Durham, B. Dorr, N. Woetzel, R. Staritzbichler, J. Meiler, Solvent accessible surface area approximations for rapid and accurate protein structure prediction. J. Mol. Model. **15**, 1093–1108 (2009). https://doi.org/10.1007/s00894-009-0454-9
74. Y. Mazola, O. Guirola, S. Palomares, G. Chinea, C. Menéndez, L. Hernández, A. Musacchio, A comparative molecular dynamics study of thermophilic and mesophilic β-fructosidase enzymes. J. Mol. Model. **21**, 228 (2015). https://doi.org/10.1007/s00894-015-2772-4
75. J.A. Marsh, S.A. Teichmann, Relative solvent accessible surface area predicts protein conformational changes upon binding. Structure **19**, 859–867 (2011). https://doi.org/10.1016/j.str.2011.03.010
76. M.S. Weiss, M. Brandl, J. Sühnel, D. Pal, R. Hilgenfeld, More hydrogen bonds for the (structural) biologist. Trends Biochem. Sci. **26**, 521–523 (2001). https://doi.org/10.1016/s0968-0004(01)01935-1
77. C.N. Pace, H. Fu, K. Lee Fryar, J. Landua, S.R. Trevino, D. Schell, R.L. Thurlkill, S. Imura, J.M. Scholtz, K. Gajiwala, J. Sevcik, L. Urbanikova, J.K. Myers, K. Takano, E.J. Hebert, B.A. Shirley, G.R. Grimsley, Contribution of hydrogen bonds to protein stability. Protein Sci. **23**, 652–661 (2014). https://doi.org/10.1002/pro.2449
78. Y. Gao, Y. Mei, J.Z.H. Zhang, Treatment of hydrogen bonds in protein simulations, in *Advanced Materials for Renewable Hydrogen Production, Storage and Utilization*, ed. by J. Liu, (InTech, Rijeka, 2015). https://doi.org/10.5772/61049
79. R. Shukla, T. Tripathi, Molecular dynamics simulation in drug discovery: opportunities and challenges, in *Innovations and Implementations of Drug Discovery Strategies in Rational Drug Design*, ed. by S.K. Singh, (Springer, Singapore, 2021), pp. 295–316. https://doi.org/10.1007/978-981-15-8936-2_12
80. R. Shukla, T. Tripathi, Molecular dynamics simulation of protein and protein-ligand complexes, in *Computer-Aided Drug Design*, ed. by D.B. Singh, (Springer, Singapore, 2020), pp. 133–161. https://doi.org/10.1007/978-981-15-6815-2_7
81. K. Prince, S. Sasidharan, N. Nag, T. Tripathi, P. Saudagar, Integration of spectroscopic and computational data to analyze protein structure, function, folding, and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 483–520

Molecular Dynamics Simulation to Study Protein Conformation and Ligand Interaction

Santanu Sasidharan, Vijayakumar Gosu, Timir Tripathi, and Prakash Saudagar

Abstract The field of molecular dynamics (MD) simulations has become indispensable today to studying the conformational flexibility and dynamics of proteins as well as protein–ligand complexes. The technique helps to replicate real-time biological events like macromolecular dynamics on a computational platform and allows us to understand the fold and conformational changes in the protein–ligand complex. In addition, MD simulations enable us to estimate the thermodynamics and kinetics associated with protein–ligand binding. In this chapter, we introduce the basics of MD simulations and the theoretical aspects of the simulations. Further, we describe the sequential steps in the process of MD simulation and the background information of the steps. The chapter also discusses ligand binding and conformational changes with the help of case studies. Though the field has advanced by leaps and bounds, there is still a necessity for better force fields and methods to accurately predict the free energy of binding. In summary, research focusing on force fields supported by advancements in computational power will help researchers have better insights into protein–ligand interactions and their conformations.

S. Sasidharan
Department of Biotechnology, National Institute of Technology Warangal, Warangal, Telangana, India

Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada

V. Gosu
Department of Animal Biotechnology, Jeonbuk National University, Jeon Ju, Republic of Korea

T. Tripathi
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

Regional Director's Office, Indira Gandhi National Open University (IGNOU), Regional Centre Kohima, Kohima, India

P. Saudagar (✉)
Department of Biotechnology, National Institute of Technology Warangal, Warangal, Telangana, India
e-mail: ps@nitw.ac.in

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_6

Keywords Protein · Molecular dynamic simulations · Ligand · Dynamics · Conformational flexibility

1 Introduction

Research in biomolecular dynamics has evolved over the last few decades. Among the macromolecules of interest, proteins are essential for the growth and structural integrity of any organism. Their three-dimensional (3D) structure and interactions with other macromolecules or ligands help them function properly. In addition, other molecules interact with proteins either in their active site or allosteric site, which may change the conformations of the protein. The study of these dynamics in detail helps us understand the underlying principles of protein function and interactions [1]. Moreover, advances in bioinformatics and computational power have led researchers to study the structural dynamics of proteins using various simulation algorithms.

Molecular dynamics (MD) simulation is a theoretical method that can analyze the protein structure, folding, and stability by visualizing it in a motion picture. MD simulations have been widely used for studying the complexity of protein folding and the interaction of proteins with ligands. This theoretical study has become an integral part of analyzing the interaction of the ligand with the protein and how the binding of the ligand influences the protein structure, dynamics, and conformation [2, 3]. Besides, it also helps in studying the interactions and changes in terms of energy and geometry over the evolved time period. Today, this method is a boon for protein fold analysis and drug discovery [4].

In principle, MD simulations consider the potential energy function of each of the atoms (force field) and determine the lowest energy state. This means that the state of the most stable conformation, which can be seen over the time period of the simulation run, is determined. Over the past few years, various refinements have been made in the forcefields used for MD simulations. Among the various forcefields used, AMBER, CHARMM, and GROMOS are the most widely used forcefields for studying the structural dynamics of proteins at different pH and temperature conditions [5–7]. These forcefields can be employed in various software like GROMACS, AMBER, and NAMD, and significant information can be obtained from the trajectory analysis [8–10].

To study the effect of the ligand on the protein conformation, one can utilize various techniques like principal component analysis (PCA), coarse-grained simulation, and umbrella sampling. Coarse-grained simulation helps overcome the time-scale and length-scale difference in the ligand–protein interaction by considering the atoms at a macroscopic scale. It does so by reducing the degrees of freedom of the atoms of the protein–ligand complex, providing reduced computational stress, thereby running smoothly. The umbrella sampling depends upon the biasing potential obtained from the mean force potential. It fixes or restrains the ligand toward an increasing center of mass distance via the umbrella sampling. This eventually helps in studying the ligand interaction with the atoms around it over a period of time.

Other than the specific protein–ligand interaction, any perturbation in the protein conformations may result in diseased conditions such as Alzheimer's disease and cancers. Thus, understanding how a protein folds and its dynamics change when interacting with other small molecules and macromolecules is of paramount importance.

2 Background of MD Simulation

We discussed that the MD simulation is a powerful computational method for the theoretical study of biomolecules through fluctuation and conformational changes at the atomic level. The technique uses Newton's second law of motion to calculate the time evolution of the molecular system. The results are obtained in the form of trajectories that are analyzed using different tools for the position and velocity of each atom in the system. In recent times, MD simulation is also being used to understand the thermodynamic properties of biological events like conformational transitions. The technique helps us understand that a protein is flexible and can thus undergo a variety of slow and fast structural rearrangements (also known as transitions), ligand binding, enzymatic regulations, and ion transport in biological systems.

For a protein to function, structural fluctuations and flexibility are very crucial. According to the Levinthal paradox [11, 12], the average time taken would be of the order of 10^{10} years if the process of protein folding was to occur randomly, considering all accessible configurations (around 10^{30} configurations) and a time of 10^{-12} s to search each configuration [4]. The fact that the protein folding process occurs in an immensely shorter time (between picoseconds and milliseconds) proves that the event of protein folding is not a result of a random search toward the correct functional form among the vast configurational space. To explore such configurational spaces, techniques like umbrella sampling have been developed [13].

Before performing MD simulations, it is essential to choose an initial configuration of the proposed system that does not have high potential energy. A velocity must be assigned to the system. To rule out instabilities during simulation, energy minimization is required. Further, a potential energy function (forcefield), which describes the forces that act between the atoms as a function of their positions, is assigned to the system. This gives an initial distribution of the velocities of the atoms and the values of the starting coordinates for the atoms in the system. During the course of the simulation, the trajectories are obtained at different time points and are analyzed. This equilibrium distribution of velocities throughout the system is done via the Maxwell–Boltzmann distribution.

2.1 Theory Behind MD Simulation

It is well known that the MD algorithms calculate the classical time evolution of the system using Newton's second law of motion, i.e.,

$$F_{\mathrm{i}} = m_{\mathrm{i}} a_{\mathrm{i}}$$

$$\frac{F_{\mathrm{i}}}{m_{\mathrm{i}}} = \frac{v_{\mathrm{i}}}{t}$$

$$\frac{F_{\mathrm{i}}}{m_{\mathrm{i}}} = \frac{d^2 x_{\mathrm{i}}}{dt^2}$$

where F_{i} = force exerted on molecule "*i*," m_{i} = mass of the particle, and a_{i} = acceleration of molecule "*i*."

The potential energy of the system can be explained as the sum of the individual contribution of both bonded and non-bonded interactions in the system, i.e.,

$$V(r) = V_{\text{bonded}} + V_{\text{nonbonded}}.$$

Bonded interactions are the sum of four simple harmonic species that describes bond stretching and angle bending. It includes all the parameters responsible for bond stretch and angular bending, including rotational torsion and improper torsion [14].

$$V_{\text{bonded}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{torsion}} + V_{\text{improper}}.$$

V_{bond} represents the energy involved in stretching the bond length in an interaction and can be explained with the help of Morse potential, a robust interatomic interaction model used for the potential energy calculation of a diatomic molecule. Morse potential is computationally expensive and requires three parameters per bond evaluation. Mathematically it can be represented as

$$V_{\text{morse}}(I) = D_{\mathrm{e}}\left[1 - e^{(-a(I - I_0))]^2}\right]$$

$$a = \omega \sqrt{\frac{\mu}{2D_{\mathrm{e}}}}$$

$$\omega = \sqrt{\frac{k}{\mu}}$$

where k = stretching constant, D_{e} = depth of potential minimum, l = bond length, l_0 = equilibrium value of bond length, μ = reduced mass, and ω = frequency of bond vibration in small displacement from l to l_0.

However, to overcome the problem of extensive mathematical calculations, a harmonic potential (Hooke's law) was proposed with an approximation that was adequate for explaining bond stretch energy. According to this method

$$V_{\text{bond}}(I) = \frac{k}{2}(I - I_0)^2$$

$$V_{\text{angle}}(\theta) = \frac{k}{2}(\theta - \theta_0)^2$$

where V_{angle} = energy due to the deviation of angles from their equilibrium values and θ = angle formed between two and three atoms.

V_{torsion} is the torsion angle term in the force field model. It represents the effective barriers for the rotation around chemical bonds. The barriers are due to the steric interactions between the atoms and a group of atoms that are separated by three covalent bonds [15].

$$V_{\text{torsion}}(\varphi) = K_{\varphi}(1 - \cos(n\varphi - \varphi_0)$$

where K_{φ} = barrier height, φ = torsional angle, φ_0 = angular position of the first minimum in the potential, and n = number of minima.

V_{improper} is the improper torsion that arises to maintain chirality.

$$V_{\text{improper}}(\omega) = K\varphi(\omega - \omega_0)^2$$

where ω = improper dihedral angle and ω_0 = improper dihedral angle at equilibrium positions.

The non-bonded interaction is composed of two components, i.e., the van der Waals interaction energy and the electrostatic interaction energy. Energy determination is considered the most time-consuming part of the simulation as they have long-range interactions of the atoms in the system to be considered.

$$V_{(\text{non}-\text{bonded})} = V_{\text{vdw}} + V_{\text{ele}}$$

V_{vdw} arises from a balance between repulsive (short-range and arises due to electron–electron interactions) and attractive forces (long-range force and arises due to electron fluctuations which generate dipole in an atom). It can be demonstrated using Lennard-Jones potential, i.e.,

$$V_{(r)} = 4 \in \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where σ = collision diameter and ϵ = well depth.

V_{ele} act at longer ranges compared to van der Waals interactions. It can be represented as

$$V_{\mathrm{ele}} = \sum\nolimits_i \sum\nolimits_j \frac{q_{\mathrm{i}} q_{\mathrm{j}}}{4\pi \in r_{\mathrm{ij}}}$$

where q_{i}, q_{j} = partial atomic charge, ϵ = dielectric constant, and r_{ij} = relative distance.

Therefore, the potential energy can be represented as

$$V_{(r)} = \sum V_{\mathrm{bond}} + \sum V_{\mathrm{angle}} + \sum V_{\mathrm{torsion}} + \sum V_{\mathrm{vdw}} + \sum V_{\mathrm{ele}}$$

Though the mathematical calculations per atom are higher for a simulation system, the molecular mechanic forcefields provide a reasonable compromise between accuracy and efficiency.

3 Steps in MD Simulation

3.1 Initialization

MD simulations for a biomolecular structure require an initial structure that can be used as a starting point. This structure can be obtained from the X-ray crystallographic or cryo-electron microscopic (cryo-EM) structures available in the protein databank (PDB). Structures from nuclear magnetic resonance (NMR) and homology models can also be used. The selection of the initial structure is critical to obtain better-quality results. Before proceeding toward simulation, energy minimization of the structure is required to eliminate structural distortions that arise due to strong van der Waals interactions and result in unstable simulation. Once the structure is obtained, the next step is to set up the periodic boundary conditions.

3.2 Periodic Boundary Conditions

Defining the periodic boundary is an important step in MD simulation. The step allows one to simulate a small part of a large system specifically. Here, all the atoms present in the computational cell or box (MD cell) are replicated to create an infinite lattice throughout the space. Each particle in the MD cell interacts not only with other particles within the computational box but also with their mirror images in the nearby boxes. Most MD simulations are done in a cubic or octahedral computational cell. The Ewald method is the most common method used for calculating the electrostatic energy of a system on the lattice with periodic boundary conditions. Total electrostatic energy from image cells can be calculated as a summation of real space (V_{r}), reciprocal space (V_{k}), correction due to excluding pairs (U_{e}), and a self-term (U_{s}) [15].

In MD simulation, constant temperature is essential that can be maintained through coupling to a Berendsen thermal bath. Velocities are scaled by a factor at every step.

$$X = \left(1 + \frac{\delta_\mathrm{t}}{\tau}\left(\frac{T}{T_0} - 1\right)\right)$$

where T = the time constant and T_0 is the reference temperature that tells the strength of the coupling between the thermal bath and the system.

3.3 Energy Minimization

Minimization algorithms are employed to identify the geometry of the system that corresponds to the minima of the potential energy surface. The minima values can be very large when a biomolecular system comprises thousands of atoms, and a large number of degrees of freedom is taken. The algorithms used for minimization are important in MD because it is essential to start a simulation from a well-minimized structure that helps avoid any high-energy interactions that might hinder the system.

$$\frac{\partial f}{\partial X_\mathrm{i}} = 0$$

$$\frac{\partial^2 f}{\partial X_\mathrm{i}^2} > 0$$

where I = [1, ..., N], given function f, which depends upon the variable x_1, x_2,. .... .x_n.

The minimum of f is the point at which the first derivative of the function corresponding to each variable is zero, and the second derivatives of the function are all positive. The energy minimization method can be divided into first and second derivative techniques.

3.3.1 First-Derivative Techniques

The first-derivative techniques include the steepest descent and conjugate gradient [16, 17]. The first derivative of energy shows where the local minima lie, and the magnitude gradient indicates the steepness of the local slope. While the second derivative indicates the function's curvature and the information that can be utilized to determine where the function will change with direction. The first-order minimization algorithm is the steepest descent, and in this, the coordinates of the atoms are changed gradually until the system moves close to the minimum energy point. A line search algorithm is used iteratively to locate the minimum point. Even when the starting initial structure is far away from the minimum, the steepest method can achieve the minimum through iterative steps. Because of this advantage, it is recommended to start with the steepest descent algorithm for energy minimization. The conjugate gradient is another first-order minimization algorithm that accumulates information about the function from one iteration to the next [15].

3.3.2 Second-Derivative Techniques

The Newton-Raphson method is a second-order derivative method used to invert the Hessian matrix for energy minimization [18]. The technique provides the curvature of the function that tells about the change in the direction of the function. For large systems, the technique requires higher computational effort and large storage requirements. In most cases, the steepest descent and the Newton-Raphson method are used in combination. However, in the steepest descent method, the structure can be brought to the minimum closely, while in the Newton-Raphson method, a few steps are required to reach the minimum.

3.4 Thermostats and Barostats

Thermostats and barostats are used for equilibrium. The objective of the equilibrium phase is to perform the simulation until the properties like structure, pressure, temperature, and energy are stable with respect to time and to bring the system to equilibrium from the initial configuration. During this phase, each atom of the system is assigned an initial velocity selected from the Maxwell–Boltzmann distribution at a low temperature. Slowly, new velocities are assigned with a gradient increase in the temperature. This process is repeated until the desired temperature is obtained. The equilibration is usually conducted using a Berendsen thermostat and a Parrinello-Rahman barostat [19, 20].

3.5 Production Stage

After the successful completion of the equilibration of the system, the desired MS simulation time length is assigned between picoseconds (ps) and milliseconds (ms). During the production run, no velocity scaling is performed, and hence the temperature becomes a calculated property. Various properties are computed during the production run and are stored for further analysis. During the production run, millions of non-bonded interactions are generated. Thus, it is necessary to evaluate the non-bonded interactions during simulation. One of the easy ways to do so is by extending the time step, which improves the simulation performance. However, we do not consider the bond vibrations during simulation because errors are generated immediately after the production run starts in bond vibration. These errors can be excluded entirely by adding bond constraints using SHAKE algorithms [21].

3.6 Analysis of the MD Data

Simulation information generated after the production run can be analyzed in different ways. One of the most important jobs during the analysis of the ligand–protein complex is to determine whether the apoprotein is stable and close to the experimentally retrieved structure or not. The basic method to check the stability and change in conformation is by calculating the root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (Rg) and hydrogen bond (H-bond), and principal component analysis (PCA) from the simulation data.

The RMSD is used to measure the structural stability of the protein–ligand complex. It provides information about the deviation produced by the complex during the MD simulation compared to the initial reference structure by calculating the Cα values of the protein backbone. Mathematically, it can be represented as

$$\left\lfloor \left(r_{\mathrm{i}}^{\alpha} - r_{\mathrm{i}}^{\beta}\right) \right\rfloor^{1/2} = \sqrt{\frac{1}{N}\sum_{i}\left(r_{\mathrm{i}}^{\alpha} - r_{\mathrm{i}}^{\beta}\right)^{2}}$$

The RMSF is used to measure the local changes that are present along the chain of the protein. It is the measure of the displacement of a particular atom or a group of atoms relative to the initial structure used for the simulation and is averaged over the number of atoms in the structure. The calculation involves a rigid alignment of structure in each frame of the simulation run with respect to the reference frame. It is mathematically represented as

$$\mathrm{RMSF} = \sqrt{\frac{1}{N_{\mathrm{f}}}\sum\nolimits_{f}\left(r_{\mathrm{i}}^{f} - r_{\mathrm{i}}^{\mathrm{avg}}\right)^{2}}$$

The Rg determines the distribution of the atoms present in a protein around the axis of the protein. Rg is given by the length that measures the distance between the point where the atom is rotating and the point where the energy transfers with maximum effect. It is mathematically represented as

$$\mathrm{Rg} = \sqrt{\frac{1}{N_{\mathrm{i}}}\sum\nolimits_{i}\left(r_{\mathrm{i}} - r_{\mathrm{cm}}\right)^{2}}$$

Hydrogen bonds are known to play a vital role in ligand binding. They are important for the effective ligand binding and conformational change in the protein's active site. Mathematically, it is calculated as

$$U_{\mathrm{HB}}(r) = \frac{A}{r^{12}} - \frac{C}{r^{10}}$$

The PCA is a machine learning tool that converts a set of correlated observations to a set of linearly independent components. This transformation to the new coordinate system represents the first coordinate with the highest variance, the second coordinate with the second highest variance, and so on. PCA is used to analyze the motion of flexible regions in the protein. Furthermore, it can also be used to analyze the poorly equilibrated regions in a protein. The calculation of PCA involves the following basic steps

1. Creation of coordinate covariance matrix—It is a 3×3 matrix that consists of the coordinates x, y, and z of the sample at different times.
2. Calculation of principal components and coordinate projections—It gives us the eigenvectors of the matrix.
3. Visualization of the principal components.

The molecular mechanics (MM) energies combined with the Poisson–Boltzmann or generalized Born and surface area continuum solvation (MM/PBSA and MM/GBSA, respectively) methods are used to estimate the ligand binding affinities in the simulation run system. They help in deciding the strength of binding of the ligand to its receptor and studying the stability of the complex. The sample is first simulated over a given period of time. Further, snapshots are taken at regular intervals in time from the simulation to calculate the free energy of the sample. For explicit solvation in water, the free energy is mathematically determined as

$$G = E_{\mathrm{int}} + E_{\mathrm{ele}} + E_{\mathrm{vdw}} + G_{\mathrm{pol}} + G_{\mathrm{np}} - TS$$

where E_{int} = molecular mechanics internal energy, E_{ele} = electrostatic internal energy, E_{vdw} = van del Walls energy, G_{pol} = polar solvation free energy, and G_{np} = nonpolar solvation free energy.

Moreover, the binding free energy between the protein and the ligand is mathematically represented as

$$\Delta G = \langle G(PL) - G(P) - G(L) \rangle_{\mathrm{PL}}$$

where PL, P, and L are protein–ligand complexes, protein, and ligand, respectively, whose free energies are calculated using the equation above. Brackets indicate the average over the snapshot taken. Depending on the protein being analyzed, the r^2 value obtained from the correlation coefficients ranges from 0.0 to 0.9.

4 Ligand Binding and Fold Transitions

The first-ever simulation of a protein was conducted as early as 1944 using a small protein bovine pancreatic trypsin inhibitor [22]. From the simulation, McCammon and his team revealed the fluidic characteristics of a protein interior for the first time. The simulation lasted for 9.2 ps and opened up a new realm in molecular biology and drug discovery. Today, the advancements in computational power allow one to perform even microsecond (μs) simulations at the atomistic level.

To start with any ligand binding, the primary requisite is the availability of the target protein structure. This may be a limitation as the number of experimentally determined structures is still less than the number of proteins existing in nature. The problem can partially be solved by its homologous proteins and by predicting the structure using homology modeling [23]. For proteins that do not exhibit any homology, their structures can be predicted de novo using Robetta, I-TASSER, or AlphaFold [24–26]. Once the structure is solved, the ligands can be docked into the rigid or semirigid target structures. The drawback, however, is that the docking process does not consider the flexibility of the target protein, and any critical fold change that occurs cannot be analyzed. To overcome this, all-atom MD simulations can be employed to obtain conformation ensembles of the target protein, which can be used later for ensemble docking. If there is a computational limitation, coarse grain simulation [27, 28] can be done, and representative conformations can be obtained. The atomistic models can then be converted using tools such as backward.py [29]. Once the structure is confirmed, the next step is to perform ensemble docking, where the ligand is docked against each structure of the conformational ensemble. Performing global docking against a conformational ensemble of the target protein with a large dataset of ligands requires computational power. However, the ligand-binding site can be identified using tools such as fpocket [30] and ConCavity [31]. Once the ligand-binding site is predicted based on the geometry of the ligand and the target protein, one can perform the docking more efficiently.

There are several ways to understand the protein–ligand interactions and fold changes with respect to the binding. The most accurate way is to perform all-atom simulations for the ligand–protein complex. However, there are other docking algorithms supported with CHARMM forcefield that employ multiple strategies to obtain better protein–ligand interactions, such as CDOCKER [32], EADock [33], etc. Even though the methods use forcefields, they fail to account for the entropy changes, and therefore the accuracy of the final results in the docking is compromised [34]. Long timescale MD simulations are an easy and effective way to sample the protein–ligand interactions. Long timescale simulations allow determining not only the interactions but also the fold changes that occur in the target protein due to ligand binding. These simulation results enable direct comparison to the experimental results and serve as a benchmark for ligand-binding studies. However, as discussed earlier, they are expensive, and most research groups cannot access them. However, coarse-grained model simulation can work around this problem. The method maps several heavy atoms into one site, reducing the total number of particles, thereby, the computational power. One can refer to Souza et al. [35] for a better understanding of the concept. The coarse grain simulations fail to provide the desired accuracy in the ligand binding, even if the back mapping is performed. Therefore, to achieve high accuracy and better binding free energy results, it is recommended to run a long timescale simulation of the ligand–protein complex.

Some simulations require the sampling of the conformational space to obtain statistically reliable results. Though this is not reliant on the resolution of the model, it is critical since there are systems with two states with high-energy crossover. To overcome this issue, there are various enhanced sampling methods such as Ligand Gaussian accelerated molecular dynamics simulations [36], metadynamics [37], Markov state models [38, 39], and replica-exchange molecular dynamics [40].

5 Case Studies

In this section, we will discuss the docking and simulation of ligand–protein, how the pipeline works, and the analysis of the simulation results. For this, we will use a study by Sasidharan et al. where natural compounds were virtually screened against the tyrosine aminotransferase (TAT) from *Leishmania donovani* [41]. The initial part of the study concentrated on the virtual screening of 1,83,659 compounds from the ZINC15 database with the protein. The top 10 compounds were then docked independently against TAT using Autodock v4.2. For the docking, authors framed the grid around the active site of the TAT enzyme housing the K286 residue, and 500 LGA docks were conducted to obtain the best-docked conformation. The top 5 compounds with the highest binding affinity and interactions (Fig. 1) with the active site cavities were chosen and carried forward for simulations.

The simulations were carried out using GROMACS v5.1.4. The complexes were energy minimized by steeped descent method and were temperature and pressure equilibrated using a modified Berendsen thermostat and Parrinello–Rahman barostat, respectively. The electrostatic interactions were computed with the help of particle mesh Ewald. The trajectory analysis showed that all complexes with the protein were stable throughout the simulation period. The RMSD of the Cα backbone (Fig. 2a) showed the stability of the complexes, while the Rg (Fig. 2c), along with solvent accessible surface area (Fig. 2d), corroborated the stability of the complexes. The RMSF analysis showed higher fluctuations in the N-terminal (Fig. 2b), and the reason for the same was deciphered by the authors in another study [42]. The study then concentrated on the binding of the ligands to the TAT. An average of 1–3 hydrogen bonds formed between the compounds and the protein (Fig. 3a). The authors eliminated the compound TI 2 from further studies owing to the presence of several metastable states (Fig. 3b). The simulation data showed that the compounds TI 1, TI 3, TI 4, and TI 5 could bind to TAT with high affinity. The authors proved the inhibitory activity of the compounds by in vitro inhibition kinetics.

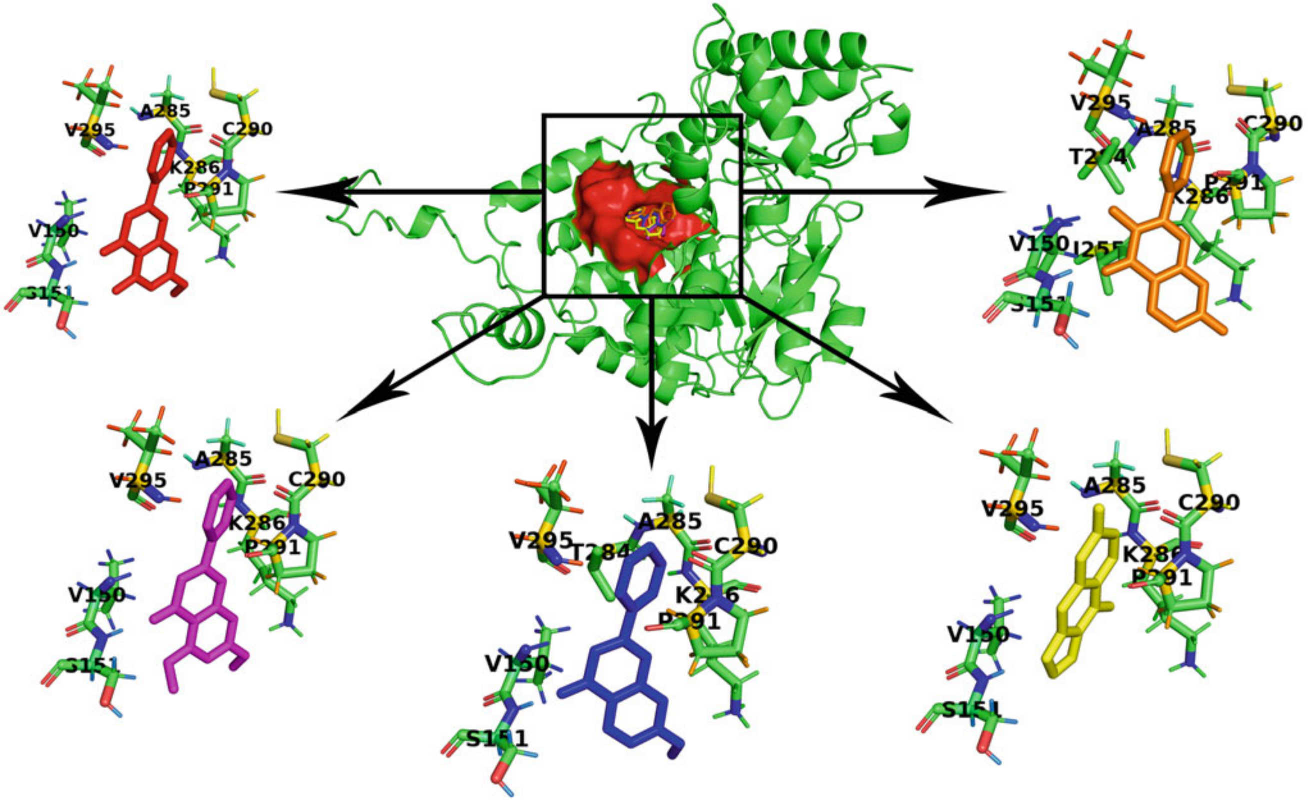

Fig. 1 Docking of the five compounds in the TAT active site. Compounds TI 1 (in red), TI 2 (in magenta), TI 3 (in blue), TI 4 (in yellow), and TI 5 (in orange) docked to the active site cavity (represented as red surface). (Figure adapted with permission from Sasidharan and Saudagar [41])

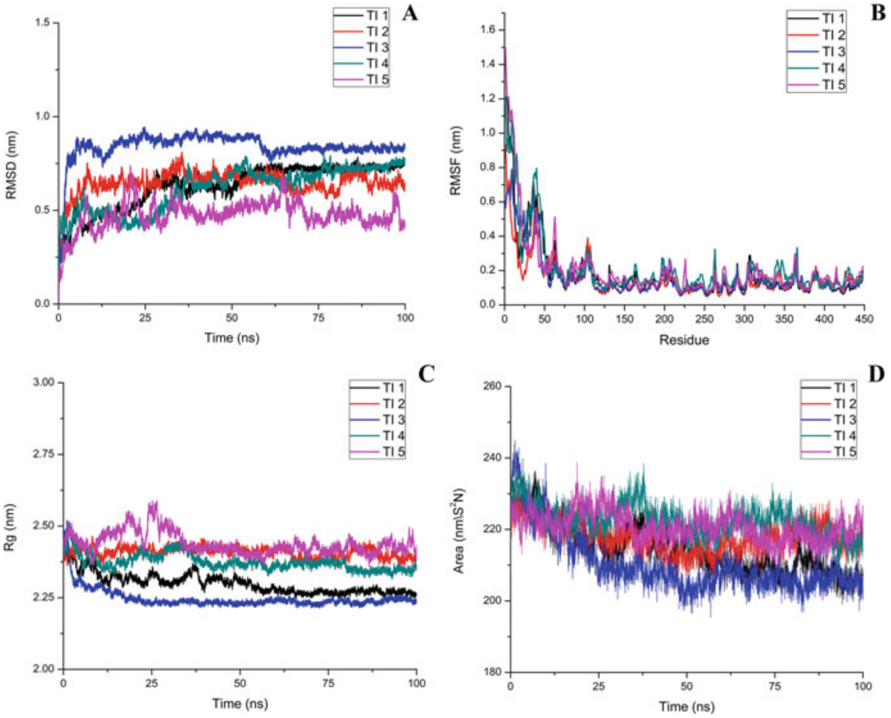

Fig. 2 Trajectory analysis of compounds TI 1–5 complexes with the TAT enzyme. (**a**) RMSD, (**b**) RMSF, (**c**) Rg, (**d**) SAS. The analysis of all four trajectories shows the stability of the TI complexes. (Figure adapted with permission from Sasidharan and Saudagar [41])

To understand the concept of PCA and MMPBSA, we use a study by Shweta et al. [43]. Here, the authors followed a similar protocol and simulated the top two protein–compound complexes. Besides RMSD, Rg, SASA, and RMSF, the authors also conducted the PCA (Fig. 4), which showed that the ligand-bound forms are more rigid than the apo-form. The changes in the large motions were limited upon binding to the ligands chrysin and genistein. Furthermore, MMPBSA calculations showed binding energies of −78 kJ/mol and −76 kJ/mol for genistein and chrysin, respectively, which were higher than the control ATP (−54 kJ/mol). The breakdown of the binding energy is given in Table 1. The binding energy contributed by each residue in the target protein using the MMPBSA tool can also be studied [44]. Several such studies can be referred to understand the protein–ligand binding analysis using MD simulations [45–50].

The studies of protein–ligand interactions and transitions are not limited to small compounds but also protein–macromolecule interactions. Gosu et al. studied the effect of mutations on the MDA5 protein responsible for Aicardi-Goutières syndrome and Singleton-Merten syndrome [51]. The effect of mutation of residues like L372F, A45T, R779H, and R822Q was studied, and the interactions of the mutated proteins with RNA were analyzed. The authors represented the PCA of the simulated

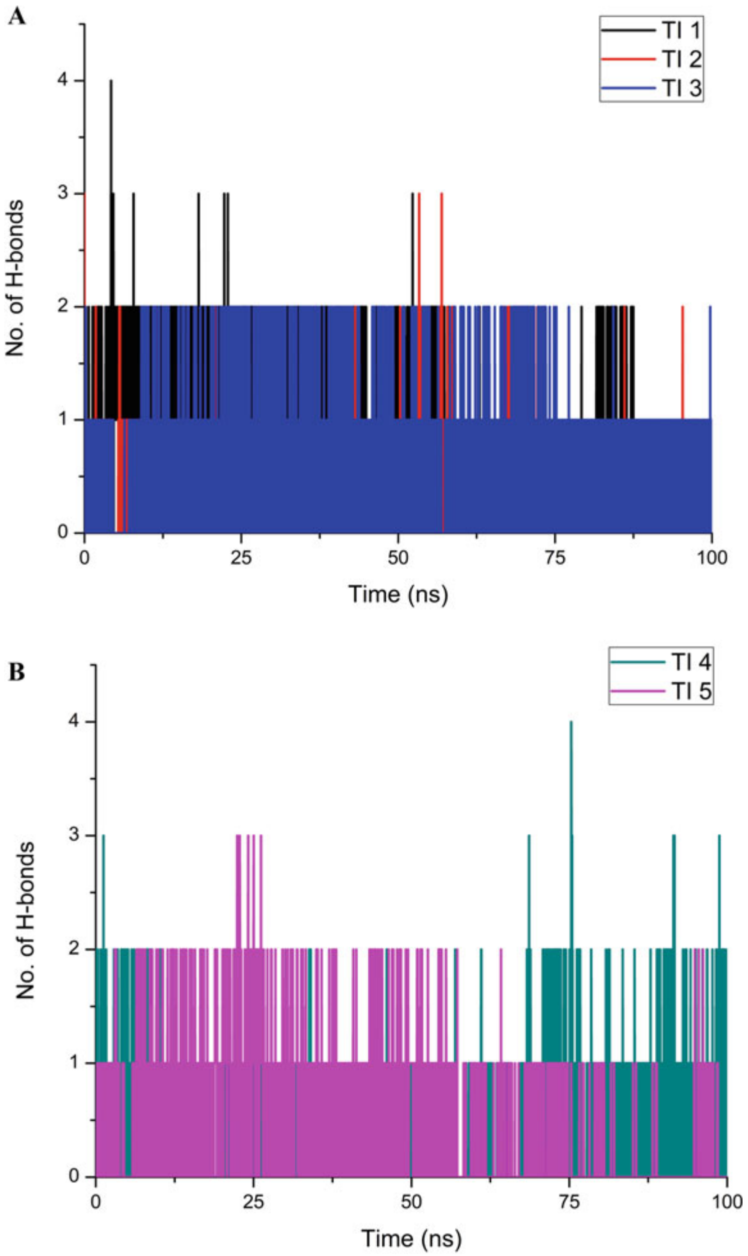

Fig. 3 Hydrogen bond analysis between TI 1–5 compounds and TAT enzyme. (**a**) H-bond analysis of TI 1–3 with TAT. (**b**) H-bond analysis of TI 4 and 5 with TAT. (Figure adapted with permission from Sasidharan and Saudagar [41])

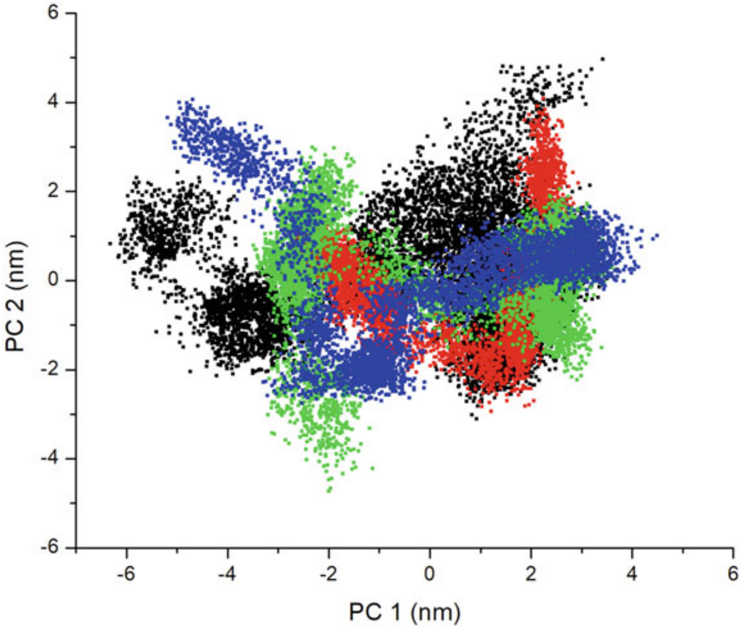

Fig. 4 Principal component analysis of large motions in the simulated structures. Apo-LdMPK4, ATP complex, GEN complex, and CHY complex with LdMPK4 are shown in black, red, green, and blue, respectively. (Figure adapted with permission from Shweta et al. [43])

Table 1 Binding free energy of MAPK4 with ligands ATP, GEN, and CHY

	Ligand		
Component	ATP	GEN	CHY
E_{vdW} (kJ/Mol)	−103.47	−104.923	−105.592
E_{elec} (kJ/Mol)	−3.588	−10.134	−11.382
G_{polar} (kJ/Mol)	60.086	46.500	49.818
$G_{\mathrm{non\text{-}polar}}$ (kJ/Mol)	−7.973	−9.653	−9.007
ΔG_{bind} (kJ/Mol)	−54.946	−78.211	−76.164

complexes in porcupine plots (Fig. 5) that revealed the effect of mutations on the large-scale motions of the whole protein as well as the fold changes occurring over the simulation period due to mutations. Hence, MD simulations can be used to study both protein–ligand and protein–macromolecule interactions for both drug discovery and mutation effects [52–57].

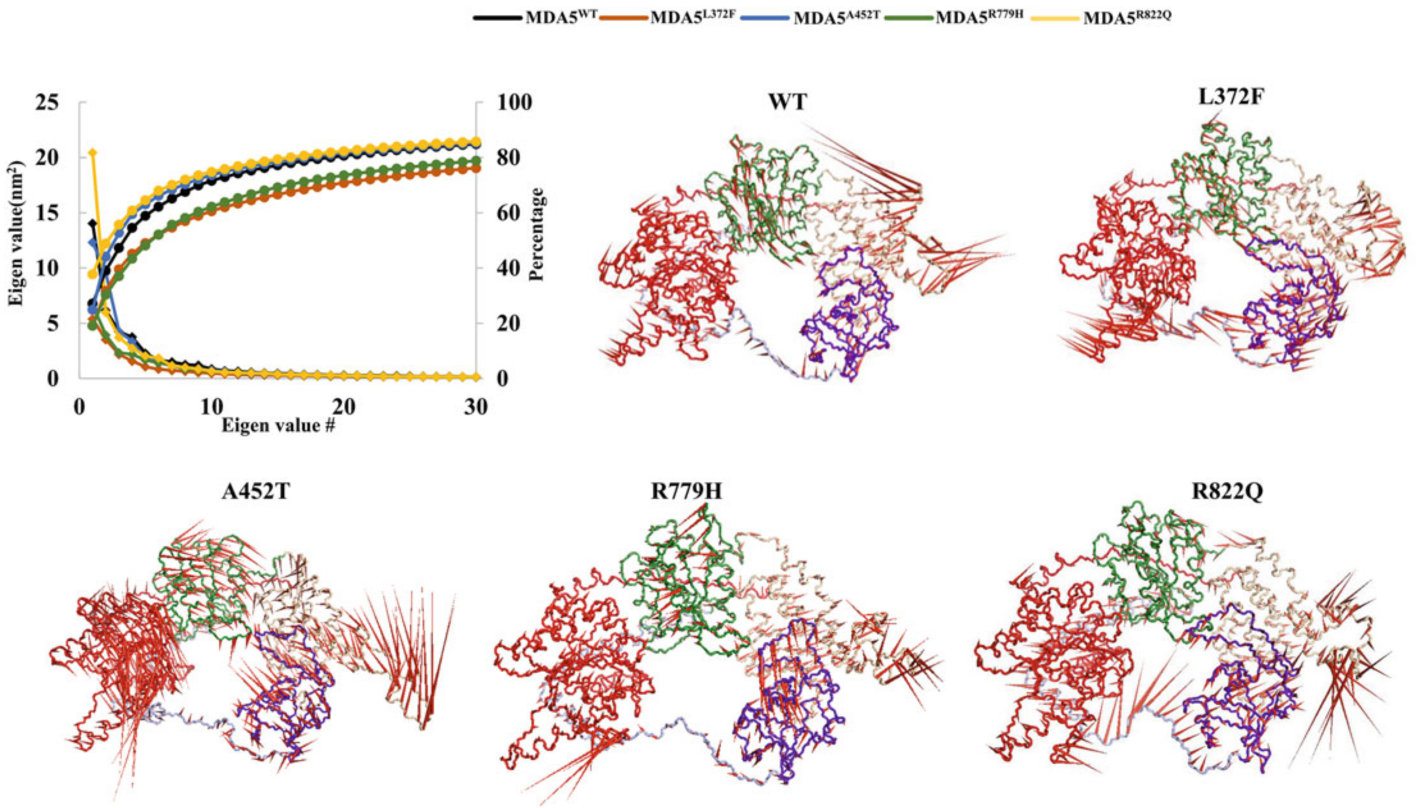

Fig. 5 Cumulative percentages and porcupine plots representing the principal component analysis. The porcupine plots were made using PyMol. (Figure adapted with permission from Gosu et al. [51])

6 Conclusions

Currently, the field of simulation is making an enormous impact in understanding the atomistic details of macromolecules. MD simulations now help researchers to drive the wet-lab experiments based on the simulation data. A detailed conformational understanding of a macromolecule explains the dynamics of the ligand binding as well as the fold transitions that accompany the ligand binding. The accuracy and free energy calculations are more accurate than the docking scores and, therefore, can be relied upon. Though the dynamics of proteins happen at msec timescale in real time, it is not possible, at least at the moment, to simulate all the proteins to that extent. Meanwhile, it is challenging to understand the dynamics using wet-lab experiments. Therefore, researchers must balance these techniques and consider the trade-off to achieve the best possible results. This research area is advancing day by day, with improvements in algorithms and forcefields. With increased computational power and refined algorithms, scientists hope to make simulations widely available at lower costs.

References

1. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Springer Nature, Singapore, 2020)
2. P. Saudagar, T. Tripathi, *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, 1st edn. (Academic Press, San Diego, 2023)
3. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods*, 1st edn. (Academic Press, Cambridge, MA, 2022)
4. S. Sasidharan, P. Saudagar, Prediction, validation, and analysis of protein structures: a beginner's guide, in *Advances in Protein Molecular and Structural Biology Methods*, ed. by T. Tripathi, V.K. Dubey, (Academic Press, Cambridge, MA, 2022), pp. 373–385
5. J.W. Ponder, D.A. Case, Force fields for protein simulations. Adv. Protein Chem. **66**, 27–85 (2003)
6. J. Lee, M. Hitzenberger, M. Rieger, N.R. Kern, M. Zacharias, W. Im, CHARMM-GUI supports the Amber force fields. J. Chem. Phys. **153**(3), 035103 (2020)
7. W.F. van Gunsteren, X. Daura, A.E. Mark, GROMOS force field, in *Encyclopedia of Computational Chemistry*, vol. 2, (Wiley, Chichester, 2002)
8. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J. Berendsen, GROMACS: fast, flexible, and free. J. Comput. Chem. **26**(16), 1701–1718 (2005)
9. D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, R.J. Woods, The Amber biomolecular simulation programs. J. Comput. Chem. **26**(16), 1668–1688 (2005)
10. J.C. Phillips, D.J. Hardy, J.D. Maia, J.E. Stone, J.V. Ribeiro, R.C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. **153**(4), 044130 (2020)
11. C. Levinthal, Are there pathways for protein folding? J. Chim. Phys. **65**, 44–45 (1968)
12. C. Levinthal, How to fold graciously. Mossbauer Spectrosc. Biol. Syst. **67**, 22–24 (1969)
13. R. Shukla, T. Tripathi, Molecular dynamics simulation in drug discovery: opportunities and challenges, in *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*, (Springer, Singapore, 2021), pp. 295–316

14. C.-E.A. Chang, Y.-M.M. Huang, L.J. Mueller, W. You, Investigation of structural dynamics of enzymes and protonation states of substrates using computational tools. Catalysts **6**(6), 82 (2016)
15. A. Kukol, *Molecular Modeling of Proteins* (Springer, New York, 2008)
16. P. Debye, Näherungsformeln für die Zylinderfunktionen für große Werte des Arguments und unbeschränkt veränderliche Werte des Index. Math. Ann. **67**(4), 535–558 (1909)
17. E. Stiefel, Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bureau Standards **49**, 409–435 (1952)
18. J. Dedieu, Newton-Raphson method, in *Encyclopedia of Applied and Computational Mathematics*, ed. by B. Engquist, (Springer, Berlin, 2015), pp. 1023–1028
19. M. Parrinello, A. Rahman, Strain fluctuations and elastic constants. J. Chem. Phys. **76**(5), 2662–2666 (1982)
20. H.J. Berendsen, J.V. Postma, W.F. Van Gunsteren, A. DiNola, J.R. Haak, Molecular dynamics with coupling to an external bath. J. Chem. Phys. **81**(8), 3684–3690 (1984)
21. J.-P. Ryckaert, G. Ciccotti, H.J. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys. **23**(3), 327–341 (1977)
22. J.A. McCammon, B.R. Gelin, M. Karplus, Dynamics of folded proteins. Nature **267**(5612), 585–590 (1977)
23. N. Eswar, B. Webb, M.A. Marti-Renom, M. Madhusudhan, D. Eramian, M.Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using modeller. Curr. Protoc. Bioinformatics **15**(1), 5.6.1–5.6.30 (2006)
24. D.E. Kim, D. Chivian, D. Baker, Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. **32**(Suppl_2), W526–W531 (2004)
25. J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER suite: protein structure and function prediction. Nat. Methods **12**(1), 7–8 (2015)
26. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, Highly accurate protein structure prediction with AlphaFold. Nature **596**(7873), 583–589 (2021)
27. S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, A.H. De Vries, The MARTINI force field: coarse grained model for biomolecular simulations. J. Phys. Chem. B **111**(27), 7812–7824 (2007)
28. D.H. de Jong, G. Singh, W.D. Bennett, C. Arnarez, T.A. Wassenaar, L.V. Schafer, X. Periole, D.P. Tieleman, S.J. Marrink, Improved parameters for the martini coarse-grained protein force field. J. Chem. Theory Comput. **9**(1), 687–697 (2013)
29. T.A. Wassenaar, K. Pluhackova, R.A. Böckmann, S.J. Marrink, D.P. Tieleman, Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. J. Chem. Theory Comput. **10**(2), 676–690 (2014)
30. P. Schmidtke, V. Le Guilloux, J. Maupetit, P. Tufféry, Fpocket: online tools for protein ensemble pocket detection and tracking. Nucleic Acids Res. **38**(Suppl_2), W582–W589 (2010)
31. J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comp. Biol. **5**(12), e1000585 (2009)
32. G. Wu, D.H. Robertson, C.L. Brooks III, M. Vieth, Detailed analysis of grid-based molecular docking: a case study of CDOCKER—a CHARMm-based MD docking algorithm. J. Comput. Chem. **24**(13), 1549–1562 (2003)
33. A. Grosdidier, V. Zoete, O. Michielin, EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. Proteins **67**(4), 1010–1025 (2007)
34. S. Yin, L. Biedermannova, J. Vondrasek, N.V. Dokholyan, MedusaScore: an accurate force field-based scoring function for virtual drug screening. J. Chem. Inf. Model. **48**(8), 1656–1662 (2008)

35. P.C. Souza, S. Thallmair, P. Conflitti, C. Ramírez-Palacios, R. Alessandri, S. Raniolo, V. Limongelli, S.J. Marrink, Protein–ligand binding with the coarse-grained martini model. Nat. Commun. **11**(1), 1–11 (2020)
36. Y. Miao, A. Bhattarai, J. Wang, Ligand Gaussian accelerated molecular dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics. J. Chem. Theory Comput. **16**(9), 5526–5547 (2020)
37. A. Barducci, M. Bonomi, M. Parrinello, Metadynamics. Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**(5), 826–843 (2011)
38. W. Wang, S. Cao, L. Zhu, X. Huang, Constructing Markov state models to elucidate the functional conformational changes of complex biomolecules. Wiley Interdiscip. Rev. Comput. Mol. Sci. **8**(1), e1343 (2018)
39. B.E. Husic, V.S. Pande, Markov state models: from an art to a science. J. Am. Chem. Soc. **140**(7), 2386–2396 (2018)
40. L.S. Stelzl, G. Hummer, Kinetics from replica exchange molecular dynamics simulations. J. Chem. Theory Comput. **13**(8), 3927–3935 (2017)
41. S. Sasidharan, P. Saudagar, Flavones reversibly inhibit Leishmania donovani tyrosine aminotransferase by binding to the catalytic pocket: an integrated in silico-in vitro approach. Int. J. Biol. Macromol. **164**, 2987–3004 (2020)
42. S. Sasidharan, P. Saudagar, Mapping N-and C-terminals of Leishmania donovani tyrosine aminotransferase by gene truncation strategy: a functional study using in vitro and in silico approaches. Sci. Rep. **10**(1), 1–15 (2020)
43. S. Raj, S. Sasidharan, V.K. Dubey, P. Saudagar, Identification of lead molecules against potential drug target protein MAPK4 from L. Donovani: an in-silico approach using docking, molecular dynamics and binding free energy calculation. PLoS One **14**(8), e0221331 (2019)
44. R. Kumari, R. Kumar, A. Lynn, g_mmpbsa—a GROMACS tool for high-throughput MM-PBSA calculations. J. Chem. Inf. Model. **54**(7), 1951–1962 (2014)
45. R. Shukla, P.B. Chetri, A. Sonkar, M.Y. Pakharukova, V.A. Mordvinov, T. Tripathi, Identification of novel natural inhibitors of opisthorchis felineus cytochrome P450 using structure-based screening and molecular dynamic simulation. J. Biomol. Struct. Dyn. **36**(13), 3541–3556 (2018)
46. R. Shukla, H. Shukla, P. Kalita, A. Sonkar, T. Pandey, D.B. Singh, A. Kumar, T. Tripathi, Identification of potential inhibitors of Fasciola gigantica thioredoxin1: computational screening, molecular dynamics simulation, and binding free energy studies. J. Biomol. Struct. Dyn. **36**(8), 2147–2162 (2018)
47. R. Shukla, H. Shukla, P. Kalita, T. Tripathi, Structural insights into natural compounds as inhibitors of Fasciola gigantica thioredoxin glutathione reductase. J. Cell. Biochem. **119**(4), 3067–3080 (2018)
48. R. Shukla, H. Shukla, A. Sonkar, T. Pandey, T. Tripathi, Structure-based screening and molecular dynamics simulations offer novel natural compounds as potential inhibitors of mycobacterium tuberculosis isocitrate lyase. J. Biomol. Struct. Dyn. **36**(8), 2045–2057 (2018)
49. R. Shukla, H. Shukla, T. Tripathi, Structural and energetic understanding of novel natural inhibitors of mycobacterium tuberculosis malate synthase. J. Cell. Biochem. **120**(2), 2469–2482 (2019)
50. R. Shukla, H. Shukla, T. Tripathi, Structure-based discovery of phenyl-diketo acids derivatives as mycobacterium tuberculosis malate synthase inhibitors. J. Biomol. Struct. Dyn. **39**(8), 2945–2958 (2021)
51. V. Gosu, S. Sasidharan, P. Saudagar, H.-K. Lee, D. Shin, Computational insights into the structural dynamics of MDA5 variants associated with Aicardi–Goutières syndrome and Singleton–Merten syndrome. Biomol. Ther. **11**(8), 1251 (2021)
52. R. Shukla, T. Tripathi, Molecular dynamics simulation of protein and protein-ligand complexes, in *Computer-Aided Drug Design*, ed. by D.B. Singh, (Springer Nature, Singapore, 2020), pp. 133–161

53. K. Prince, S. Sasidharan, N. Nag, T. Tripathi, P. Saudagar, Integration of spectroscopic and computational data to analyze protein structure, function, folding, and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 483–502
54. J. Kalita, H. Shukla, T. Tripathi, Engineering glutathione S-transferase with a point mutation at conserved F136 residue increases the xenobiotic-metabolizing activity. Int. J. Biol. Macromol. **163**, 1117–1126 (2020)
55. P. Kalita, H. Shukla, K.C. Das, T. Tripathi, Conserved Arg451 residue is critical for maintaining the stability and activity of thioredoxin glutathione reductase. Arch. Biochem. Biophys. **674**, 108098 (2019)
56. R. Shukla, H. Shukla, T. Tripathi, Activity loss by H46A mutation in mycobacterium tuberculosis isocitrate lyase is due to decrease in structural plasticity and collective motions of the active site. Tuberculosis (Edinb.) **108**, 143–150 (2018)
57. A. Sonkar, D.L. Lyngdoh, R. Shukla, H. Shukla, T. Tripathi, S. Ahmed, Point mutation A394E in the central intrinsic disordered region of Rna14 leads to chromosomal instability in fission yeast. Int. J. Biol. Macromol. **119**, 785–791 (2018)

Monte Carlo Approaches to Study Protein Conformation Ensembles

Nidhi Awasthi, Rohit Shukla, Devesh Kumar, Arvind Kumar Tiwari, and Timir Tripathi

Abstract The molecular dynamics (MD) simulation method is widely used to determine the protein folding sampling by applying various force fields. The results obtained by these methods give sufficiently good accuracy but are time consuming. The Monte Carlo (MC) algorithm is an efficient method to provide the protein sampling results within a short time period, as it calculates the average of ensembles. This chapter discusses the MC simulations as a promising approach for revealing protein folding dynamics and conformations. These methods offer reliable results as it employs statistical approaches. Using these methods, the thermodynamic properties can be investigated by averaging the ensembles of a protein. Further, to understand the reliability of results obtained by MC simulations, a case study of the properties of Trp-cage protein is also discussed. One MC integration step of Trp-cage protein was found to reveal the excellent sampling of MC. The simplicity and efficiency of the MC method enable studying involving larger systems of proteins and polypeptides.

Keywords Protein folding · Conformation · Dynamics · Monte Carlo simulations · Molecular dynamics · Statistical approaches

1 Introduction

The mechanism of conformational changes plays an essential role in the function and regulation of proteins [1–3]. To analyse protein function, it is necessary to understand the stability of the whole protein as well as its individual conformational

N. Awasthi · A. K. Tiwari
Department of Physics, B.S.N.V.P.G. College, Lucknow, Uttar Pradesh, India

R. Shukla · T. Tripathi (✉)
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

D. Kumar (✉)
Babasaheb Bhimrao Ambedkar University, Lucknow, India

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_7

sub-ensembles. Experimentally, it is challenging to directly observe the conformational changes at the single molecule level due to its dynamic nature at the microscopic level [4, 5]. However, computational methods have emerged as more convenient tools for these studies by involving timescale steps [6–9]. Generally, simulation methods are used for understanding the mechanism of protein folding, which provides detailed information on folding like-intermediate complexes, barrier heights, etc.

In the last two decades, a large number of studies have been performed that focused on understanding the mechanistic properties of folding of small peptides and proteins using certain force fields in simulation methods [10–16]. Various well-known simulation methods exist, but the molecular dynamics (MD) method has been the most explored. This method has been limited by some admissible timesteps to track the atomistic changes, which are very small [17], while natural processes have larger inherent timescales. These timescales range from microseconds to seconds [18]. Hence, for capturing even a single folding step of a protein, it is essential to perform large numbers of MD simulation steps, which require high computational costs or supercomputers [19]. A specialized supercomputer can easily observe many fold transitions of small or large fast-folding proteins [20, 21] by using biophysical force fields like AMBER and CHARMM in certain solvent models [22, 23]. If supercomputers are not available, then multiple strategies can be used to perform the calculations using various force fields to improve the simulation protocol [24–26]. Amongst the possible simulating methods, MD simulations with solvent model enable to perform faster conformational sampling of proteins [27]. Although MD simulations have some limitations, they can provide accurate results for approximately 100 protein residues [28–30]. The numbers of residues depend on the complexity of the protein and the quality of the solvent [31, 32]. Apart from these methods, other techniques, like enhanced sampling methods, give more accurate results even at longer timescales [33]. Unfortunately, none of these methods can provide a straightforward analysis of the protein folding mechanism at larger timescales as it occurs in nature [34].

There is an alternate method known as Monte Carlo (MC) simulation, which essentially extends to explore the simulation approach as it has no inherent timescale. MC simulation provides all thermodynamic data, which can be reconstructed and provides kinetic information at a large timescale [35]. These thermodynamic data are usually extracted directly from the MD simulations [36]. MC simulation is based on some special moves (conformational changes), which can be designed not to follow the local force field so that in each step, the conformational change per energy evaluation may be larger for simulation. Hence the simulation focuses on a few specialized degrees of freedom of protein, such as the dihedral angle of proteins. These advanced features of the MC simulation methods are able to accelerate the molecular simulations of proteins and peptides by providing a suitable forcefield to the system. Generally, the calculation of multiparticle moves for a large number of molecules becomes complex and expensive, so the MC algorithm calculates only a small part of the protein and peptide system in a single move. Hence, implicit solvent models are more suited than explicit solvent models for MC simulations

[37, 38]. MC simulation calculations with an implicit solvent model can increase the simulation speed and give more accurate results in a short time. However, it has some limitations, such as hydrogen-bond representation, over-stabilized salt bridges, incorrect distribution of ions, temperature independency of free energy, etc. [39, 40].

Since the 1990s, various program packages have been developed to understand the folding and unfolding behaviour of proteins and peptides by applying MC folding algorithms. The most popular model is the Rosetta model, which uses an all-atom as well as coarse-grained representation of proteins [41, 42]. This model implements a knowledge-guided MC sampling approach using various energy functions. Its result heavily relies on the data of experimental structures. So, for users, it is not compatible with other biochemical data obtained from large and complex protein structures. Usually, it is tough to define the topology of a structure that has never or rarely been observed in the protein data bank (PDB) [43], MC-based C++ code, and SMMP [44], a FORTRAN, which are known for simulations of proteins. These are computationally very fast methods and have an excellent ability to capture the structural and thermodynamic properties of a set of sequences. Similarly, coarse-grained models like CABS (C-alpha, beta and side-chain) [45] for protein folding [46, 47] have been successfully applied to study the binding studies of intrinsically disordered proteins (IDPs) [48]. Due to large-scale dynamics, these models provide significant structural transitions and good conformational results with sufficiently good accuracy [42]. All known multiscale modelling techniques, like all-atom/coarse-grained, are used for the conformation and folding of proteins and peptides [49, 50]. MC methods are also used to predict missing protein chains or fragments [51]. However, MC simulations have some limitations as it is limited to employing specifically designed force fields and algorithms. These methods may impact their common usage.

2 Monte Carlo Simulations

In MC simulations, many conformations of ensembles are generated for solving the complex macromolecular system under a specific thermodynamics condition [52]. These configurations can be generated by applying some perturbations. These applied perturbations are extensively large, feasible and have a sufficiently high probability. MC simulation provides an ensemble of representative conformational changes rather than information on time evolution. MC simulation also plays a vital role in designing new algorithms of MD simulations for complex and hybrid protein structures [53].

The following sub-sections are dedicated to the underlying principles of MC simulation algorithms. In Sect. 2.1, we review some important notions about Lagrangian and Hamiltonian dynamics, which are common for both MC and MD simulations. In Sect. 2.2, we introduce the partition function and the probability density function, as well as the calculation of thermodynamics observables associated with a macromolecule, such as the hemagglutinin or the neuraminidase. The partition function is instrumental in computing such observables. In Sect. 2.3, we explain how to efficiently sample the representative space. For this, we discuss the approaches of emission probability, transition probability, acceptance probability, and detailed balance theory. Sampling is useful only when performed in realistic experimental conditions. Hence, we explain how to sample in a canonical ensemble (with a constant number of particles, volume, and temperature) and in an isothermal-isobaric ensemble (with a constant number of particles, pressure, and volume) in Sects. 2.4 and 2.5, respectively. Finally, in Sect. 2.6, we address the sampling problem in the presence of numerous minima. This is a significant problem, particularly when studying *Haemophilus influenzae* proteins, such as hemagglutinin and neuraminidase.

2.1 Lagrangian and Hamiltonian Dynamics (or How to Formulate the Problem)

Lagrangian and Hamiltonian dynamics are applicable to both MD and MC simulations. These dynamics are functions of some generalized coordinates, which provide a simple framework for understanding complex proteins or peptides [54]. The Lagrangian of any system gives the difference between kinetic energy and potential energy of that system, as shown in Eq. (1), while Hamiltonian define the sum of potential energy and kinetic energy of the system, as shown in Eq. (2).

$$L\Big(q, q\dot{)} = K(\dot{q}) - U(q) \tag{1}$$

$$H\Big(q, q\dot{)} = K(\dot{q}) + U(q) \tag{2}$$

where $K(\dot{q})$ is the kinetic energy and $U(q)$ is the potential energy of the atomic system. The $(\dot{q})$ and (q) are generalized velocities and generalized coordinates. The generalized momentum is denoted by Eq. (3):

$$p = \frac{\partial L}{\partial \dot{q}} \tag{3}$$

The momentum (p) is the function of Hamiltonian, which is obtained by the Legendre transformation of the Lagrangian, defined in Eq. (4):

$$H(q, p) = \sum_{n=1}^{3N} p^n\, q_{\mathrm{n}} - L\left(q^{3N}, \dot{q}^{3N}\right) \tag{4}$$

The Hamiltonian is given by Hamilton's equations, as shown in Eq. (5):

$$
\begin{aligned}
\frac{\partial H}{\partial t} &= 0, \\
\dot{q}_{\mathrm{n}} &= \frac{\partial H}{\partial p^n}, \\
\dot{p^n} &= -\frac{\partial H}{\partial q_{\mathrm{n}}}
\end{aligned}
\tag{5}
$$

where q_{n}, p^n are generalized position and generalized momentum.

2.2 Partition Functions, Probability Density Functions, and Expectation (or How to Compute Observables)

The partition function of a system is used to determine the microstates associated with a macrostate and other thermodynamical properties such as free energies, enthalpy, ensemble average, and probability of occurrence of specific conformation. Hence, from the partition function, one can determine many other thermodynamical parameters. Due to these properties, partition functions are widely used in all MC simulations. The number of microstates is given by Eq. (6):

$$
\Omega(N, V, E) = E_0 C_{\{\mathrm{N}\}} \int dp^N dr^N \delta\left(H\left(r^N, r^P\right) - E\right) \tag{6}
$$

where

$$
C_{\{\mathrm{N}\}} = \frac{1}{h^{3N}(N_{\mathrm{A}}!N_{\mathrm{B}}!\ldots)} \tag{7}
$$

where $C_{\{\mathrm{N}\}}$ is a well-defined quantum factor, which accounts for the indiscernibility of the various atoms A, B, and so on, ⍰ is the Planck constant, and $\delta(x)$ is the Dirac delta function. The number of states of constant energy E of any atomic system is counted by the function $\Omega(N, V, E)$, which is directly related to the entropy as defined in Eq. (8):

$$
S(N, E) = k_{\mathrm{B}} \ln \Omega\ (N, V, E) \tag{8}
$$

where k_{B} is the Boltzmann constant. The partition function is defined in Eq. (9):

$$
Z(N, V, T) = \int dp^N dr^N\ \exp\left[-\beta H\left(r^N, p^N\right)\right] \tag{9}
$$

where

$$\beta = \frac{1}{k_{\mathrm{B}}T} \tag{10}$$

The canonical partition function is a function determined by the Hamiltonian of the corresponding macromolecular system. The probability of having the macromolecular system in any state is given by Eq. (11):

$$P\left(r^N, p^N\right) dp^N dr^N = \frac{\exp[-\beta H(r^N, p^N)] dp^N dr^N}{Z(N, V, E)}. \tag{11}$$

Therefore, the average of any observable O is obtained by

$$\langle O \rangle = \int dp^N dr^N O\left(r^N, p^N\right) P\left(r^N, p^N\right) \tag{12}$$

So, the standard deviation associated with the observable is given by

$$\sigma(O) = \sqrt{\left\langle O^2 \right\rangle - \langle O \rangle^2} \tag{13}$$

But, due to large number of degrees of freedom, it is impossible to integrate the partition function easily. It would be integrated into a multidimensional space.

2.3 How to Sample Efficiently Thermodynamical Quantities

The Monte Carlo integration method is the most popular method to solve multidimensional integration of partition function and probability function, using the equation:

$$P(i)T(i \rightarrow f)A(i \rightarrow f) = P(f)T(f \rightarrow i)A(f \rightarrow i) \tag{14}$$

where $P(i)$ is the emission probability that the system is in the initial state, $T(i \rightarrow f)$ is the transition probability from the state $i \rightarrow f$, and $A(i \rightarrow f)$ is the acceptance probability of such a transition state

$$T(i \rightarrow f) = T(f \rightarrow i) \tag{15}$$

The detailed balance equation then reduces to

$$\frac{A(i \rightarrow f)}{A(f \rightarrow i)} = \frac{P(f)}{P(i)} = \exp[-\beta(u(f) - u(i))] \tag{16}$$

A possible solution to this equation is

$$A(i \rightarrow f) = min\{1, \exp[-\beta(u(f) - u(i))]\} \tag{17}$$

2.4 Canonical Ensemble (NVT) Sampling (or How to Sample in Realistic Experimental Conditions)

The configurational canonical partition function associated with such an ensemble is obtained by

$$Z(N, V, T) = \int dr^N \exp\left[-\beta u\left(r^N\right)\right] \tag{18}$$

The above equation can also be given as:

$$Q(N, V, T) = M_{\{\mathrm{N}\}}\, Z(N, V, T) \tag{19}$$

where $M_{\{\mathrm{N}\}}$ is constant, as:

$$M_{\{\mathrm{N}\}} = \frac{1}{\left(\sqrt{h^2\beta/2\pi m_{\mathrm{A}}}\right) N_{\mathrm{A}}! \left(\sqrt{h^2\beta/2\pi m_{\mathrm{B}}}\right) N_{\mathrm{B}}! \cdots}. \tag{20}$$

This constant takes into account the indiscernibility of the constituent atoms. For example, in the case of microcanonical ensembles, the probability is given as:

$$Pr_{NVT}\left(r^N\right) dr^N = \frac{\exp[-\beta u(r^N)] dr^N}{Z(N, V, T,)} \tag{21}$$

Hence the average value of the observable is given by

$$\langle O \rangle = \frac{1}{Z(N, V, T)} \int dr^N \exp\left[-\beta u\left(r^N\ \right)\right] o\left(r^N\right) \tag{22}$$

For the canonical partition function, the acceptance probability associated with the MC method is given as follows:

$$A_{NVT}\left(r^N \rightarrow r^{N'} = \min\left\{1, \exp\left[-\beta\left(u\left(r^N\right) - u\left(r^{N\prime}\right)\right)\right]\right\} \tag{23}$$

From the above canonical partition function, various thermodynamical quantities can be obtained, such as Helmholtz free energy, using the Eq. (24):

$$F(N,V,T) = -k_{\mathrm{B}}T \ \ln \ Q\ (N,V,T) \tag{24}$$

But there are many quantities which are obtained at constant temperature and pressure.

2.5 Isobaric-Isothermal Ensemble (NPT) Sampling (or How to Sample in Even More Realistic Experimental Conditions)

Many experimental conditions are represented by isobaric-isothermal ensembles. For these ensembles, microcanonical partition function is

$$Z(N,P,T) = \int dV \ \exp[-\beta PV] \int dr^N \exp \ \left[-\beta u\left(r^N\right)\right]. \tag{25}$$

In case of indiscernibility of atoms, the partition function becomes

$$Q(N,P,T) = \frac{M_{\{\mathrm{N}\}}}{V_0} \ Z(N,P,T) \tag{26}$$

In the above partition function, the isotropic macromolecular structures are assumed deformed to maintain the pressure constant. In the case of anisotropic deformations, the partition function is modified as follows:

$$Z(N,P,T) = \int dHZ'(N,P,T,H)\delta(\mathrm{det}H - V) \tag{27}$$

where H is the tensor associated with an elementary parallelepiped volume. The probability that a macromolecular system is in a state r^N is given by

$$Pr_{NPT}\left(r^N\right)dr^N = \frac{\exp[-\beta PV]\exp[-\beta u(r^N)]dr^N}{Z(N,P,T)} \tag{28}$$

The Gibbs free energy can also be obtained by partition function, using the equation:

$$G(N,P,T) = -k_B T \ln \ Q(N,P,T) \tag{29}$$

The isobaric-isothermal acceptance probability associated with the MC method is

$$A_{NPT}\left(r^N, V \to r'^N, V'\right) = \min\left\{1, \exp\left[-\beta\left(u\left(r'^N, V'\right) - u\left(r^N, V\right)\right)\right] * \exp\left[-\beta P(V' - V) + N \ln \frac{V'}{V}\right]\right\} \tag{30}$$

Here the Metropolis algorithm is impaired by local minima [55]. The acceptance probability is observed by the local minimum of the potential energy, which is performed by sampling the macromolecular states [56].

2.6 Sampling and Local Minima (or When Temperature May Help to Escape Local Minima)

In nature, there are various biomolecular processes in which a high energy barrier exists between the initial and final states [57]. For efficient macromolecular sampling, overcoming this type of barrier height is essential. To overcome barrier height, a computationally expensive simulation method is used for sampling called Replica exchange [58]. Replica exchange (or parallel tempering) involves a certain number of non-interacting simulations called replicas. Each simulation is parallelly performed at its own temperature. Low-temperature simulation tends to local minima, while high-temperature simulation tends to overcome barrier height and move between local minima.

To better explore the sampling of macromolecular states, the replicas are periodically exchanged according to the following acceptance probability:

$$A_{\mathrm{R}}(i \to f) = \min\left\{1, \exp\left[-\left(\beta_f - \beta_i\right) * (u(f)|Tf - u(i)|Ti)\right]\right\} \tag{31}$$

After completing one exchange, the simulations resume normal r unless another exchange is performed. Generally, symmetrical functions are used for sampling, but instead of restricting, one can consider a nonsymmetrical sampling function. In the case of nonsymmetrical functions, the final conformation is given as:

$$T(i \to f) = \pi(u(f)) \tag{32}$$

Then the acceptance probability becomes

$$A(i \to f) = \min\left\{1, \frac{\pi(u(i))}{\pi(u(f))} \exp[-\beta(u(f) - u(i))]\right\} \tag{33}$$

The above sampling is known as the bias sampling algorithm [58]. This consideration increases the efficiency of large macromolecular chains.

3 Advantages and Limitations of MC Simulations

3.1 Advantages

The MD simulation method is based on classical mechanics, where Newton's equations of motion are used. In contrast, MC simulations are free from these restrictions. This helps to generate the new conformation of the ensemble of choice. MC method is based on statistical mechanics, where moves are nontrivial, and they can sample a large number of individual steps up to 10^{10} or more in equilibrium. Some specific MC moves can provide great flexibility to solve some specific problem by combining several simulations. Additionally, MC methods can be performed parallelly by using multiple CPU clusters. These advantages make MC simulations more convenient and useful than other simulation methods.

3.2 Limitations

Since MC simulations do not use Newton's equation of motion, they cannot observe information on dynamics. The main disadvantage of MC simulation of proteins is the explicit solvent effect. The explicit solvent effect induces difficulties in large-scale movement. Various moves of simulations of proteins that change the internal coordinates without moving the solvent particles help form a large overlap of atoms. It results in the rejection of trial configurations. However, the simulations with implicit solvent models do not show these drawbacks; hence these models are the most popular method for MC simulations of proteins. Another disadvantage of MS simulations is that there is no general, reasonable, and freely available program for protein simulations. This is because the choice of simulation moves and attempting rates vary for a specific problem. Nowadays, the MC module has been added to CHARMM software [59]. In the following section, we present a case study discussing methods and properties of MC simulations for Trp-cage proteins.

4 Case Study of the MC Simulations of a Trp-Cage Protein

The Trp-cage protein (PDB ID: 1L2Y) is a mini-protein of 20 amino acid residues. It has the property to fold rapidly and spontaneously and has been of high interest to both experimentalists and theoreticians [60]. Trp-cage plays a vital role in understanding the enhancement of protein stability and improving drug binding efficiency via mutations of different proteins [61, 62]. Additionally, for the last two decades, this protein has been used as a benchmark of force fields and modelling techniques to provide detailed structural and thermodynamic data [63]. We discuss a case study where authors performed the MC simulations of Trp-cage at 200 million MC steps

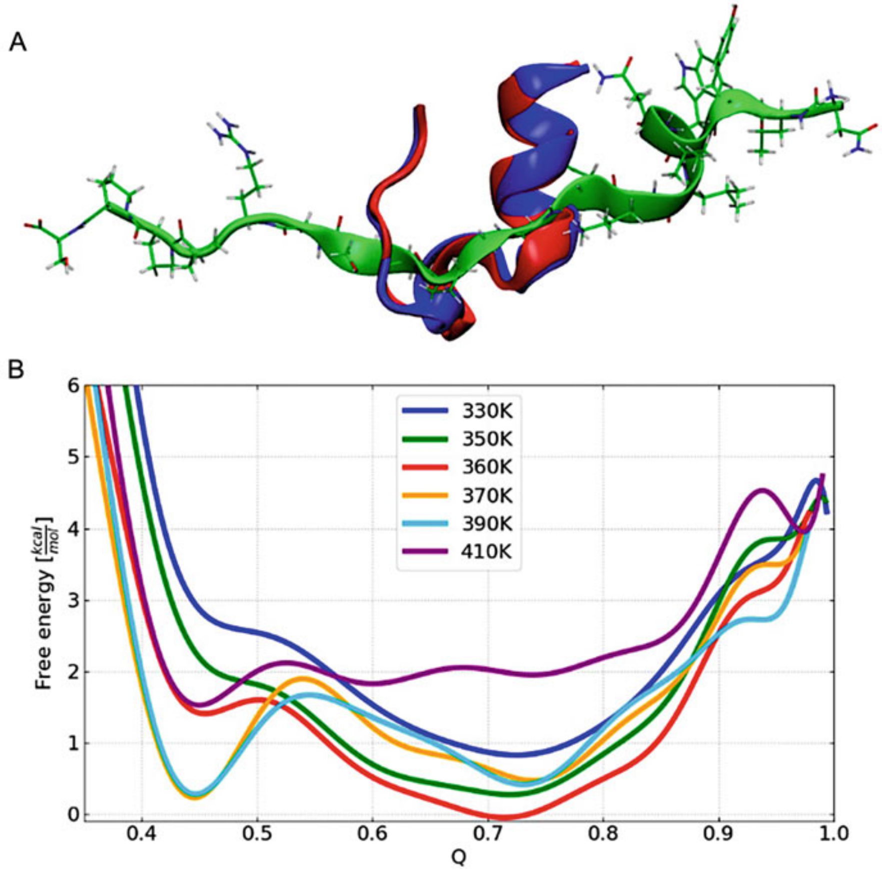

Fig. 1 Conformational landscape sampling of the Trp-cage protein (PDB ID: 1L2Y). (**a**) Superimposition of the native structure of Trp-cage protein (in blue) with the refolded structure (red) obtained from the MC simulation at 370 K (Full trajectory is shown in Fig. 2). (**b**) Free energy profiles as a function of the reaction coordinate Q (fraction of native contacts) at different temperatures calculated from the MC simulations. The refolded and intermediate ensembles were observed at $Q \sim 0.73$ and 0.45, respectively. The figure is adapted with permission from [59]

on the AMD EPYC 7551P node using 15 to 30 cores of 181 and 108 h of CPU time at 330 and 410 K temperatures. The structure of Trp-cage consists of α-helix (2–9 residues), a single turn of 3^{10} helix (11–14 residues,) and a hydrophobic core made of proline residues (Pro12, Pro18, Pro19) and Tyr3, Trp6 (Fig. 1a).

Conformational folding is modulated via interactions between polar groups of the protein and water molecules. Hence, to understand the correct mechanism of folding, it is also essential to properly treat the solvation environment [64]. There are various solvent models used in simulations to correctly refold the structure of the Trp-cage protein. Here, the authors used a generalized Born-based implicit solvent model with the AMBERff99SB-ILDN force field. Figure 1a depicts the overlay of the native (blue), refolded (red), and completely unfolded structures of the Trp-cage protein (green). The MC simulation run was performed from the unfolded Trp-cage with the fraction of native contacts (*Q*) of 0.07 (Fig. 2). The refolded structures matched well with the native Trp-cage as the Cα RMSD of the refolded protein was 0.86 Å at 370 K with only minor deviations around the turn of the helix (Fig. 1a, red). This indicated the high quality of the force field as well as the sampling efficiency of the MC method with an accumulated acceptance ratio of 60%.

The free energy profile was estimated using the potential of mean force (PMF) projected on the fraction of native contacts (*Q*) to determine the folding temperature of the protein. This process is widely used in reaction coordinates for the folding process (Fig. 1b) [54]. It indicates how similar the native and predicted structures of a protein are, with $Q \sim 0.9$–1.0 being the closest to the native structure obtained in NMR. While the two primary states of the conformational ensemble describing its folded ($Q \sim 0.73$) and unfolded states were observed, a partially unfolded state ($Q \sim 0.45$) with an energy barrier of 1.20 and 0.80 kcal.mol^{-1} at 370 and 390 K, respectively, was also observed. The folded protein structure minimum ($Q \sim 0.73$) was broad, resulting from its weak stabilization due to implicit solvation. To resolve this issue, MD simulations with an explicit water solvent were used, where a *Q* of ~0.9–1.0 for the folded Trp-cage was reported [64].

A broad range of folding temperatures of the protein was observed starting from 370 K. At 330 K, only a single energy minimum was observed before a free energy surface with two minima appeared at about 350 K. The estimated folding temperature, i.e. the temperature when both minima are equally probable, is significantly higher than was experimentally observed, i.e. 315–317 K [62], or calculated using all-atom force fields with explicit solvation, i.e. 321–326 K [65], but it is consistent with the folding temperatures obtained using implicit solvent models (375–400 K). This is due to the lack of temperature dependence of the implicit solvent model. The generalized Born solvent-accessible surface area (GBSA)-type implicit solvation models used here are known to over-stabilize the folded states of proteins, especially those stabilized with solvent-exposed hydrogen-bond salt bridges [66, 67]. The breaking or formation of such hydrogen-bond salt bridges is the primary regulator of Trp-cage folding and refolding, inducing the observed increase in the folding temperature.

This work on the refolded Trp-cage indicates the accuracy of the force field and MC method. A few refolded structures are shown in Figs. 3 and 4. The Cα RMSD of the refolded protein was 0.97 Å. The secondary structures α-helix and proline were present correctly in the refolded protein (Fig. 3) with Cα RMSD of 0.73 and 0.47 Å, respectively. Arg16 is among the most flexible residues in the refolded protein and cannot form the salt bridge in the refolded structure (while in the native structure, it forms a salt bridge) (Figs. 3 and 4). This is because the H⋯O distance between Asp9 and Arg16 is far more than in the native structure. This outcome most likely results from the limitations of the implicit solvation model. Although the salt bridge was unstable in the MC simulations, the refolded Trp-cage structure conserves the two main secondary elements, with the most noticeable difference being an N–H⋯O hydrogen bond between Trp6 (H-bond donor) and Arg16 (H-bond acceptor) present

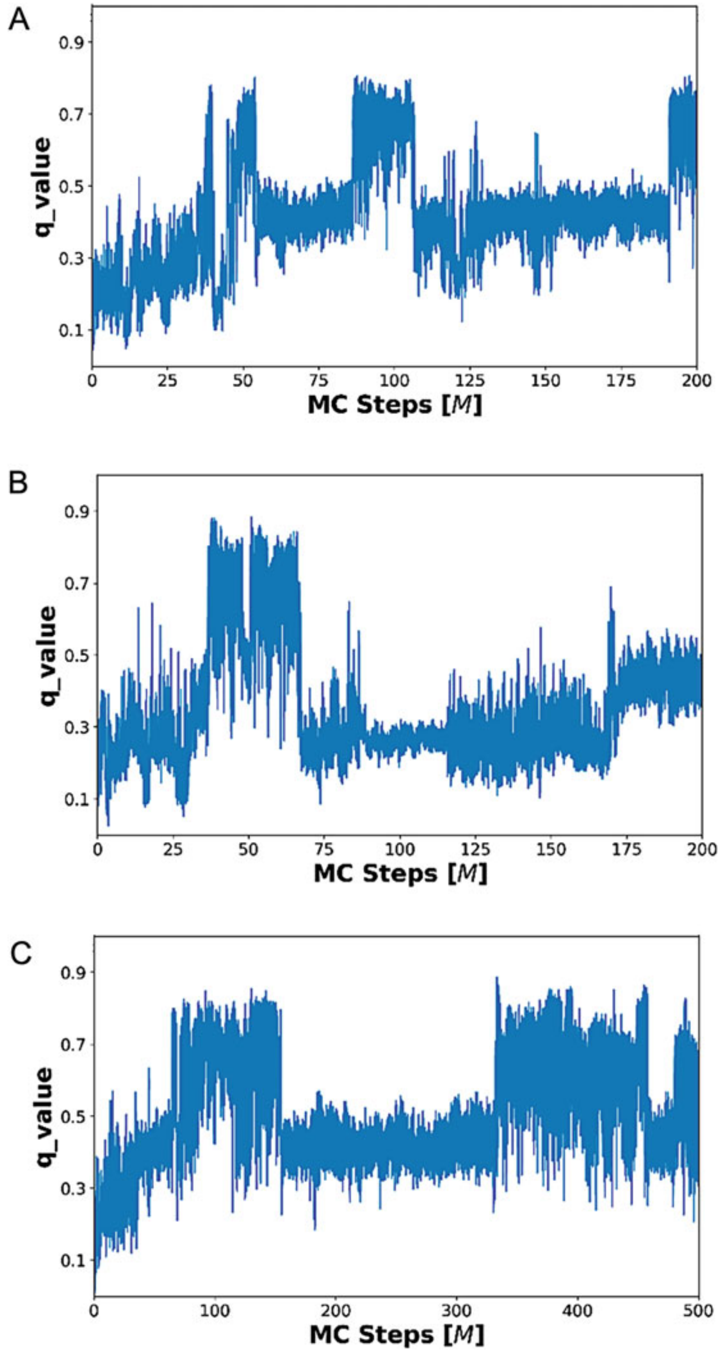

Fig. 2 MC simulations of Trp-cage starting from the unfolded structure. MC trajectories at the transition temperature of 370 K reveal the change in the Q. Multiple folding and unfolding events were observed. The MC simulation started from the unfolded structure with $Q \sim 0.07$ (Fig. 1, green colour) (**a**) and $Q \sim 0.12$ (**b, c**). The figure is adapted with permission from [59]

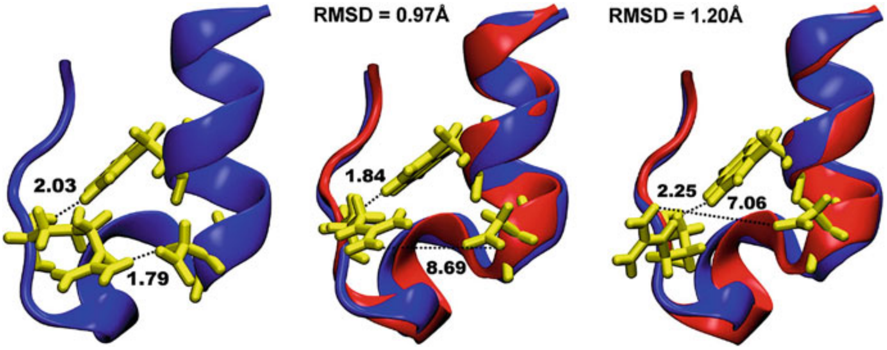

Fig. 3 Local minima representing the refolded conformers of the Trp-cage. The MC simulation started from the unfolded protein structure (see Fig. 2). The native and refolded structures are shown in blue and red, respectively. Asp9 and Arg16 form a hydrogen-bonded salt bridge (distance in the native state ~1.79 Å), and Trp6 forms a hydrogen bond with Arg16 (distance in the native state ~2.03 Å), shown in yellow. The figure is adapted with permission from [59]

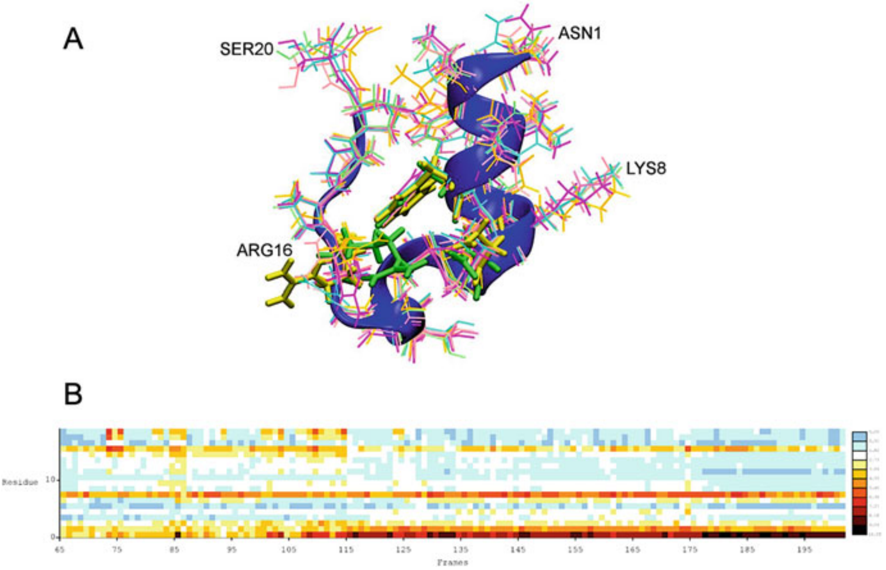

Fig. 4 (**a**) Overlay of the different refolded conformers of Trp-cage, obtained from the MC simulations at the folding temperature. Asn1, Lys8, Arg16, and Ser20 are the most flexible residues and are marked. Trp6, Asp9, and Arg16 in the native structure are shown in green sticks, while those residues in the refolded state are shown in yellow sticks. (**b**) Heatmap of a particular residue RMSD changes in the different refolded conformers as a fluctuation from its position in the native form. The changes close to the native-like structure are shown in blue, while residues with high RMSD are in brown and chestnut brown. The figure is adapted with permission from [59]

at a distance of 2.03 Å (Fig. 2), which is stable in the refolded Trp-cage. This hydrogen bond, along with the salt bridge between Asp9 and Arg16, regulates the fast folding of the Trp-cage protein.

5 Conclusions

The MC simulation is a widely used method in the ab initio protein structure prediction. In this chapter, we discussed the advantages of MC simulations for revealing the conformations of protein folding and its structural dynamics. As MC simulations are based on statistical physics, they can provide thermodynamic properties of protein complexes by calculating the average of ensembles. It is also not limited by the force field parameters like the conventional MD simulation methods. In the MC simulation method, the partition function can also be obtained by the average of ensembles and provide information on various free energies of the complex. The MC method is used to explore the folding mechanism of the Trp-cage protein, where it performs well with less computational power. It will be interesting to apply the efficient MC algorithm with concerted rotations to larger systems and investigate its performance when replica exchange moves are included.

References

1. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins. Nature **450**(7172), 964–972 (2007)
2. E.Z. Eisenmesser, O. Millet, W. Labeikovsky, et al., Intrinsic dynamics of an enzyme underlies catalysis. Nature **438**, 117–121 (2005)
3. L.V. Bock, C. Blau, G.F. Schröder, et al., Energy barriers and driving forces in tRNA translocation through the ribosome. Nat. Struct. Mol. Biol. **20**(12), 1390–1396 (2013)
4. C.D. Snow, H. Nguyen, V.S. Pande, M. Gruebele, Absolute comparison of simulated and experimental-folding dynamics. Nature **420**(6911), 102–106 (2002)
5. C. Cecconi, E.A. Shank, C. Bustamante, S. Marqusee, Direct observation of the three-state folding of a single protein molecule. Science **4174**, 2057–2060 (2005)
6. V. Duaw, P.A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science **282**(5389), 740–744 (1998)
7. H. Lei, C. Wu, H. Liu, Y. Duan, Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc. Natl. Acad. Sci. U S A **104**(12), 4925–4930 (2007)
8. T. Herges, W. Wenzel, Free-energy landscape of the villin headpiece in an all-atom force field. Structure **13**, 661–668 (2005)
9. A. Schug, T. Herges, W. Wenzel, Reproducible protein folding with the stochastic tunneling method. Phys. Rev. Lett. **91**(15), 158102 (2003)
10. J.A. Vila, D.R. Ripoll, H.A. Scheraga, Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. Proc. Natl. Acad. Sci. U S A **100**(25), 14812–14816 (2003)
11. R.D. Hills Jr., L. Lu, G.A. Voth, Multiscale coarse-graining of the protein energy landscape. PLoS Comput. Biol. **6**(6), e1000827 (2010)

12. E. Suárez, J.L. Adelman, D.M. Zuckerman, J.L. Adelman, D.M. Zuckerman, Accurate estimation of protein folding and unfolding times: beyond Markov state models. J. Chem. Theory Comput. **12**(8), 3473–3481 (2016)
13. A. Irbäck, S. Mitternacht, S. Mohanty, An effective all-atom potential for proteins. PMC Biophys. **24**(1), 2 (2009)
14. C.S. Division, PROFASI: a Monte Carlo simulation package for protein folding and aggregation. J. Comput. Chem. **27**(13), 1548–1555 (2006)
15. F. Ding, D. Tsao, H. Nie, N.V. Dokholyan, Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure **16**(7), 1010–1018 (2008)
16. J.A.N.H. Meinke, U.H.E. Hansmann, Free-energy-driven folding and thermodynamics of the 67-residue protein GS-α3W—a large-scale Monte Carlo study. J. Comput. Chem. **30**(11), 1642–1648 (2009)
17. B. Hess, S. Uppsala, D. van der Spoel, E. Lindahl, GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. **4**, 435–447 (2008)
18. T. Veitshans, D. Klimov, D. Thirumalai, Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. Fold. Des. **2**(1), 1–22 (1997)
19. S. Scheindlin, The duplicitous nature of inorganic arsenic. Mol. Interv. **5**(2), 60–64 (2005)
20. D.E. Shaw, Atomic-level characterization of the structural dynamics of proteins. Science **330**, 341–346 (2010)
21. R.O. Dror, T.J. Mildorf, D. Hilger, et al., SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins. Science **348**(6241), 1361–1365 (2015)
22. H. Lee, Structures, dynamics, and hydrogen-bond interactions of antifreeze proteins in TIP4P/Ice water and their dependence on force fields. PLoS One **13**(6), e0198887 (2018)
23. D.H. De Jong, G. Singh, W.F.D. Bennett, et al., Improved parameters for the Martini coarse-grained protein force field. J. Chem. Theory Comput. **9**(1), 687–697 (2013)
24. H. Nguyen, J. Maier, H. Huang, V. Perrone, C. Simmerling, Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. J. Am. Chem. Soc. **136**(40), 13959–13962 (2014)
25. N. Onuchic, A.E. García, Folding a protein in a computer: an atomic description of the folding-unfolding of protein A. Proc. Natl. Acad. Sci. U S A **100**(24), 13898–13903 (2003)
26. F. Noé, C. Schütte, E. Vanden-eijnden, L. Reich, T.R. Weikl, Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc. Natl. Acad. Sci. U S A **106**(45), 19011–19016 (2009)
27. J.L. Sessler, D. Seidel, Synthetic expanded porphyrin chemistry. Angew. Chem. Int. Ed. Engl. **42**(42), 5134–5175 (2003)
28. P. Ferrara, J. Apostolakis, A. Caflisch, Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins **46**, 24–33 (2002)
29. A.H. Follmer, M. Mahomed, D.B. Goodin, T.L. Poulos, Substrate-dependent allosteric regulation in cytochrome P450cam (CYP101A1). J. Am. Chem. Soc. **140**(47), 16222–16228 (2018)
30. Q. Shao, W. Zhu, How well can implicit solvent simulations explore folding pathways? A quantitative analysis of α—helix bundle proteins. J. Chem. Theory Comput. **13**(12), 6177–6190 (2017)
31. J.D. Durrant, J.A. McCammon, Molecular dynamics simulations and drug discovery. BMC Biol. **9**, 71 (2011)
32. P. Tao, Y. Xiao, Using the generalized born surface area model to fold proteins yields more effective sampling while qualitatively preserving the folding landscape. Phys. Rev. E **101**(6), 62417 (2020)
33. R. Harada, Y. Shigeta, Temperature shuffled structural dissimilarity sampling based on a root-mean square deviation. J. Chem. Inf. Model. **58**(7), 1397 (2018)
34. J.A. McCammon, M. Karplus, Simulation of protein dynamics. Ann. Rev. Phys. Chem. **31**, 29–45 (1980)
35. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087 (1953)

36. F. Liang, W.H. Wong, Evolutionary Monte Carlo for protein folding simulations. J. Chem. Phys. **115**(7), 3374–3380 (2001)
37. P.S. Nerenberg, T. Head-Gordon, ScienceDirect New developments in force fields for biomolecular simulations. Curr. Opin. Struct. Biol. **49**, 129–138 (2018)
38. A. Perez, J.A. Morrone, C. Simmerling, K.A. Dill, ScienceDirect advances in free-energy-based simulations of protein folding and ligand binding. Curr. Opin. Struct. Biol. **36**, 25–31 (2016)
39. J. Kleinjung, F. Fraternali, ScienceDirect design and application of implicit solvent models in biomolecular simulations. Curr. Opin. Struct. Biol. **25**, 126–134 (2014)
40. R. Anandakrishnan, A. Drozdetski, R.C. Walker, A.V. Onufriev, Article speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. Biophys. J. **108**(5), 1153–1164 (2015)
41. J. Shimada, E.I. Shakhnovich, The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. Proc. Natl. Acad. Sci. U S A **99**(17), 11175–11180 (2002)
42. S. Kmiecik, D. Gront, M. Kolinski, et al., Coarse-grained protein models and their applications. Chem. Rev. **116**(14), 7898–7936 (2016)
43. C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein structure prediction using Rosetta. Methods Enzymol. **383**, 66–93 (2004)
44. F. Eisenmenger, U.H.E. Hansmann, S. Hayryan, C. Hu, [SMMP] A modern package for simulation of proteins. Comput. Phys. Commun. **138**, 192–212 (2001)
45. W. Pulawski, M. Jamroz, M. Kolinski, A. Kolinski, S. Kmiecik, Coarse-grained simulations of membrane insertion and folding of small helical proteins using CABS model. J. Chem. Inf. Model. **56**(11), 2207–2215 (2016)
46. M. Khalili, A. Liwo, H.A. Scheraga, Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains. J. Mol. Biol. **355**(3), 536–547 (2006)
47. A. Liwo, M. Baranowski, C. Czaplewski, E. Go, Y. He, D. Jagie, A unified coarse-grained model of biological macromolecules based on mean-field multipole—multipole interactions. J. Mol. Model. **20**(8), 2306 (2014)
48. M. Kurcinski, A. Kolinski, S. Kmiecik, Mechanism of folding and binding of an intrinsically disordered protein as revealed by ab initio simulations. J. Chem. Theory Comput. **10**(6), 2224–2231 (2014)
49. S. Fiorucci, M. Zacharias, Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. Proteins **78**(15), 3131–3139 (2010)
50. M. Feig, J. Karanicolas, C.L. Brooks III, MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. J. Mol. Graph. Model. **22**, 377–395 (2004)
51. J. Nasica-labouze, M. Meli, P. Derreumaux, G. Colombo, A multiscale approach to characterize the early aggregation steps of the amyloid-forming peptide GNNQQNY from the yeast prion sup-35. PLoS Comput. Biol. **7**(5), e1002051 (2011)
52. K.A. Fichthorn, W.H. Weinberg, K.A. Fichthorn, Theoretical foundations of dynamical Monte Carlo simulations theoretical foundations of dynamical Monte Carlo simulations. J. Chem. Phys. **95**, 1090 (1991)
53. S. Duane, A.D. Kennedy, B.J. Pendleton, D. Roweth, Hybrid Monte Carlo. Phys. Lett. B **195**(2), 216–222 (1987)
54. G.J. Martyna, D.J. Tobias, M.L. Klein, G.J. Martyna, D.J. Tobias, M.L. Klein, Constant pressure molecular dynamics algorithms. J. Chem. Phys. **101**, 4177 (1994)
55. M. Shu, Z. Lin, Y. Zhang, Y. Wu, Molecular dynamics simulation of oseltamivir resistance in neuraminidase of avian influenza H5N1 virus. J. Mol. Model. **17**, 587–592 (2011)
56. P. Taylor, J.I. Siepmann, Configurational bias Monte Carlo: a new sampling scheme for flexible chains. Mol. Phys. **75**, 37–41 (2013)
57. P. Minary, M.E. Tuckerman, G.J. Martyna, Dynamical spatial warping: a novel method for the conformational sampling of biophysical structure. SIAM J. Sci. Comput. **30**(4), 2055–2083 (2008)

58. E. Paquet, H.L. Viktor, M.C. Simulations, Molecular dynamics, Monte Carlo simulations, and Langevin dynamics: a computational review. Biomed. Res. Int. **2015**, 183918 (2015)
59. J.I.E. Hu, A.O. Ma, A.R. Dinner, Monte Carlo simulations of biomolecules: the MC module in CHARMM. J. Comput. Chem. **27**(2), 203–216 (2006)
60. J.W. Neidigh, R.M. Fesinmeyer, N.H. Andersen, Designing a 20-residue protein. Nat. Struct. Biol. **9**(6), 425–430 (2002)
61. D.J. Drucker, J.B. Buse, K. Taylor, et al., Exenatide once weekly versus twice daily for the treatment of type 2 diabetes: a randomised, open-label, non-inferiority study. Lancet **372**(9645), 1240–1250 (2008)
62. B. Barua, J.C. Lin, V.D. Williams, P. Kummler, J.W. Neidigh, N.H. Andersen, The Trp-cage: optimizing the stability of a globular miniprotein. Protein Eng. Des. Sel. **21**(3), 171–185 (2008)
63. C. Simmerling, B. Strockbine, A.E. Roitberg, All-atom structure prediction and folding simulations of a stable protein. J. Am. Chem. Soc. **124**(38), 11258–11259 (2002)
64. R.B. Best, J. Mittal, Balance between alpha and beta structures in ab initio protein folding. J. Phys. Chem. B **114**(26), 8790–8798 (2010)
65. R. Day, D. Paschek, A.E. Garcia, Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage miniprotein. Proteins **78**(8), 1889–1899 (2010)
66. D. Lili, M.E.I. Ye, L.I. Yongle, Z. Qinggang, Z. Dawei, Z.J. Zenghui, Simulation of the thermodynamics of folding and unfolding of the Trp-cage mini-protein TC5b using different combinations of force fields and solvation models. Sci. China Chem. **53**(1), 196–201 (2010)
67. B.D. Bursulaya, C.L.B. Brooks, Comparative study of the folding free energy landscape of a three-stranded-sheet protein with explicit and implicit solvent models. J. Phys. Chem. B **104**(51), 12378–12383 (2000)

Markov State Models of Molecular Simulations to Study Protein Folding and Dynamics

Vivek Junghare, Sourya Bhattacharya, Khalid Ansari, and Saugata Hazra

Abstract Proteins are essential units of life that govern several functions. Understanding their behavior is closely related to their conformations, native folds, and change in conformations. Thus, the dynamic information of protein becomes essential to understand its properties at the molecular level. The molecular dynamics (MD) simulation approach provides atomistic-level dynamic information about proteins. However, more extended or complex MD simulations of protein are challenging to analyze and to gather meaningful confirmation from several snapshots of the dynamic system. To achieve it, i.e., analyzing MD simulation data, Markov State Model (MSM) is a powerful tool that has a statistical background. It represents the MD simulation system as a combination of finite memoryless states, i.e., states that are not dependent on prior states and transition probability among such states. MSM applications have grown from peptides to membrane protein simulations. The present book chapter sheds light on MD simulation's role in protein dynamics and why MSM is required. The brief theoretical aspects of MSM techniques are demonstrated. Lastly, the chapter discusses the application of MSM in different protein folding and dynamics.

Keywords Molecular dynamics (MD) simulation · Markov state model · Dynamics · Sampling · Statistical approach · Transition count matrix

V. Junghare · S. Bhattacharya
Department of Biosciences and Bioengineering, Indian Institute of Technology Roorkee, Roorkee, India

K. Ansari
Department of Physics, Indian Institute of Technology Roorkee, Roorkee, India

S. Hazra (✉)
Department of Biosciences and Bioengineering, Indian Institute of Technology Roorkee, Roorkee, India

Center for Nanotechnology, Indian Institute of Technology Roorkee, Roorkee, India
e-mail: saugata.hazra@bt.iitr.ac.in

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_8

1 Introduction

Protein dynamics and folding have been challenging phenomena essential for the molecular-level understanding of protein function. Molecular dynamics (MD) simulation is a valuable tool that comprehends macromolecular structural and functional insights. Data assembled after the MD simulation study can confer good knowledge about the macromolecular structure and provide detailed informational insights [1].

1.1 Importance of Molecular Dynamics

Proteins and nucleic acids are dynamic entities, and their dynamics play a significant role in their functions. Crystal structures stored at the PDB provide a halfway and limited perspective on three-dimensional (3D) construction. Especially protein molecules undergo crucial conformational changes during a particular function [2, 3]. One such change is the structural rearrangement in the protein molecule upon binding a substrate or inhibitor [4, 5]. This can be effectively verified by comparing apo and ligand-bound 3D protein structures. The conformational changes are usual parameters of enzymes' catalytic mechanisms [6]. One of the common instances is loop movement or domain rearrangements that change the local composition of the active site's chemical environment to perform a function. Sometimes, these alterations activate the catalytic process by bringing protein subunits together. Moreover, one can correlate protein function only when dynamic properties are considered [7–9].

There are several ways to deal with the conformation correlated with the relevant macromolecular function. One of the conventional ways is to gather experimentally determined structures covering the conformational space using X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (cryo-EM) methods. These methods can be used to study structures of macromolecules in different environments or bound with other substrates or ligands. However, these experimental studies are time taking and need specific high-end instruments.

On the other hand, theoretical strategies are the most helpful method for getting an image of the macromolecular dynamic properties of a protein. Protein folding occurs in a timescale of a few microseconds, allosteric transitions in microseconds to milliseconds, relative motions of protein domains in nanoseconds to seconds, and dynamics of side chains in picoseconds to nanoseconds (Fig. 1) [10]. Additionally, it is observed that longer timescale motions can influence shorter timescale dynamics and vice versa. Hence, long timescale simulations have always been a well-chosen option [11, 12]. Long-time simulations provide an opportunity to understand the flexibility of proteins and their related ensemble of alternative structural states, which are crucial for understanding the folding and dynamics of proteins [13, 14].

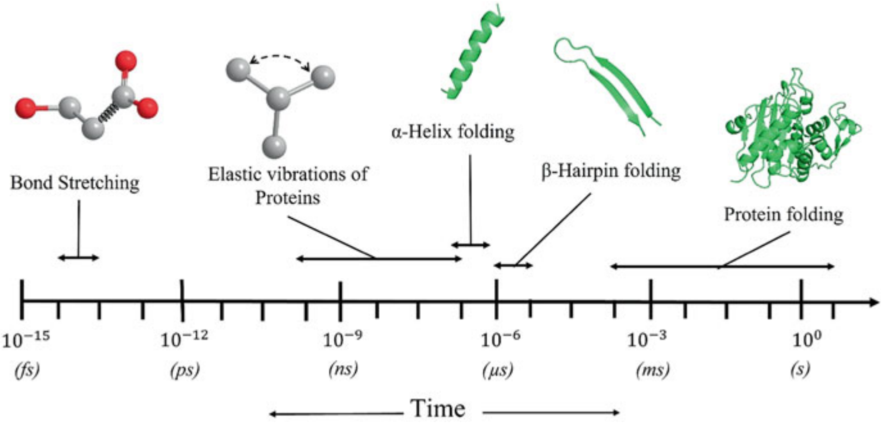

Fig. 1 The figure represents the protein motion concerning the time axis. The MD simulations must be performed at femtosecond time steps to capture the bond stretching motion and similarly for other represented motions that are more time-scaled atomistic. Hence, more computational power is required

Protein-conformational changes play a vital role in its functioning [15, 16]. Hence it is not enough to study just one PDB conformer. Modern-day advances in simulation algorithms and calculations have promoted the idea of "conformational ensembles" as an option in contrast to examining a single structure from PDB. These ensembles or conformers can be examined to determine thermodynamic properties, entropy, free energy, conformational changes, or protein folding phenomenon [16–18]. There are two significant difficulties in analyzing MD simulations of biomolecules: adequate conformational sampling and exact physical force fields. Despite remarkable improvements in modern computing capacity, conventional MD (cMD) simulations are still essentially constrained to shorter timescales than those demonstrated by various biomolecular movements and functions [19–22]. Hence, to gather multiple conformations, a specified tool is required.

Furthermore, protein folding remains one of biology's fundamental and least understood phenomena. This fascinating phenomenon of conversion of the primary sequence of a protein to the native 3D structure remains less understood. Small molecular weight proteins with ~10–100 amino acid residues fold in the microsecond to sub-millisecond timescales, known as "fast-folding" proteins. They are magnificent model systems to study and analyze protein folding through long timescale cMD simulations in explicit water [23]. Protein folding needs a broad measure of conformational examination and computational ability to describe the free energy landscape appropriately. Advancement in computation with more extended simulations is insufficient to expand the conformational sampling in the molecular framework. The complicated state of the free energy landscape makes the majority of the simulations investigate only a small region around the energy least near to the initial conformation. With the accessibility of the current advanced HPC systems, a conspicuous methodology is to play out a series of parallel simulations with several initial energy-minimized conformations. Although this could be proficient, it requires detailed information on the framework to simulate and cannot be applied as an overall strategy.

Nevertheless, protein folding has been analyzed using cMD and utilizing productive examining methods such as replica-exchange MD [24], Markov State Models (MSM), biasing MD simulations such as bias-exchange metadynamics [25], and transition path sampling [26]. This chapter sheds light on how MSM helps tackle protein dynamics and folding problems.

1.2 Motivation Behind Using MSM Technique

At times, protein folding and dynamics require long timescale simulations, or the system becomes highly complex or enormous (such as in the case of membrane protein simulation). The first microsecond-length all-atom MD simulation of a small protein was carried out by Duan and Kollman [27]. Further advancements in computer power open up possibilities of MD simulations of thousands of protein atoms, long time-scaled simulation of proteins, etc. Biomacromolecules frequently perform their functions through dynamic transitions between conformational states. For instance, the AdeB efflux pump undergoes carbapenem resistance through conformational modifications [28]. By performing long timescale dynamics based on several short MD simulations, MSM has emerged as a prominent method for bridging this timescale gap [2, 29].

Representing physical, chemical, or biological systems using stochastic processes is standard practice. The objective is to analyze the stochastic model and roughly compute the exciting properties of the system. Direct sampling and building a coarse-grained model of the system are two methods for carrying out such analysis. In a direct sampling strategy, one attempt to produce a statistically significant number of occurrences representing the system property in question. Here, making sufficient statistics for accurate estimates requires much computation. Estimation through direct numerical simulation is impossible, especially if the state space is continuous and has a high dimension [30]. In the coarse-grained model, discretization of the systems state space is used. This is achievable using MSM. The advantage is that it uses discrete finite space. Due to this, the vast systems became finite discrete models that can be solved numerically to find their properties. It uses transition path theory (TPT) to analyze systems' discrete states. In summary, the analysis of the ensemble of reactive trajectories, or trajectories that originate from a specific set of states A and go to B. Hence using such a technique provides a more comprehensive analysis of biological protein simulations.

2 Markov State Model

A theoretical model, often known as the Markov State Model (MSM), is frequently used to study the dynamic nature of biological systems. The basic idea of MSM is making the square matrix known as the transition probability matrix (TPM). In the case of protein dynamics, MSM can be used after obtaining initial data from MD simulation trajectories.

2.1 Building of MSM

To develop MSM, an adaptive sampling algorithm is frequently used. Adaptive sampling is a statistical approach for solving protein dynamics on large timescales (100 μs to the ms) to sample conformational transitions. The adaptive sampling algorithm is based on iterations, which are used until the desired sampling criteria are reached [19]. The adaptive sampling process is divided into three steps: (i) to run an MD simulation and get many short trajectories, (ii) build an MSM using trajectories, and (iii) run a simulation trajectory based on obtained results from the MSM. MSM uses a matrix, so it needs microstates that can be prepared in two ways: one is based on geometric distributions (distance metric), and the other is based on a free energy map (kinetic-based metric). The preferred one is to choose free energy minima, i.e., kinetic distribution, instead of the geometric distribution. The pathway of MSM is illustrated in Fig. 2.

2.2 Microstates and Macrostates Generation

Microstates are required to construct MSM. They are the nonoverlapping discrete configurational space. Every transition among these microstates is not dependent on the previous state. This phenomenon is known as memoryless transition. In this regard, one needs microstates where shifts can happen smoothly and rapidly. For this, there is a requirement to group configurations, often known as clustering. Since many clustering techniques are available, one must choose them wisely. One of the clustering techniques is choosing a distance metric. The k-centers, k-medoids, and

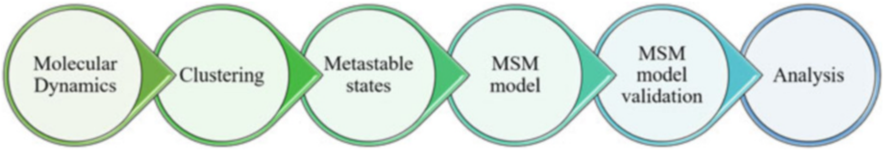

Fig. 2 The schematic pathway of the Markov State Model (MSM)

hybrid k-centers/k-medoids clustering are some of the essential clustering algorithms. To determine states, one needs to go through the MD simulations first and then find the suitable conformations based on either the root mean square deviation (RMSD) chosen appropriately 2 to 3 Å or based on the energy barriers. Most of the time, it is assumed that as the degree of structural similarity is higher, the corresponding kinetic similarity is also higher. It is known as the kinetic clustering of microstates into larger macrostates [31].

In Markovian microstate formation, there is a timeframe difference at which the states occur, often known as lag time or Markovian lag (τ). Hence, after lag time τ, the state will not be dependent on the previous state. MSM building requires a transition probability among these microstates, which depends on the number of microstates and lag time. Markovian lag should be large enough but not too large so that it does not alter significantly from other trajectories, which are often considered microstates. Markovian lag is just a method of selecting steps for trajectories that must be chosen carefully.

Additionally, in the case of tens of thousands of microstates or huge system sizes (such as membrane protein simulation), kinetic-based clustering can be performed that are supersets of microstates and are named macrostates. These macrostates are obtained using coarse-graining the model. This method collects microstates that are quickly clumping together and are collected to form macrostates. Available lumping procedures from microstates to macrostates are perron cluster cluster analysis (PCCA), their improved version (PCCAC), Bayesian agglomerative clustering engine (BACE), and super level set hierarchical clustering (SHC).

2.3 MSM Model and Validation

After obtaining the microstates, the next step is constructing the transition count matrix (TCM). It is a matrix that describes the transition from one state to another. The transition count matrix in general form is shown below:

$$M = \begin{bmatrix} a_{11} & & a_{12}\cdots\cdots & & a_{1n} \\ a_{21} & \ddots & a_{22}\cdots\cdots & \cdots & a_{2n} \\ \vdots & & \cdots & & \\ a_{n1} & a_{n2} & \ddots & \cdots & a_{nn} \end{bmatrix}$$

where a_{ij} denotes the transition from ith state to jth state. For example, if the states chosen from trajectories named A, B, and the trajectory are given as:

$$\text{Trajectory} : AABBBABABAABB.$$

Also, if the trajectory is chosen one step, then the number of transitions from A to A is 2 ($N_{AA} = 2$), from A to B is 4 ($N_{AB} = 4$), from B to A is 3 ($N_{BA} = 3$), and from B to B is 3 ($N_{BB} = 3$). Then the TCM can be written as mentioned in Table 1.

Table 1 Transition count matrix representing the transition between states *A* and *B*

From\To	*A*	*B*
A	2	4
B	3	3

Table 2 Transpose of the transition count matrix

From\To	*A*	*B*
A	2	3
B	4	3

The transition count matrix is usually not symmetric, so it is necessary to make a symmetric matrix and any symmetric matrix. One must follow the symmetry property of the matrix, which is defined as any (square) matrix. It is written as the sum of a symmetric matrix and an antisymmetric matrix [32].

$$M = \frac{[M + M^T]}{2} + \frac{[M - M^T]}{2}$$

where M^T is the transpose of M, $[M + M^T]$ is symmetric, and $[M - M^T]$ is antisymmetric.

This matrix should be symmetric because the transition between states depends not only on the forward direction but also on the reverse direction and is transposable. The transpose matrix describes moving from one state to another in either a forward or reverse direction. The transpose of TCM is shown below:

$$M^T = \begin{bmatrix} a_{11} & a_{21}\cdots\cdots & & a_{n1} \\ a_{12} & \ddots \; a_{22}\cdots\cdots & \cdots & a_{n2} \\ \vdots & \cdots & \cdots & \\ a_{1n} & a_{2n} \quad \ddots & \cdots & a_{nn} \end{bmatrix}$$

For the transpose matrix, the row (horizontal elements) is changed into a column (vertical components) and vice versa, as shown in Table 2.

Averaging the transition matrix counts by adding a transition matrix, and their transpose matrix gives symmetry.

$$M^{\text{symm}} = \frac{M + M^T}{2}$$

The symmetry matrix is shown below:

$$M_{\text{ij}}^{\text{symm}} = \frac{1}{2}\begin{bmatrix} a_{11}+a_{11} & a_{12}+a_{21}\cdots\cdots & & a_{1n}+a_{n1} \\ a_{21}+a_{12} & a_{22}+a_{22}\cdots\cdots & \cdots & a_{2n}+a_{n2} \\ \vdots & \ddots \quad \cdots & \cdots & \\ a_{n1}+a_{1n} & a_{n2}+a_{2n} \quad \ddots & \cdots & a_{nn}+a_{nn} \end{bmatrix}$$

For the present example, the symmetric matrix is shown in Table 3.

After this, reversible TPM will be calculated for each element of the matrix. There are two requirements for the TPM that must be rigorously followed. First, the total

Table 3 Symmetry matrix for present trajectory

From\To	*A*	*B*
A	2	3.5
B	3.5	3

Table 4 The transition probability matrix for a given trajectory

From\To	*A*	*B*
A	0.222	0.778
B	0.636	0.364

probability in each row is equal to unity, and second, elements should be nonnegative. There is no negative value meaning because probability only contains values between zero and one. Another essential point about transition probability is that it depends only on the time difference, i.e., the transition should be homogeneous [33].

$$P_{\mathrm{ij}} = \frac{M_{\mathrm{ij}}^{\mathrm{symm}}}{\sum_{i}^{j}\left(M_{\mathrm{ij}}^{\mathrm{symm}}\right)}$$

The transition probability matrix is shown below:

$$M_{\mathrm{prob}} = \begin{bmatrix} P_{11} & & P_{12}\ldots\ldots\ldots & & P_{1\mathrm{n}} \\ P_{21} & & P_{22}\ldots\ldots\ldots & & P_{2\mathrm{n}} \\ \vdots & \ddots & \cdots & \cdots & \\ P_{\mathrm{n}1} & P_{\mathrm{n}2} & \ddots & \cdots & P_{\mathrm{nn}} \end{bmatrix}$$

For the present example, the transition probability matrix will be shown in Table 4.

$$\text{Auxiliary equation}: |M - \lambda I| = 0;$$

where I is an identity matrix, and λ is for eigenvalues.

After solving the auxiliary equation for the TPM, one can get the eigenvectors and corresponding eigenvalues. The total sum of eigenvalues is to be zero. From eigenvalues data, one can analyze that the most positive value gives the most fluctuation from the equilibrium states, and the least negative value is in the most equilibrium states. There are several methods and tests to validate the models, such as Chapman–Kolmogorov equation model-based test, correlation function test, Bayesian Model selection, Swope–Pitera eigenvalue test, etc.

3 MSM to Understand Protein Folding and Dynamics

The initial studies of using MSM were started by studying peptide folding [34–36] and other small systems [37]. Further, it was applied in protein folding, protein–ligand binding, nucleic acids, and other biological problems (Fig. 3). It is used to analyze small-timescale and large-timescale simulations to gather relevant information. We now discuss how MSM is used to understand protein folding and dynamics, focusing on ensemble sampling and conformational fluctuations.

3.1 Peptide Modeling

Researchers have tried to address the issues related to understanding the mechanism of protein folding and finding the nature of folds. MD simulations have been regularly used along with experimental studies. In 2004, Swope et al. developed an algorithm to study the kinetics of protein folding. They applied it to a small peptide, a C-terminal alpha-hairpin motif from protein G. They used a Boltzmann-weighted ensemble to formulate the transition function from MD simulation [35]. They found the pattern and number of hydrogen bonds in a peptide. The Markov model depends on finding the finite number of metastable states; thus, identifying them is a critical and essential step. Hence, the clustering algorithm

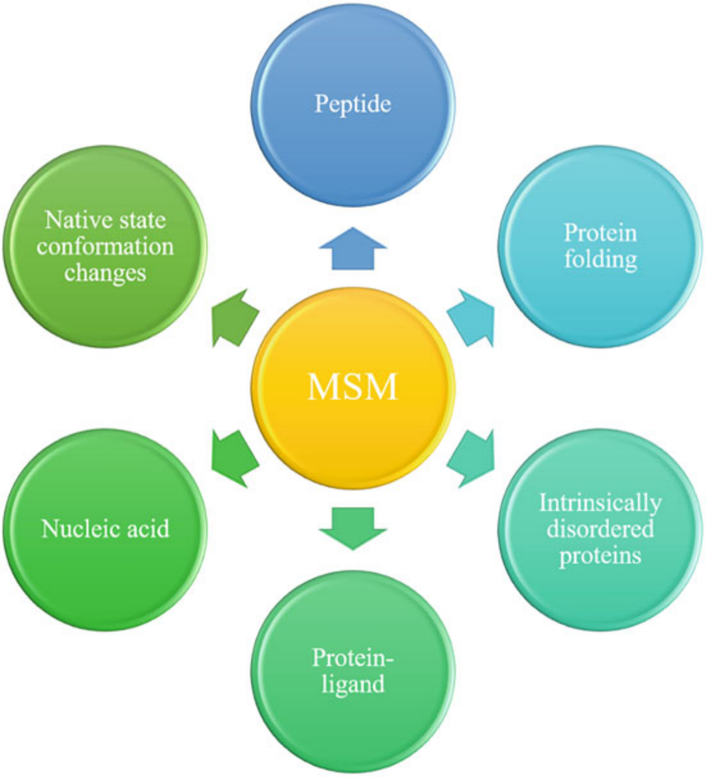
Peptide
Native state conformation changes
Protein folding
MSM
Nucleic acid
Intrinsically disordered proteins
Protein-ligand

Fig. 3 Applications of MSM in protein folding and protein dynamics

was applied to get kinetics-based states that were long-lived in dynamic systems. This kinetics-based clustering was used by Noe et al., who tested ALA_8 and ALA_{12} peptides [36]. This study, by Noé et al., brought a new direction to form metastable states, which consider dynamic behavior and not geometric proximity. Following this method, the automated algorithm was proposed, which detects the kinetically metastable states and was tested on three peptides [38]. After this, the master equation was developed by Buchete & Hummer for studying MD simulation of peptide folding at an atomistic level [39]. ALA_5 peptide was used for the study, which was intended to form a small helix. In recent studies, this technique has been used to study peptides like amyloid-β peptide (Aβ), which is responsible for Alzheimer's disease [40].

3.2 Protein Folding

Protein folding prediction through an in silico approach has been a mystery since the inception of protein simulation. Protein folds have numerous possibilities, as stated by the Leventhal paradox [41]. However, protein folds within a few microseconds in natural states and retains its native fold to function [42]. At the same time, predicting protein folding, understanding different folding conformations, and the folding rates also matter [43]. Several mechanisms have been proposed to explain the protein folding process, from a simple two-state model [44] to more complex models [45]. Also, it has been observed that some proteins do not fold and exist in an intrinsically disordered state [46].

Additionally, the misfolding of protein also occurs and has been observed in neurodegenerative disorders [47]. Thus, gathering the information on the folded and unfolded states is not enough, but the intermediate, misfolded, and disordered should also be analyzed. MSM uses the MD simulation data to find transition probability between different finite states. Initially, the model is constructed using geometric conformation similarity [48, 49]. The obvious choice is to use RMSD between the conformations by limiting it to a smaller cut-off value [50]. However, the RMSD is based on a protein backbone and is used to generate distance metrics. Hence, side chain and dihedral angle flip may hinder the results. The assumption is that the conformations with smaller deflections may have similar kinetic stability. However, finding more kinetically relevant metastable states should be carried out. Different clustering algorithms have been used [51], such as k-centers clustering, k-medoids clustering, and a hybrid of both k-clustering methods. The k-center clustering algorithm aims to find clusters with approximately the same radius and map different conformations to the nearest center of the cluster so that the distance from a distance is minimum. Li et al. & Voelz et al. used this clustering algorithm to improve the microstate generation efficiently [29, 52]. In the case of k-medoid clustering, the optimization is performed for the average distance between the center and other cluster points. In protein folding, this algorithm creates many clusters in the folded scenario and very few in the unfolded system [53]. The hybrid approach of both the k-clustering techniques was used to build MSMBuilder2 [54].

3.3 Protein–Ligand Binding

Analyzing the interaction of a protein with its substrate/inhibitor can provide critical information about the protein's function [5, 55]. The binding of small molecules to proteins or detecting new binding sites could be performed using MSM methods. Earlier, binding kinetics has been studied by constructing MSM to find long-lived intermediates of trypsin inhibitors [56]. The induced fit model (conformation changes due to ligand binding) and conformation selection model (ligand bind to protein without changing in protein's conformation) are used to detect protein–ligand recognition [57–59]. But later, it was observed that both are found in real-life scenarios [60–62]. In an earlier study to find the contribution of both methods, an analytical model based on a three-pronged approach of MD simulation, flux, and MSM was developed [63]. The choline-binding protein (ChoX) was used as a case study, and MD and MSM methods were used to find parameters for flux analysis [61].

3.4 Analyzing Intrinsically Disordered Proteins

Intrinsically Disordered Proteins (IDPs) are proteins that do not have a stable 3D structure. They bind to nucleic acids or other proteins for their functions. IDPs are dynamic ensembles that continuously change their internal conformation with high structural heterogeneity [64, 65]. However, IDPs are responsible for several cellular functions and are involved in many diseases like diabetes, cancer, neurodegenerative diseases, and cardiovascular diseases [66–69]. While interacting with partners, IDPs are coupled binding and folding reactions, which is essential for their function. Similar to ligand binding, induced fit, conformational selection, and a combination of both models are used to study IDPs. However, the kinetics of the binding-folding reaction, specifically binding to a partner or conformation without a partner, requires detailed investigation [70, 71]. Here, MD simulation can provide a contemporary way to analyze IDP folding at the atomistic level. To achieve this, MD simulations of IDPs should be performed so that the whole binding-folding pathways can be analyzed. Such simulation trajectories are complex to study; however, MSM techniques can help to identify metastable states in the pathway and the transition probability [72, 73].

3.5 Native State Conformation Changes

Generally, the rational structure-based drug design does not take into account protein-conformational changes. Approximately 15% of proteins have deep active sites related to their activity [74]. Hence, conformational heterogeneity is essential to understand protein behavior. This could provide information on the novel active sites or transient catalytic sites, which are allosteric or can block protein–protein interaction [75–78]. Since MD simulation can provide the system's dynamical behavior, if coupled with MSM, it can provide a set of ensembles where the metastable state is in an equilibrium state. Also, the advancement in MSM to capture kinetic and thermodynamic properties makes it a more viable option to identify the transient active site. There are several examples where similar approaches have been used to find cryptic pockets and allosteric sites. Among such studies, the TEM-1 beta-lactamase was used and observed that several such allosteric sites were present [79]. Such studies could also be performed with novel proteins to find active or allosteric sites.

4 Summary

Advancements in computational power, such as parallel programming and GPUs, have made the MD simulation more achievable. However, analyzing the simulation data is challenging. MSM is based on finite ensembles and uses clustering methods to create ensembles. Before MSM, geometric clustering was used, but MSM provides enhanced metastable states, which means it is the kinetic energy-based state. It is a coarse-graining of a system's dynamics, which depicts the underlying free energy landscape that governs the system's structure and dynamics. Identifying states in a kinetically relevant scheme and effectively using state decomposition to construct a transition matrix are the two main issues for creating an MSM. To build the MSM model, the traditional geometric clustering method is used to develop microstates. These microstates are further used to build a transition matrix. This step takes care of finding kinetically related microstates. This information is used to build MSM. However, adaptive sampling is used to improve the MSM model. Further, validations can be done by Bayesian Model selection, Swope–Pitera eigenvalue test, and other such tests (Fig. 4).

Protein folding and the dynamics of the native 3D structure are critical biological phenomena [80]. MD simulation can provide a way to understand these processes in millisecond simulations [81–83]; however, analyzing such data requires sophisticated protocols and methods [84, 85]. MSM provides a convenient and interpretable solution [86]. With the current advancement in computational power and algorithm, the use of MSM has increased and will continue to grow. This technique can also analyze and comprehend complicated systems such as membrane proteins, peptide

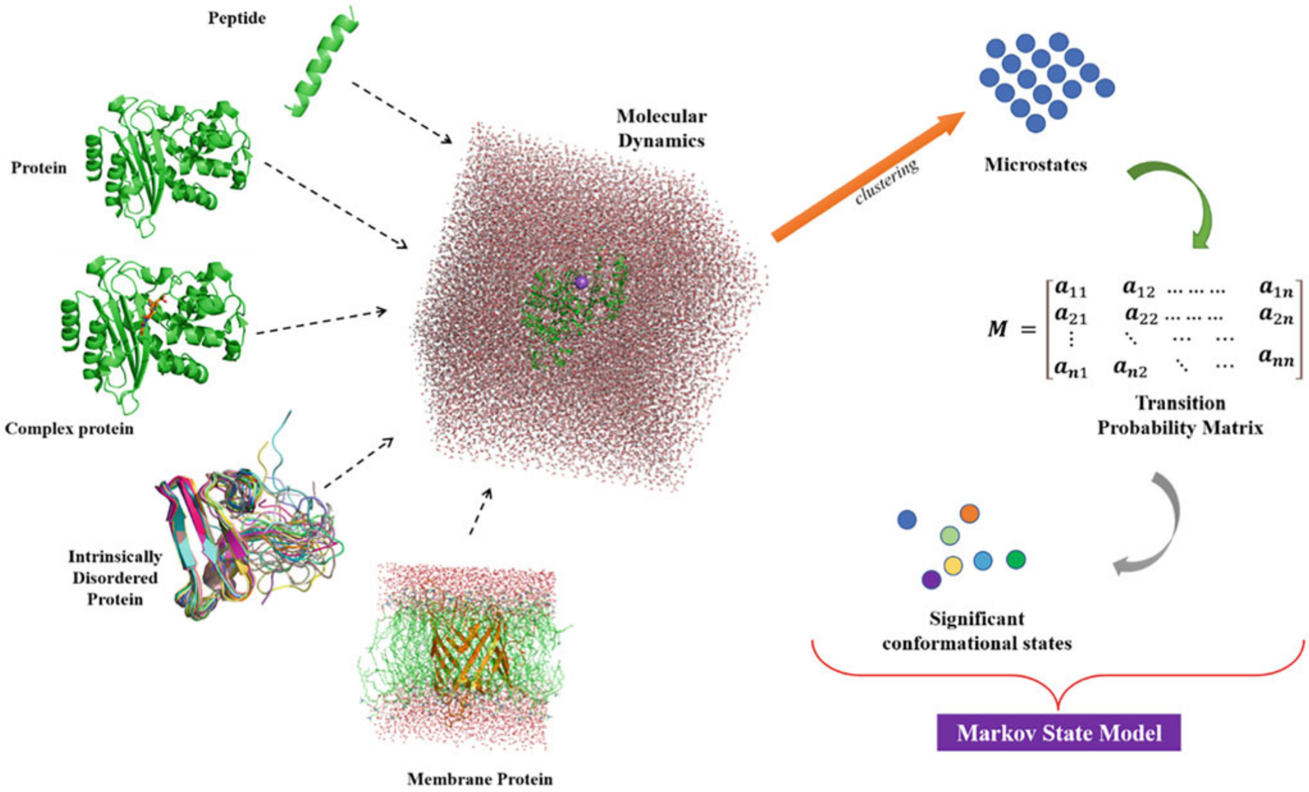
Peptide
Protein
Complex protein
Intrinsically Disordered Protein
Membrane Protein
Molecular Dynamics
clustering
Microstates
M = [a11 a12 ……… a1n; a21 a22 ……… a2n; ⋮ ⋱ … …; an1 an2 ⋱ … ann]
Transition Probability Matrix
Significant conformational states
Markov State Model

Fig. 4 Various applications which could be studied using Markov State Model

folding, IDPs, and other biological systems; hence, it is emerging as a critical in silico approach.

References

1. J. Gelpi, A. Hospital, R. Goñi, M. Orozco, Molecular dynamics simulations: advances and applications. Adv. Appl. Bioinforma. Chem. **8**, 37 (2015). https://doi.org/10.2147/AABC.S70333
2. J. Chodera, F. Noé, Markov state models of biomolecular conformational dynamics. Curr. Opin. Struct. Biol. **25**, 135–144 (2014)
3. S. Hazra, A. Szewczak, S. Ort, M. Konrad, A. Lavie, Post-translational phosphorylation of serine 74 of human deoxycytidine kinase favors the enzyme adopting the open conformation making it competent for nucleoside binding and release. Biochemistry **50**, 2870–2880 (2011). https://doi.org/10.1021/bi2001032
4. M.F. Chek, S.-Y. Kim, T. Mori, H.T. Tan, K. Sudesh, T. Hakoshima, Asymmetric open-closed dimer mechanism of polyhydroxyalkanoate synthase PhaC. IScience **23**, 101084 (2020). https://doi.org/10.1016/j.isci.2020.101084
5. S. Hazra, H. Xu, J.S. Blanchard, Tebipenem, a new Carbapenem antibiotic, is a slow substrate that inhibits the β-lactamase from *mycobacterium tuberculosis*. Biochemistry **53**, 3671–3678 (2014). https://doi.org/10.1021/bi500339j
6. M. Kokkinidis, N.M. Glykos, V.E. Fadouloglou, Protein flexibility and enzymatic catalysis. Adv. Protein. Chem. Struct. Biol. **87**, 181–218 (2012). https://doi.org/10.1016/B978-0-12-398312-1.00007-X
7. M. Karplus, Role of conformation transitions in adenylate kinase. Proc. Natl. Acad. Sci. U S A **107**, E71 (2010). https://doi.org/10.1073/pnas.1002180107
8. R.C. Stevens, W.N. Lipscomb, Allosteric control of quaternary states in E. coli aspartate transcarbamylase. Biochem. Biophys. Res. Commun. **171**, 1312–1318 (1990). https://doi.org/10.1016/0006-291X(90)90829-C
9. S. Bhattacharya, A.K. Padhi, V. Junghare, N. Das, D. Ghosh, P. Roy, K.Y.J. Zhang, S. Hazra, Understanding the molecular interactions of inhibitors against Bla1 beta-lactamase towards unraveling the mechanism of antimicrobial resistance. Int. J. Biol. Macromol. **177**, 337–350 (2021). https://doi.org/10.1016/j.ijbiomac.2021.02.069
10. S.A. Adcock, J.A. McCammon, Molecular dynamics: survey of methods for simulating the activity of proteins. Chem. Rev. **106**, 1589–1615 (2006). https://doi.org/10.1021/cr040426m
11. K.A. Henzler-Wildman, M. Lei, V. Thai, S.J. Kerns, M. Karplus, D. Kern, A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature **450**, 913–916 (2007). https://doi.org/10.1038/nature06407
12. S. Hammes-Schiffer, S.J. Benkovic, Relating protein motion to catalysis. Annu. Rev. Biochem. **75**, 519–541 (2006). https://doi.org/10.1146/ANNUREV.BIOCHEM.75.103004.142800
13. G.M. Lee, C.S. Craik, Trapping moving targets with small molecules. Science **324**(2009), 213–215 (1979). https://doi.org/10.1126/SCIENCE.1169378
14. S.J. Teague, Implications of protein flexibility for drug discovery. Nat. Rev. Drug Discov. **2**, 527–541 (2003). https://doi.org/10.1038/nrd1129
15. M. Pal, S. Bhattacharya, G. Kalyan, S. Hazra, Cadherin profiling for therapeutic interventions in epithelial Mesenchymal transition (EMT) and tumorigenesis. Exp. Cell Res. **368**, 137–146 (2018). https://doi.org/10.1016/j.yexcr.2018.04.014
16. G. Kalyan, V. Junghare, S. Bhattacharya, S. Hazra, Understanding structure-based dynamic interactions of antihypertensive peptides extracted from food sources. J. Biomol. Struct. Dyn. **39**, 635–649 (2021). https://doi.org/10.1080/07391102.2020.1715836

17. M.C. Baxa, E.J. Haddadian, J.M. Jumper, K.F. Freed, T.R. Sosnick, Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. Proc. Natl. Acad. Sci. **111**, 15396–15401 (2014). https://doi.org/10.1073/pnas.1407768111
18. A.N. Naganathan, M. Orozco, The native ensemble and folding of a protein molten-globule: Functional consequence of downhill folding. J. Am. Chem. Soc. **133**, 12154–12161 (2011). https://doi.org/10.1021/ja204053n
19. T.J. Lane, D. Shukla, K.A. Beauchamp, V.S. Pande, To milliseconds and beyond: challenges in the simulation of protein folding. Curr. Opin. Struct. Biol. **23**, 58–65 (2013). https://doi.org/10.1016/j.sbi.2012.11.002
20. S. Piana, J.L. Klepeis, D.E. Shaw, Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr. Opin. Struct. Biol. **24**, 98–105 (2014). https://doi.org/10.1016/j.sbi.2013.12.006
21. S. Bhattacharya, V. Junghare, N.K. Pandey, D. Ghosh, H. Patra, S. Hazra, An insight into the complete biophysical and biochemical characterization of novel class a beta-lactamase (Bla1) from bacillus anthracis. Int. J. Biol. Macromol. **145**, 510–526 (2020). https://doi.org/10.1016/j.ijbiomac.2019.12.136
22. S. Bhattacharya, V. Junghare, N.K. Pandey, S. Baidya, H. Agarwal, N. Das, A. Banerjee, D. Ghosh, P. Roy, H.K. Patra, S. Hazra, Variations in the SDN loop of class a beta-lactamases: a study of the molecular mechanism of BlaC (mycobacterium tuberculosis) to Alter the stability and catalytic activity towards antibiotic resistance of MBIs. Front. Microbiol. **12**, 710291 (2021). https://doi.org/10.3389/fmicb.2021.710291
23. K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold. Science **334**(2011), 517–520 (1979). https://doi.org/10.1126/science.1208351
24. Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. **314**, 141–151 (1999). https://doi.org/10.1016/S0009-2614(99)01123-9
25. F. Marinelli, F. Pietrucci, A. Laio, S. Piana, A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. PLoS Comput. Biol. **5**, e1000452 (2009). https://doi.org/10.1371/journal.pcbi.1000452
26. J. Juraszek, P.G. Bolhuis, Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. Proc. Natl. Acad. Sci. U S A **103**, 15859–15864 (2006). https://doi.org/10.1073/pnas.0606692103
27. Y. Duan, P.A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science **282**(1998), 740–744 (1979). https://doi.org/10.1126/SCIENCE.282.5389.740
28. S. Roy, V. Junghare, S. Dutta, S. Hazra, S. Basu, Differential binding of carbapenems with the AdeABC efflux pump and modulation of the expression of AdeB linked to novel mutations within two-component system AdeRS in carbapenem-resistant acinetobacter baumannii. mSystems **7**, e0021722 (2022). https://doi.org/10.1128/msystems.00217-22
29. V.A. Voelz, G.R. Bowman, K. Beauchamp, V.S. Pande, Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1−39). J. Am. Chem. Soc. **132**, 1526–1528 (2010). https://doi.org/10.1021/ja9090353
30. C. Hartmann, R. Banisch, M. Sarich, T. Badowski, C. Schütte, Characterization of rare events in molecular dynamics. Entropy **16**, 350–376 (2013). https://doi.org/10.3390/e16010350
31. V.S. Pande, K. Beauchamp, G.R. Bowman, Everything you wanted to know about Markov state models but were afraid to ask. Methods **52**, 99–105 (2010). https://doi.org/10.1016/j.ymeth.2010.06.002
32. G.B. Arfken, H.J. Weber, F.E. Harris, *Mathematical Methods for Physicists* (Elsevier, Amsterdam, 2013). https://doi.org/10.1016/C2009-0-30629-7
33. N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 2007). https://doi.org/10.1016/B978-0-444-52965-7.X5000-4

34. N. Singhal, C.D. Snow, V.S. Pande, Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. J. Chem. Phys. **121**, 415 (2004). https://doi.org/10.1063/1.1738647
35. W.C. Swope, J.W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B.G. Fitch, R.S. Germain, A. Rayshubski, T.J.C. Ward, Y. Zhestkov, R. Zhou, Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β-hairpin peptide. J. Phys. Chem. B **108**, 6582–6594 (2004). https://doi.org/10.1021/jp037422q
36. F. Noé, I. Horenko, C. Schütte, J.C. Smith, Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. J. Chem. Phys. **126**, 155102 (2007). https://doi.org/10.1063/1.2714539
37. F. Noé, S. Doose, I. Daidone, M. Löllmann, M. Sauer, J.D. Chodera, J.C. Smith, Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. Proc. Natl. Acad. Sci. U S A **108**, 4822–4827 (2011). https://doi.org/10.1073/pnas.1004646108
38. J.D. Chodera, N. Singhal, V.S. Pande, K.A. Dill, W.C. Swope, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. J. Chem. Phys. **126**, 155101 (2007). https://doi.org/10.1063/1.2714538
39. N.-V. Buchete, G. Hummer, Coarse master equations for peptide folding dynamics. J. Phys. Chem. B **112**, 6057–6069 (2008). https://doi.org/10.1021/jp0761665
40. A. Paul, S. Samantray, M. Anteghini, M. Khaled, B. Strodel, Thermodynamics and kinetics of the amyloid-β peptide revealed by Markov state models based on MD data in agreement with experiment. Chem. Sci. **12**, 6652–6669 (2021). https://doi.org/10.1039/D0SC04657D
41. C. Levinthal, How to fold graciously, in *Mossbauer Spectroscopy in Biological Systems*, ed. by P. DeBrunner, J. Tsibris, E. Munck, (University of Illinois Press, Urbana, 1969), pp. 22–26
42. H. White Frederick, J. Bello, D. Harker, E. de Jarnette, Regeneration of native secondary and tertiary structures by air oxidation of reduced Ribonuclease. J. Biol. Chem. **236**, 1353–1360 (1961). https://doi.org/10.1016/S0021-9258(18)64176-6
43. V.A. Voelz, V.R. Singh, W.J. Wedemeyer, L.J. Lapidus, V.S. Pande, Unfolded-state dynamics and structure of protein L characterized by simulation and experiment. J. Am. Chem. Soc. **132**, 4702–4709 (2010). https://doi.org/10.1021/ja908369h
44. J. Kubelka, T.K. Chiu, D.R. Davies, W.A. Eaton, J. Hofrichter, Sub-microsecond protein folding. J. Mol. Biol. **359**, 546–553 (2006). https://doi.org/10.1016/j.jmb.2006.03.034
45. P.S. Kim, R.L. Baldwin, Intermediates in the folding reactions of small proteins. Annu. Rev. Biochem. **59**, 631–660 (1990). https://doi.org/10.1146/annurev.bi.59.070190.003215
46. H.J. Dyson, P.E. Wright, Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol. **12**, 54–60 (2002). https://doi.org/10.1016/S0959-440X(02)00289-0
47. J. Hardy, D.J. Selkoe, The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science **297**(2002), 353–356 (1979). https://doi.org/10.1126/science.1072994
48. G.R. Bowman, X. Huang, V.S. Pande, Using generalized ensemble simulations and Markov state models to identify conformational states. Methods **49**, 197–201 (2009). https://doi.org/10.1016/j.ymeth.2009.04.013
49. M. Senne, B. Trendelkamp-Schroer, A.S.J.S. Mey, C. Schütte, F. Noé, EMMA: a software package for Markov model building and analysis. J. Chem. Theory Comput. **8**, 2223–2238 (2012). https://doi.org/10.1021/ct300274u
50. L.-T. Da, F.K. Sheong, D.-A. Silva, X. Huang, Application of Markov state models to simulate long timescale dynamics of biological macromolecules. Adv. Exp. Med. Biol. **805**, 29–66 (2014). https://doi.org/10.1007/978-3-319-02970-2_2
51. J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J.D. Chodera, C. Schütte, F. Noé, Markov models of molecular kinetics: generation and validation. J. Chem. Phys. **134**, 174105 (2011). https://doi.org/10.1063/1.3565032

52. Y. Li, Z. Dong, Effect of clustering algorithm on establishing Markov state model for molecular dynamics simulations. J. Chem. Inf. Model. **56**, 1205–1215 (2016). https://doi.org/10.1021/acs.jcim.6b00181
53. G.R. Bowman, An overview and practical guide to building Markov state models. Adv. Exp. Med. Biol. **797**, 7–22 (2014). https://doi.org/10.1007/978-94-007-7606-7_2
54. K.A. Beauchamp, G.R. Bowman, T.J. Lane, L. Maibaum, I.S. Haque, V.S. Pande, MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. J. Chem. Theory Comput. **7**, 3412–3419 (2011). https://doi.org/10.1021/ct200463m
55. G. Kalyan, V. Junghare, M.F. Khan, S. Pal, S. Bhattacharya, S. Guha, K. Majumder, S. Chakrabarty, S. Hazra, Anti-hypertensive peptide predictor: a machine learning-empowered web server for prediction of food-derived peptides with potential angiotensin-converting enzyme-I inhibitory activity. J. Agric. Food Chem. **69**, 14995–15004 (2021). https://doi.org/10.1021/acs.jafc.1c04555
56. U. Kahler, A.S. Kamenik, F. Waibl, J. Kraml, K.R. Liedl, Protein-protein binding as a two-step mechanism: preselection of encounter poses during the binding of BPTI and trypsin. Biophys. J. **119**, 652–666 (2020). https://doi.org/10.1016/j.bpj.2020.06.032
57. B. Ma, M. Shatsky, H.J. Wolfson, R. Nussinov, Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein Sci. **11**, 184–197 (2002). https://doi.org/10.1110/ps.21302
58. D.E. Koshland, Application of a theory of enzyme specificity to protein synthesis. Proc. Natl. Acad. Sci. U S A **44**, 98–104 (1958). https://doi.org/10.1073/pnas.44.2.98
59. V. Junghare, R. Alex, A. Baidya, M. Paul, R.R. Alyethodi, G.S. Sengar, S. Kumar, U. Singh, R. Deb, S. Hazra, *In silico* modeling revealed new insights into the mechanism of action of enzyme 2′-5′-oligoadenylate synthetase in cattle. J. Biomol. Struct. Dyn. **40**, 14013–14026 (2021). https://doi.org/10.1080/07391102.2021.2001373
60. H.-X. Zhou, From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. Biophys. J. **98**, L15–L17 (2010). https://doi.org/10.1016/j.bpj.2009.11.029
61. S. Gu, D.-A. Silva, L. Meng, A. Yue, X. Huang, Quantitatively characterizing the ligand binding mechanisms of choline binding protein using Markov state model analysis. PLoS Comput. Biol. **10**, e1003767 (2014). https://doi.org/10.1371/journal.pcbi.1003767
62. M.S. Formaneck, L. Ma, Q. Cui, Reconciling the "old" and "new" views of protein allostery: a molecular simulation study of chemotaxis Y protein (CheY). Proteins **63**, 846–867 (2006). https://doi.org/10.1002/prot.20893
63. G.G. Hammes, Y.-C. Chang, T.G. Oas, Conformational selection or induced fit: a flux description of reaction mechanism. Proc. Natl. Acad. Sci. **106**, 13737–13741 (2009). https://doi.org/10.1073/pnas.0907195106
64. P. Tompa, Intrinsically disordered proteins: a 10-year recap. Trends Biochem. Sci. **37**, 509–516 (2012). https://doi.org/10.1016/j.tibs.2012.08.004
65. V.N. Uversky, Dancing protein clouds: the strange biology and chaotic physics of intrinsically disordered proteins. J. Biol. Chem. **291**, 6681–6688 (2016). https://doi.org/10.1074/jbc.R115.685859
66. H. Xie, S. Vucetic, L.M. Iakoucheva, C.J. Oldfield, A.K. Dunker, Z. Obradovic, V.N. Uversky, Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J. Proteome Res. **6**, 1917–1932 (2007). https://doi.org/10.1021/pr060394e
67. V.N. Uversky, C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human diseases: introducing the D^2 concept. Annu. Rev. Biophys. **37**, 215–246 (2008). https://doi.org/10.1146/annurev.biophys.37.032807.125924
68. V.N. Uversky, Intrinsic disorder-based protein interactions and their modulators. Curr. Pharm. Des. **19**, 4191–4213 (2013). https://doi.org/10.2174/1381612811319230005
69. S.J. Metallo, Intrinsically disordered proteins are potential drug targets. Curr. Opin. Chem. Biol. **14**, 481–488 (2010). https://doi.org/10.1016/j.cbpa.2010.06.169

70. L. Mollica, L.M. Bessa, X. Hanoulle, M.R. Jensen, M. Blackledge, R. Schneider, Binding mechanisms of intrinsically disordered proteins: theory, simulation, and experiment. Front. Mol. Biosci. **3**, 52 (2016). https://doi.org/10.3389/fmolb.2016.00052
71. T. Kiefhaber, A. Bachmann, K.S. Jensen, Dynamics and mechanisms of coupled protein folding and binding reactions. Curr. Opin. Struct. Biol. **22**, 21–29 (2012). https://doi.org/10.1016/j.sbi.2011.09.010
72. J.C. Ezerski, P. Zhang, N.C. Jennings, M.N. Waxham, M.S. Cheung, Molecular dynamics ensemble refinement of intrinsically disordered peptides according to deconvoluted spectra from circular dichroism. Biophys. J. **118**, 1665–1678 (2020). https://doi.org/10.1016/j.bpj.2020.02.015
73. G. Pérez-Hernández, F. Paul, T. Giorgino, G. de Fabritiis, F. Noé, Identification of slow molecular order parameters for Markov model construction. J. Chem. Phys. **139**, 015102 (2013). https://doi.org/10.1063/1.4811489
74. A.L. Hopkins, C.R. Groom, The druggable genome. Nat. Rev. Drug Discov. **1**, 727–730 (2002). https://doi.org/10.1038/nrd892
75. M.R. Arkin, M. Randal, W.L. DeLano, J. Hyde, T.N. Luong, J.D. Oslob, D.R. Raphael, L. Taylor, J. Wang, R.S. McDowell, J.A. Wells, A.C. Braisted, Binding of small molecules to an adaptive protein-protein interface. Proc. Natl. Acad. Sci. U S A **100**, 1603–1608 (2003). https://doi.org/10.1073/PNAS.252756299
76. D.F. Ceccarelli, X. Tang, B. Pelletier, S. Orlicky, W. Xie, V. Plantevin, D. Neculai, Y.C. Chou, A. Ogunjimi, A. Al-Hakim, X. Varelas, J. Koszela, G.A. Wasney, M. Vedadi, S. Dhe-Paganon, S. Cox, S. Xu, A. Lopez-Girona, F. Mercurio, J. Wrana, D. Durocher, S. Meloche, D.R. Webb, M. Tyers, F. Sicheri, An allosteric inhibitor of the human Cdc34 ubiquitin-conjugating enzyme. Cell **145**, 1075–1087 (2011). https://doi.org/10.1016/J.CELL.2011.05.039
77. J.A. Hardy, J.A. Wells, Searching for new allosteric sites in enzymes. Curr. Opin. Struct. Biol. **14**, 706–715 (2004). https://doi.org/10.1016/J.SBI.2004.10.009
78. J.R. Horn, B.K. Shoichet, Allosteric inhibition through core disruption. J. Mol. Biol. **336**, 1283–1291 (2004). https://doi.org/10.1016/J.JMB.2003.12.068
79. G.R. Bowman, P.L. Geissler, Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proc. Natl. Acad. Sci. U S A **109**, 11681–11686 (2012). https://doi.org/10.1073/PNAS.1209309109/SUPPL_FILE/PNAS.1209309109_SI.PDF
80. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Springer Nature, Singapore, 2020)
81. R. Shukla, T. Tripathi, Molecular dynamics simulation of protein and protein-ligand complexes, in *Computer-Aided Drug Design*, ed. by D.B. Singh, (Springer Nature, Singapore, 2020), pp. 133–161
82. R. Shukla, T. Tripathi, Molecular dynamics simulation in drug discovery: opportunities and challenges, in *Innovations and Implementations of Drug Discovery Strategies in Rational Drug Design*, ed. by S.K. Singh, (Springer Nature, Singapore, 2021), pp. 295–316
83. K. Prince, S. Sasidharan, N. Nag, T. Tripathi, P. Saudagar, Integration of spectroscopic and computational data to analyze protein structure, function, folding, and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 483–502
84. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods*, 1st edn. (Academic Press, Cambridge, MA, 2022)
85. P. Saudagar, T. Tripathi, *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, 1st edn. (Academic Press, San Diego, 2023)
86. D. Ensign, P. Kasson, V.S. Pande, Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. J. Mol. Biol. **374**(3), 806–816 (2007)

Enhanced Sampling and Free Energy Methods to Study Protein Folding and Dynamics

Muthuraja Arun Pravin and Sanjeev Kumar Singh

Abstract A virtual study of the physical and chemical behaviour of particles in the energy space is referred to as computer simulation. The interaction of biomolecules and atoms during conformational changes is studied through molecular dynamics (MD) simulation. MD simulation complements the experimental results by providing a theoretical perspective of the real-time environment. However, the sampling of configuration is limited to a definite timescale due to free energy barriers. This free energy barrier arises due to the energy gap between initial and closing entropy in biomolecular structural transition. To deal with this biophysical problem, various enhanced sampling methods have been developed that are classified into collective variable-based and collective variable-free approaches based on the algorithm of the sampling method. This chapter discusses the numerical aspects of sampling methods, followed by a review of some of the most commonly used techniques in MD simulation and enhanced sampling. Lastly, a combined enhanced sampling method has been discussed.

Keywords Molecular dynamics · Collective variables · Free energy calculation · Accelerated molecular dynamics · Metadynamics · Umbrella sampling

1 Introduction

Proteins are a major component of all living organisms and play pivotal roles in biological processes such as enzymatic reaction, replication of nucleic acids, cellular organization, stimuli, and carrying molecules within and outside the cell. The structure of the protein is made up of a linear sequence of amino acids. This amino acid sequence gives protein the native conformation and folding in the cellular environment [1, 2]. The arrangement of the tertiary structure of proteins determines their location in the cell and overall function. Although proteins are

M. A. Pravin · S. K. Singh (✉)
Department of Bioinformatics, Alagappa University, Karaikudi, Tamilnadu, India

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_9

visualized and represented in a static form, their actual functions are dictated by their dynamic nature. To study the dynamic nature of proteins, computer-based simulation methods were developed to study the movements of atoms and molecules [3]. Since the first simulation of biomolecules, both power and methodology have been improved tremendously, which helped researchers to solve complex problems such as diseases related to protein misfolding and structural changes due to mutations [4].

A molecular dynamics (MD) simulation involves the movement of atoms in a biological system for a particular period of time to study real-time dynamics. By comparing MD trajectories with experimental results such as neutron scattering and nuclear magnetic resonance (NMR), one can get information on the dynamic properties of biomolecules [4, 5]. In normal-sized systems, MD simulations take a few microseconds, and in larger systems, they are even shorter. Simulations using MD are only reliable if they last long enough to cover all relevant components [6]. Due to unexplored regions in configurational space, most MD trajectories are not ergodic. In MD, timescales often differ from biological processes at the cellular level [7]. As a result, we get inadequate sampling and convergent simulations of biomolecular systems, which also leads to unreliable calculations of free energies. Since the mid-1980s, numerous enhanced sampling methods have been tested to resolve the problem of insufficient sampling and conformational studies [8]. Biomolecular conformational changes have historically been studied using molecular simulations as a method of sampling [9]. However, biomolecules exhibit rough energy landscapes with high energy barriers [10], which can lead to a nonfunctional state affecting conventional simulation. Protein activity that depends on large conformational changes, such as catalysis, is characterized by a large amplitude [5]. When transporting through a membrane, transporters undergo large conformational changes as they gate substrates [11]. The conventional MD simulation method cannot be used for long-scale simulation. An enhanced sampling approach has been developed to resolve free energy barriers in protein dynamics [12]. Conventional simulation methods have been unreliable with energy calculations [13]. Therefore, advanced sampling methods have been developed to study the flexibility of protein at its true biological level [14]. This chapter describes the theoretical understanding of enhanced sampling methods and their application in biomolecular simulations [15, 16]. Sampling methods are categorized into two broad classes: one that adds bias potentials to predefined collective variables (CVs) (CV-based) and one that does not involve CVs (CV-free). Further, a brief overview of algorithms used in various sampling methods is discussed.

2 Protein Folding and Dynamics

Protein folding is a biological process which determines the protein structure and subsequent function [17]. Protein folding is a unimolecular reaction that occurs between microseconds to hours at room temperature [18]. Thus, understanding the physicochemical process involved in protein folding and dynamics is essential

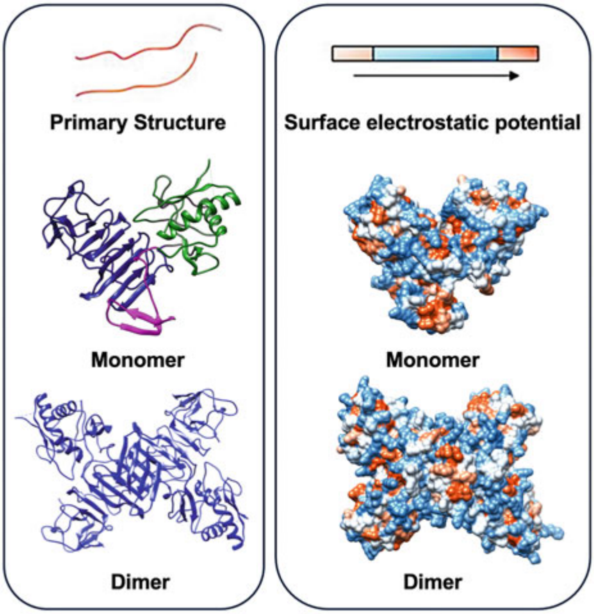

Fig. 1 Cartoon and electrostatic surface representation of the monomeric and dimeric forms of a protein

[19]. The major physiochemical process which underlines protein folding are secondary structure formation and kinetics of the protein folding [20]. The kinetics and free energy are studied through the protein energy landscape funnel, which illustrates the enthalpy and entropy of the folding process [15]. A complete understanding of protein folding requires governing all the factors, including conformational states of protein in the presence of water at varying temperatures [21]. Protein and water molecules form hydrogen bonds, which correlates with the hydrophobic effect. The hydrophobicity of a protein depends on surface charges, as shown in Fig. 1. The folding of protein involves an ensemble of structures with a small number of uniquely defined structural intermediates [22]. These ensemble structures are very crucial in understanding the process of protein folding [21]. However, due to limitations in computer simulations, major structural intermediates are not sampled well [23]. Therefore, methods for calculating free energy using enhanced sampling were discovered to assist in solving the problem of studying protein folding and dynamics [24]. Enhanced sampling techniques generally increase sampling efficiency. In order to modify the effective temperature, bias potentials are introduced, and the potential energy is modified [25]. This chapter briefly discusses sampling and free energy calculations using enhanced sampling. It also examines methods based on collective variables, such as metadynamics and steered enhanced sampling [26, 27]. We provide not only the hypothetical perspective but also their numerical implementation and projection for advancement in enhanced sampling methods.

3 Free Energy and Sampling Methods

Molecular dynamics is an efficient technique to sample the conformational space of a system with ergodic behaviour where all the important configurations can be accessed. While sampling the phase space of a quasi-non-ergodic system, one needs to analyse the nature of the initial and closing state of simulation in order to compute the change in free energy of intermediate states. In thermodynamics, the free energy of a system is called Helmholtz energy or Gibb's free energy of the system. The free energy calculation consists of the following elements:

1. A Hamiltonian distribution model.
2. An enhanced sampling method.
3. A method to estimate the thermodynamics of a system.

The Gibbs energy of a system in a constant volume and temperature (NVT) ensemble is given by

$$G = \frac{1}{\beta} \ln Q \tag{1}$$

$$\beta = \frac{1}{k_{\mathrm{\beta}}} T \tag{2}$$

where Q is the partition function of the system.

The partition function provides the number of states that are accessible to a particular temperature. A protein may have N number of dynamic states and energy at different conformations. The probability of finding the system in the state K is given by the Boltzmann distribution

$$P_{\mathrm{K}} = e \frac{-\beta EK}{\sum_k} e^{-\beta EK} \tag{3}$$

$$Q = \sum_k e^{-\beta EK} \tag{4}$$

For a macroscopic system in thermodynamic equilibrium (no energy flow along the system)

$$Q = \frac{1}{h^{3n}} N! \iint e^{-\beta H(p,q)} \, dpdr \tag{5}$$

where h is the Planck constant, N is the number of particles, and H is the Hamiltonian describing the total energy of the system.

$$H(p,q) = \frac{\sum (i=0)^{NP_{\mathrm{i}}^2}}{2m} + v(q) \tag{6}$$

where V is the potential, which is given by a force field in classical MD.

For complex systems, an analytical expression for Q (Eq. 5) cannot be derived due to the high free energy barrier in the intermediate state. Thus, enhanced sampling techniques are used to study the intermediate states of a protein. The enhanced sampling has been categorized into CV-based sampling and CV-free sampling.

3.1 Collective Variables and Free Energy

MD simulations are used for the in silico studies of the dynamic nature of biological molecules using atomic coordinates. Since atomic coordinates are associated with a dimensional problem, collective variables (CVs) are introduced to quantify the particular property of a simulated system. CVs are atomic coordinate functions used to describe certain motions or transitions. For example, the study of the atomic distance is given by CV, which can represent the bond formation and arrangement [28]. For different properties of a system, CV can be used to describe the system more efficiently.

The collective variable of a system at a given point is given by s_{i} and is equal to the configuration of q mapping to s_{i}.

$$P(s_{\mathrm{i}}) = \langle \delta[s(q) - s] \rangle \tag{7}$$

The delta function takes s_{i} considering all possibilities $s(q)$, and the bracket $\langle\rangle$ represents the all-canonical ensemble. The probability of obtaining free energy is given by

$$G(S_{\mathrm{i}}) = k_{\mathrm{B}}T \;\ln\langle \mathrm{\delta}[s(q) - \mathrm{s}] \rangle \tag{8}$$

where K_{B} represents the Boltzmann constant and T denotes temperature.

Based on the Gibbs free energy $G(s)$, the transition state is calculated by moving from one CV region to the adjacent one, which is represented in Fig. 2. The figure illustrates the energy landscape and the state of transition from A to B, with C as the intermediate state. The overall transition rate depends on the level of the free energy barrier compared to the thermal energy $K_{\mathrm{B}}T$, according to the Arrhenius equation.

$$V_{A\to B} = v_0 \;\exp\left(\frac{-\Delta G^{\ddagger}}{K_{\mathrm{B}}T}\right) \tag{9}$$

where V_0 is a prefactor and $\Delta G^{\ddagger}$ is the free energy difference between state A and the transition state.

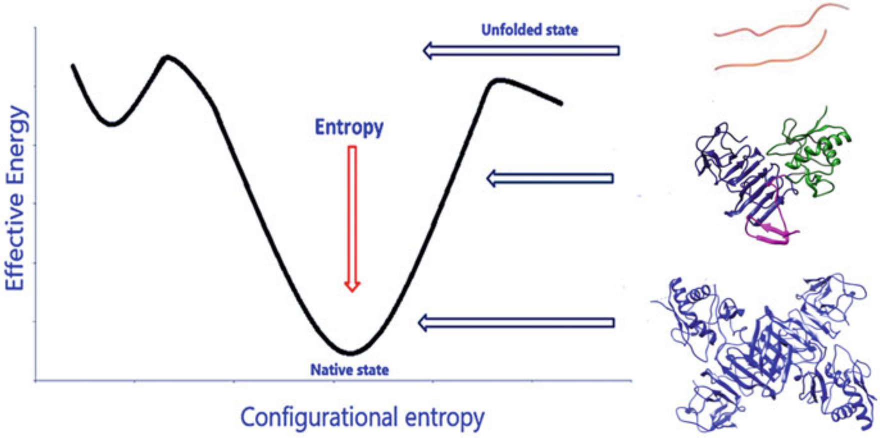

Fig. 2 A folding funnel representing the thermodynamics of protein folding during translation

$$\Delta G^{\ddagger} = G_{\mathrm{C}} - G_{\mathrm{A}} = -K_{\mathrm{B}} T \ln \frac{\langle \delta[s(q) - s_{\mathrm{C}}] \rangle}{\langle \delta[s(q) - s_{\mathrm{A}}] \rangle} \tag{10}$$

A complete configuration of CV space is possible in a perfectly ergodic system. The reaction coordinate suggests that there is enough transition which results in the crossing of *A* and *B*. The free energy can be calculated in the histogram into a probability $P(s_{\mathrm{i}})$ via Eqs. (8) and (9). The histograms are used to estimate the differences in free energy between discrete states. It is possible to calculate thermodynamic components of energies using these histograms [29]. Therefore, CV-based methods are essential for enhanced sampling in an MD environment. To study the enthalpy and entropy of the system with varying energies, the sampling technique is used along with MD simulation (Fig. 3).

4 Collective Variable-Based Sampling

This method involves the addition of bias to the numerical function. It is used to calculate the potential energy of unaccessed regions in the energy landscape [30]. The bias selection is much more important than the overall prediction, as the bias determines the efficiency of sampling [31]. There are two widely used CV-based methods discussed in this segment.

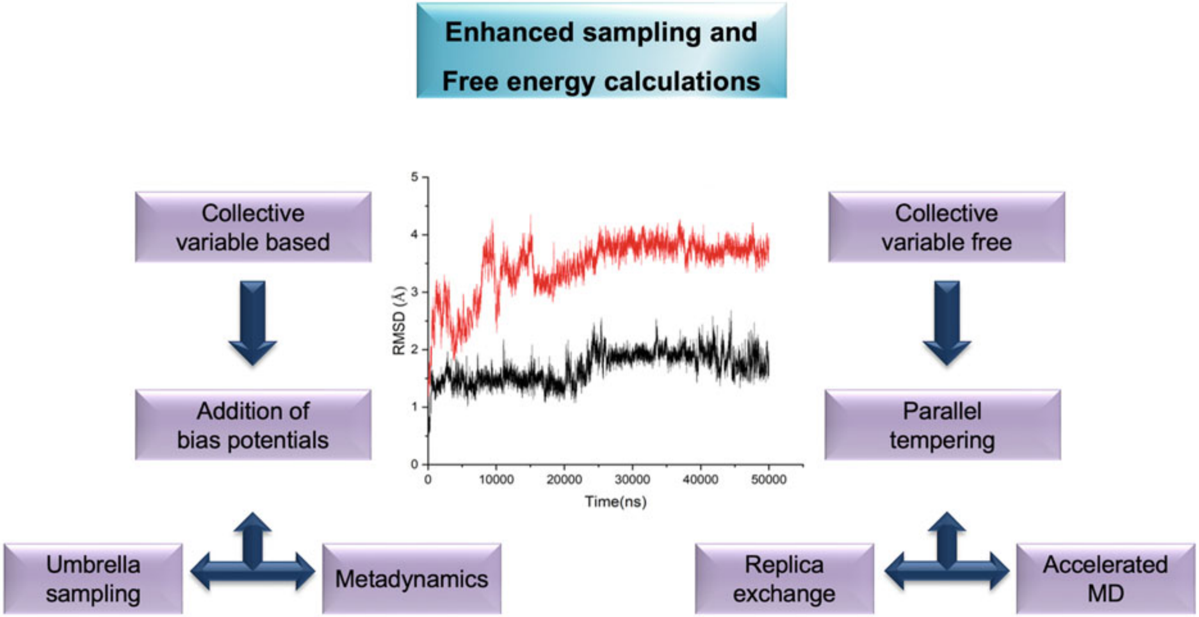

Fig. 3 Enhanced sampling and free energy calculation techniques in molecular dynamics

4.1 Umbrella Sampling

Umbrella sampling is a widely used MD technique in computational physics and chemistry. It is used to enhance the sampling of different systems where dynamic equilibrium is hindered by the free energy landscape [32, 33]. To calculate accurate thermodynamics data with a sufficient range of configurational changes, the umbrella sampling method has been adopted to overcome the potential barrier in free energy landscape study during protein folding dynamics. It was first proposed by Torrie and Valleau [34]. The thermodynamics of a system that involves low and high energy barriers is poorly sampled by conventional Monte Carlo methods, which can leave crucial confirmation unsampled in a dynamic simulation [35]. Umbrella sampling helps bridge the gap between low and high energy barriers in a simulation study [36, 37]. This method involves introducing a biased potential in the existing Hamiltonian sampling distribution. In order to calculate the system's Hamiltonian, a simple harmonic potential $\Delta V_{\mathrm{i}}\ (q)$, which is determined by the force constant k, is added to each window.

$$\Delta V_{\mathrm{i}}\ (q) = \frac{K}{2*}\ (s(q) - si)^2 \tag{11}$$

With its fast convergence and ability to run simulations independently of each other, umbrella sampling has emerged as one of the most successful tools for improving convergence [38]. A reasonable overlapping position between each window alongside the CVs is obtained by tuning the harmonic potentials manually for each window so that the harmonic potentials for all the windows are tuned [39]. Computationally it is time-consuming and challenging to determine increased CVs or complexity of the system of interest [40, 41]. To further improve sampling efficiency, the sampling methods can be combined to obtain the desired result.

4.2 Metadynamics

To accelerate the rare event sampling, Parrinello developed Metadynamics (MetaD). In MetaD simulation, a system is subjected to an external CV-dependent bias potential [42]. The Gaussians, along with the CV space, are added to help the system visit configurations that have not been tested [43]. Bias potentials of Gaussian types are defined as $(s(q),\ t)$, where τ is the rate of Gaussian deposition, σ_{i} and $W(k_{\tau})$ is referred to as the width and height of the Gaussian and time k_{τ} of simulation. An increased positive bias potential encourages to explore configurations that are not explored in the CVs space, resulting in a system escaping the local minimum of the CVs space. Eventually, CVs predict a bias potential convergent to negative free energy. In addition to the added bias potential, the high sampling efficiency allows us to easily traverse the energy barriers that separate different local minima

[44]. During a standard MetaD simulation, the bias potential of Gaussian remains constant. Consequently, using the bias potential, the landscape of free energy is analysed. In addition, the system may be driven into physically irrelevant phases of phase space if it oscillates around the real values.

$$V\lambda\ (q,t) = \frac{K(t)}{2\ (s(q,t)-\lambda(t))^2} = \frac{(k(t))^2}{2\ (s(q,t)-s_0-vt)^2} \tag{12}$$

where $K(t)$, V denotes harmonic potential and pulling speed, $s(q)$ is used to define the CVs of pulling direction, and for the correlation of $s(q)$, λ is used as a parameter. The free energy calculation in SMD simulation between two states, i and j, can be calculated by work $W_{\mathrm{i}\,\to\,\mathrm{j}}$. Using the Jarzynski equation, the work done by the system is calculated as

$$\Delta G = -\beta^{-1}\ \ln\ \big\langle \exp\big(-\beta W_{\mathrm{i-j}}\big)\big\rangle 0 \tag{13}$$

Later on, Crooks proposed a new version

$$\Delta G = -\beta^{-1}\ \ln \frac{\langle \int \big(W_{\mathrm{i}\to\mathrm{j}}\ \big)\rangle \mathrm{I}}{\langle \int \big(W_{\mathrm{i}\to\mathrm{j}}\big)\rangle \mathrm{j}} \tag{14}$$

where $\int(W_{\mathrm{i}\,\to\,\mathrm{j}})$ defines the finite function of work. The proposed Crooks equation is utilized for deriving the bars equation. Apart from metaD, steered molecular dynamics (SMD) is extensively used for protein conformation and ligand binding study. SMD can be used for sampling CV space and involves shorter simulation time [45]. Therefore, it is used along with other sampling methods, such as umbrella sampling and MetaD simulations.

5 Collective Variable-Free Sampling

Enhanced sampling algorithms based on CVs can dramatically broaden the timescale of MD simulations. Though CVs are necessary for these algorithms to accurately represent biological events, there are certain limitations in calculating hidden layers of energies [46]. In MD simulations, determining the optimal CVs is not trivial when processes are complex, especially when the transition processes are complex [47]. Thus, CV-free sampling can be used to solve hidden barriers issues in the CV-based biased sampling method. Some methods, such as accelerated MD and replica exchange molecular dynamics (REMD), are often used to resolve CV-related problems [48]. In this section, an overview of CV-free enhanced sampling methods has been discussed.

5.1 Replica Exchange Molecular Dynamics

To study the conformational of various states of protein, replica exchange molecular dynamics (REMD) was developed [49]. It is also called parallel tempering (PT) and involves independent replicas of the system of interest. These replicas are parallel simulated at different temperatures. REMD consists of the exchange of replicas at regular intervals of time. The exchange of replicas occurs when the condition is satisfied according to the Metropolis criterion.

$$P\left(q_{\mathrm{i}} \leftrightarrow q_{\mathrm{j}}\right) = \min\left(1, \exp\left[\left(\frac{1}{K_{\mathrm{B}}}T_{\mathrm{j}} - \frac{1}{K_{\mathrm{B}}}T_{\mathrm{i}}\right)\left[H(q_i) - H(q_i)\right]\right]\right) \tag{15}$$

where $H(q_\mathrm{i})$ and $H(q_\mathrm{i})$ represent potential energies of replica i, j and T_i, T_j represents the temperature of i and j.

For the calculation of free energy landscapes, configurations are replicated at low temperatures. As the number of replicas increases, the simulation becomes larger, and temperature REMD (T-REMD) requires higher computational power. In T-REMD, the MD simulation **has** a temperature difference. There are certain parameters, such as the Hamiltonian system of different replicas changes with temperature, which is then expressed as

$$P\left(q_{\mathrm{i}} \leftrightarrow q_{\mathrm{j}}\right) = \min\left[1, \exp\left(\frac{H_{\mathrm{i}}(q_{\mathrm{i}}) - H_{\mathrm{i}}\left(q_{\mathrm{j}}\right)}{K_{\mathrm{B}}T_{\mathrm{i}}} + \frac{H_{\mathrm{j}}\left(q_{\mathrm{j}}\right) - H_{\mathrm{j}}(q_{\mathrm{i}})}{K_{\mathrm{B}}T_{\mathrm{j}}}\right)\right] \tag{16}$$

Various methods have been proposed to modify the Hamiltonian equation. In the T-REMD method altering solute gives good exchange probabilities, which only require a relatively small number of replicas. A sufficient overlap in energy contributions is needed for replicas to exchange successfully. REST2 had Ra much greater sampling efficiency than T-REMD [50]. Later, Bussi designed a new REST2 variation with improved scaling and additional flexibility in terms of which elements of the system the scaling is applied to the existing system [51]. An example of other computational approaches would be constant pH replica exchange which is frequently used with replica exchange.

5.2 Accelerated Molecular Dynamics

Accelerated molecular dynamics (aMD) is a sampling method developed by the McCammon group in 2004 to promote the bridging of energy barriers between various conformational states [52]. In this method, an enhanced potential is enforced to the existing potential function $V(r)$ to examine the potential energy of the system [53]. The boost potential $V^{*}(r)$ activates when the system's potential energy decreases below a threshold energy E [54]. The change in the potential function has helped eliminate the sampling barriers observed in traditional sampling techniques. Here the existing potential function is replaced with a negative function, which allows calculating the potential when a system has low threshold energy E. This method is excellent in identifying key regions of protein folding in a large-scale simulation. Overall sampling depends on the superiority of the true potential, which needs to be varied when the threshold energy reaches below a certain level. The aMD involves modifying potential $V^{*}(r) = V(r) + \Delta V(r)$. This method lowers the energy barrier while preserving key information about the potential energy landscape [55, 56]. Thus, the efficacy of aMD is in improving the biomolecular sampling of systems, which has been demonstrated in a wide range of applications involving interactions between proteins and peptides, small molecule and protein binding behaviour, and protein conformational changes. The simulation of systems includes dipeptides, membrane proteins, and globular proteins [57, 58]. To find the potential binding poses of protein–ligand docking, aMD was used [59]. In the previous section, we have described a few popular enhanced sampling approaches and conformational space sampling, which can be further optimized by combining methods from CV-based and CV-free approaches since these approaches differ in methodology.

6 Conclusion and Outlook

Numerous strategies have been developed in the past 20 years to increase the sampling methods in MD simulation and the calculation of energies [60–63]. They enhance the convergence of free energy calculations in addition to probing the conformational space of biomolecular systems. The theories, most recent advancements, and examples of three CV-based and two CV-free enhanced sampling approaches are covered in this chapter. While simulations can follow specified routes using a CV-based technique (such as US, MetaD, or SMD), they are not necessary with a CV-free method (such as REMD or aMD). These techniques can be used to execute an improved sampling biomolecular simulation without needing to adhere to a strict procedure. It is generally advisable to describe biological events of interest using US/MetaD. REMD (or HREX) or aMD is used whenever there is minimal knowledge about the process that needs to be mimicked (such as protein folding or investigating a dynamically disordered protein). In computational enzymology, the QM/MM (quantum mechanics/molecular mechanics) technique can also be used in conjunction with improved sampling and free energy calculations. In general, a large system with millions of atoms will need longer timescales and more computing power, and modelling them using REMD (including HREX) will further raise the computation cost for improved sampling and free energy calculations for macromolecular simulations. The complex CVs of large systems may prevent CV-based simulation techniques from being effective compared to how they perform in smaller systems. The timesteps are larger for all-atom simulation; the fastest degrees of freedom can be eliminated by employing virtual hydrogen sites or hydrogen mass repartition. Machine learning has gained a lot of popularity recently. More intelligent sampling strategies have emerged due to new, potent unsupervised and reinforced deep learning algorithms. We hope to see additional advancements made in this area and more bimolecular simulations using machine learning-related techniques.

Acknowledgements The authors acknowledge the research facility provided by Alagappa University. The authors also thank the Department of Biotechnology (DBT-BIC) Project Grant/Award (No.BT/PR40154/BTIS/137/34/2021), RUSA-Phase 2.0 Policy (TNmulti-Gen), Dept. of Edn, Govt. of India (Grant No: F.24-51/2014-U), Department of Biotechnology (DBT), New Delhi Grant/Award (No.BT/PR40154/BTIS/137/34/2021), The Higher Education, Govt. of Tamil Nadu for the Grant (No. 5594/H1/2020-1), The Tamil Nadu State Council for Higher Education (TANSCHE) for the research grant (Au/S.o. (P&D): TANSCHE Projects: 117/2021), DST-PURSE 2nd Phase Programme Order no. SR/PURSE phase 2/28 (G dated 21.02.2017) and FIST (SR/FST/LSI—667/2016) for financial support.

References

1. C.B. Anfinsen, Principles that govern the folding of protein chains. Science **181**(4096), 223–230 (1973)
2. P.D. Sun, C.E. Foster, J.C. Boyington, Overview of protein structural and functional folds. Curr. Protoc. Protein Sci. **Chapter 17**(1), Unit 17.1 (2004). https://doi.org/10.1002/0471140864.ps1701s35
3. S.K. Tripathi, C. Selvaraj, S.K. Singh, K.K. Reddy, Molecular docking, QPLD, and ADME prediction studies on HIV-1 integrase leads. Med. Chem. Res. **21**(12), 4239–4251 (2012)
4. J.C. Smith, G.R. Kneller, Combination of neutron scattering and molecular dynamics to determine internal motions in biomolecules. Mol. Simul. **10**(2–6), 363–375 (1993). https://doi.org/10.1080/08927029308022173
5. P. Vijayalakshmi, C. Selvaraj, S.K. Singh, J. Nisha, K. Saipriya, P. Daisy, Exploration of the binding of DNA binding ligands to Staphylococcal DNA through QM/MM docking and molecular dynamics simulation. J. Biomol. Struct. Dyn. **31**(6), 561–571 (2013)
6. E.I. Shakhnovich, Proteins with selected sequences fold into unique native conformation. Phys. Rev. Lett. **72**(24), 3907 (1994)
7. T. Komatsu et al., Real-time measurements of protein dynamics using fluorescence activation-coupled protein labeling method. J. Am. Chem. Soc. **133**(17), 6745–6751 (2011)
8. Y. Miao, J.A. McCammon, Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. Mol. Simul. **42**(13), 1046–1055 (2016). https://doi.org/10.1080/08927022.2015.1121541
9. X. Chu, Y. Wang, P. Tian, W. Li, D. Mercadante, Editorial: advanced sampling and modeling in molecular simulations for slow and large-scale biomolecular dynamics. Front. Mol. Biosci. **8**, 795991 (2021). https://doi.org/10.3389/fmolb.2021.795991
10. K. Röder, D.J. Wales, The energy landscape perspective: encoding structure and function for biomolecules. Front. Mol. Biosci. **9**, 820792 (2022)
11. S. Shaikh, P.-C. Wen, G. Enkavi, Z. Huang, E. Tajkhorshid, Capturing functional motions of membrane channels and transporters with molecular dynamics simulation. J. Comput. Theor. Nanosci. **7**(12), 2481–2500 (2010). https://doi.org/10.1166/jctn.2010.1636
12. A. Mitsutake, Y. Mori, Y. Okamoto, Enhanced sampling algorithms. Methods Mol. Biol. **924**, 153–195 (2013)

13. M.P. Allen, Introduction to molecular dynamics simulation. Comput. Soft Matter **23**(1), 1–28 (2004)
14. A. Barducci, M. Bonomi, M. Parrinello, Metadynamics. Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**(5), 826–843 (2011)
15. J.N. Onuchic, H. Nymeyer, A.E. García, J. Chahine, N.D. Socci, The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. Adv. Protein Chem. **53**, 87–152 (2000)
16. R. Khandelwal et al., Structure-based virtual screening for the identification of high-affinity small molecule towards STAT3 for the clinical treatment of osteosarcoma. Curr. Top. Med. Chem. **18**(29), 2511–2526 (2018)
17. C. Selvaraj, S.K. Singh, S.K. Tripathi, K.K. Reddy, M. Rama, In silico screening of indinavir-based compounds targeting proteolytic activity in HIV PR: binding pocket fit approach. Med. Chem. Res. **21**(12), 4060–4068 (2012). https://doi.org/10.1007/s00044-011-9941-5
18. D. Pradiba, M. Aarthy, V. Shunmugapriya, S.K. Singh, M. Vasanthi, Structural insights into the binding mode of flavonols with the active site of matrix metalloproteinase-9 through molecular docking and molecular dynamic simulations studies. J. Biomol. Struct. Dyn. **36**(14), 3718–3739 (2018). https://doi.org/10.1080/07391102.2017.1397058
19. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins. Nature **450**(7172), 964–972 (2007)
20. K. Patidar et al., Virtual screening approaches in identification of bioactive compounds akin to delphinidin as potential HER2 inhibitors for the treatment of breast cancer. Asian Pacific J. Cancer Prev. **17**(4), 2291–2295 (2016)
21. K.K. Reddy, S.K. Singh, Combined ligand and structure-based approaches on HIV-1 integrase strand transfer inhibitors. Chem. Biol. Interact. **218**, 71–81 (2014)
22. K. Patidar et al., An in silico approach to identify high affinity small molecule targeting m-TOR inhibitors for the clinical treatment of breast cancer. Asian Pacific J. Cancer Prev. **20**(4), 1229 (2019)
23. S.S. Plotkin, J.N. Onuchic, Understanding protein folding with energy landscape theory part I: basic concepts. Q. Rev. Biophys. **35**(2), 111–167 (2002)
24. S. Vijayakumar, P. Manogar, S. Prabhu, R.A. Sanjeevkumar Singh, Novel ligand-based docking; molecular dynamic simulations; and absorption, distribution, metabolism, and excretion approach to analyzing potential acetylcholinesterase inhibitors for Alzheimer's disease. J. Pharm. Anal. **8**(6), 413–420 (2018). https://doi.org/10.1016/j.jpha.2017.07.006
25. D.S. Malar, V. Suryanarayanan, M.I. Prasanth, S.K. Singh, K. Balamurugan, K.P. Devi, Vitexin inhibits Aβ25-35 induced toxicity in Neuro-2a cells by augmenting Nrf-2/HO-1 dependent antioxidant pathway and regulating lipid homeostasis by the activation of LXR-α. Toxicol. In Vitro **50**, 160–171 (2018). https://doi.org/10.1016/j.tiv.2018.03.003
26. Y. Inagaki, Generalized simulated annealing algorithms using Tsallis statistics: application to the discrete-time optimal growth problem. Rev. Econ. Bus. Admin. **37**(2), 1–11 (2007)
27. S. Sharda et al., A virtual screening approach for the identification of high affinity small molecules targeting BCR-ABL1 inhibitors for the treatment of chronic myeloid leukemia. Curr. Top. Med. Chem. **17**(26), 2989–2996 (2017)
28. S. Kirkpatrick, C.D. Gelatt Jr., M.P. Vecchi, Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
29. S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, P.A. Kollman, THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem. **13**(8), 1011–1021 (1992). https://doi.org/10.1002/jcc.540130812
30. H. Szu, R. Hartley, Fast simulated annealing. Phys. Lett. A **122**(3–4), 157–162 (1987)
31. H. Zang, S. Zhang, K. Hapeshi, A review of nature-inspired algorithms. J. Bionic Eng. **7**(4), S232–S237 (2010)
32. K. Hamacher, W. Wenzel, Scaling behavior of stochastic minimization algorithms in a perfect funnel landscape. Phys. Rev. E **59**(1), 938 (1999)

33. A. Laio, M. Parrinello, Escaping free-energy minima. Proc. Natl. Acad. Sci. **99**(20), 12562–12566 (2002). https://doi.org/10.1073/pnas.202427399
34. G.M. Torrie, J.P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J. Comput. Phys. **23**(2), 187–199 (1977). https://doi.org/10.1016/0021-9991(77)90121-8
35. N. Metropolis, S. Ulam, The Monte Carlo method. J. Am. Stat. Assoc. **44**(247), 335–341 (1949)
36. C. Chen, Y. Huang, Y. Xiao, Enhanced sampling of molecular dynamics simulation of peptides and proteins by double coupling to thermal bath. J. Biomol. Struct. Dyn. **31**(2), 206–214 (2013). https://doi.org/10.1080/07391102.2012.698244
37. C. Selvaraj et al., Structural insights into the binding mode of d-sorbitol with sorbitol dehydrogenase using QM-polarized ligand docking and molecular dynamics simulations. Biochem. Eng. J. **114**, 244–256 (2016). https://doi.org/10.1016/j.bej.2016.07.008
38. M.D. De Andrade, M.A.C. Nascimento, K.C. Mundim, A.M.C. Sobrinho, L.A.C. Malbouisson, Atomic basis sets optimization using the generalized simulated annealing approach: new basis sets for the first row elements. Int. J. Quantum Chem. **108**(13), 2486–2498 (2008)
39. K.C. Mundim, C. Tsallis, Geometry optimization and conformational analysis through generalized simulated annealing. Int. J. Quantum Chem. **58**(4), 373–381 (1996)
40. S.S. da Rocha Pita, T.V.A. Fernandes, E.R. Caffarena, P.G. Pascutti, Studies of molecular docking between fibroblast growth factor and heparin using generalized simulated annealing. Int. J. Quantum Chem. **108**(13), 2608–2614 (2008)
41. S. Bandaru et al., Identification of small molecule as a high affinity β2 agonist promiscuously targeting wild and mutated (Thr164Ile) β 2 adrenergic receptor in the treatment of bronchial asthma. Curr. Pharm. Des. **22**(34), 5221–5233 (2016)
42. M.C.R. Melo, R.C. Bernardi, T.V.A. Fernandes, P.G. Pascutti, GSAFold: a new application of GSA to protein structure prediction. Proteins Struct. Funct. Bioinform. **80**(9), 2305–2310 (2012)
43. M.A. Moret, P.M. Bisch, K.C. Mundim, P.G. Pascutti, New stochastic strategy to analyze helix folding. Biophys. J. **82**(3), 1123–1132 (2002)
44. M.A. Moret, P.G. Pascutti, P.M. Bisch, K.C. Mundim, Stochastic molecular optimization using generalized simulated annealing. J. Comput. Chem. **19**(6), 647–657 (1998)
45. Y. Xiang, X.G. Gong, Efficiency of generalized simulated annealing. Phys. Rev. E **62**(3), 4473 (2000)
46. C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. **52**(1), 479–487 (1988)
47. C. Abrams, G. Bussi, Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. Entropy **16**(1), 163–199 (2014)
48. G. Helles, A comparative study of the reported performance of ab initio protein structure prediction algorithms. J. R. Soc. Interface **5**(21), 387–396 (2008)
49. M. Majhi et al., An in silico investigation of potential EGFR inhibitors for the clinical treatment of colorectal cancer. Curr. Top. Med. Chem. **18**(27), 2355–2366 (2018)
50. L. Wang, R.A. Friesner, B.J. Berne, Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). J. Phys. Chem. B **115**(30), 9431–9438 (2011)
51. G. Bussi, Hamiltonian replica exchange in GROMACS: a flexible implementation. Mol. Phys. **112**(3–4), 379–384 (2014)
52. D. Hamelberg, J. Mongan, J.A. McCammon, Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J. Chem. Phys. **120**(24), 11919–11929 (2004)
53. L.-H. Hung, S.-C. Ngan, T. Liu, R. Samudrala, PROTINFO: new algorithms for enhanced protein structure predictions. Nucleic Acids Res. **33**(Suppl 2), W77–W80 (2005)
54. D.C. Rapaport, D.C.R. Rapaport, *The Art of Molecular Dynamics Simulation* (Cambridge University Press, Cambridge, 2004)
55. D. Hamelberg, J.A. McCammon, Fast peptidyl cis-trans isomerization within the flexible gly-rich flaps of HIV-1 protease. J. Am. Chem. Soc. **127**(40), 13778–13779 (2005)

56. P.R.L. Markwick, G. Bouvignies, M. Blackledge, Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. J. Am. Chem. Soc. **129**(15), 4724–4730 (2007)
57. B. Zhao, M.A. Cohen Stuart, C.K. Hall, Navigating in foldonia: using accelerated molecular dynamics to explore stability, unfolding and self-healing of the β-solenoid structure formed by a silk-like polypeptide. PLoS Comput. Biol. **13**(3), e1005446 (2017)
58. Y. Miao, W. Sinko, L. Pierce, D. Bucher, R.C. Walker, J.A. McCammon, Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. J. Chem. Theory Comput. **10**(7), 2677–2689 (2014)
59. S.F. Sousa, P.A. Fernandes, M.J. Ramos, Protein–ligand docking: current status and future challenges. Proteins Struct. Funct. Bioinform. **65**(1), 15–26 (2006)
60. R. Shukla, T. Tripathi, Molecular dynamics simulation of protein and protein-ligand complexes, in *Computer-Aided Drug Design*, ed. by D.B. Singh, (Springer Nature, Singapore, 2020), pp. 133–161
61. R. Shukla, T. Tripathi, Molecular dynamics simulation in drug discovery: opportunities and challenges, in *Innovations and Implementations of Drug Discovery Strategies in Rational Drug Design*, ed. by S.K. Singh, (Springer Nature, Singapore, 2021), pp. 295–316
62. K. Prince, S. Sasidharan, N. Nag, T. Tripathi, P. Saudagar, Integration of spectroscopic and computational data to analyze protein structure, function, folding, and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 483–502
63. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Springer Nature, Singapore, 2020)

Investigating Protein Unfolding and Stability Using Chaotropic Agents and Molecular Dynamics Simulation

Rohit Shukla and Timir Tripathi

Abstract Protein folding and unfolding processes follow a thermodynamically favourable transitional path. The folding process occurs on a timescale in the order of milliseconds; therefore, observing the correct transitional pathway is challenging. However, with the advancement of computer science, it is now possible to decipher the structural level changes in the folding pathway of the protein using the molecular dynamics (MD) simulation. The MD simulation can provide detailed information about various energetic terms, structural parameters, etc. One can calculate the secondary structure changes with respect to time using MD simulation and correlate them with the CD spectra results. It can also generate thousands of snapshots that can be used to determine accurate unfolding pathways through structure visualization. In this chapter, we describe how chaotropic agents and MD simulation can be used in combination to study the stability and unfolding process of a protein. We also discuss the software used in the MD simulation with a detailed methodology of the GROMACS tool. Lastly, we take two case studies to show the process of urea and GdnHCl-induced denaturation of proteins analysed through MD simulation.

Keywords Urea · Guanidine hydrochloride · Unfolding · Stability · Dynamics · Folding · Denaturation

R. Shukla
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

T. Tripathi (✉)
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

Regional Director's Office, Indira Gandhi National Open University, Regional Centre Kohima, Kohima, India

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_10

1 Introduction

The protein unfolding studies often involves the use of chaotropic agents such as urea and guanidinium hydrochloride (GdnHCl), which reduce the stability of the native protein by destabilizing the hydrophobic interactions between various amino acids [1]. They are widely used for protein unfolding analysis, but the exact mechanism of action is still a mystery. It is well established that protein stability depends on the hydrogen bonding network of the protein with the solvent and intramolecular hydrogen bond interactions [2]. A proper hydrogen bond network is required for a protein to function correctly. Several studies have shown that chaotropic agents can directly bind to the protein or bind with the solvent and alter the properties of the solvent [3–9]. In other cases, the presence of chaotropic molecules also breaks the hydrogen bond network between water molecules, which induces the weakening of the hydrophobic effects. The effect of hydrogen bond disruption due to chaotropic agents is similar to the temperature and pressure-induced hydrogen bond network disruption for the denaturation of the protein [10, 11]. Additionally, the direct binding of the chaotropic molecules with the proteins may weaken the hydrophobic interactions between the non-polar amino acids responsible for stabilizing native proteins.

The folding energy difference between the well-folded and unfolded proteins is typically between 5 and 10 kcal/mol. The unfolded protein is 5–10 kcal/mol less stable than the corresponding native protein. During folding, multiple forces weaken simultaneously with several conformations between native to unfolding transitional states [12], indicating the level of complexity in understanding the protein unfolding and folding process. It also suggests that multiple factors are involved in unfolding/folding processes that should be carefully examined [13, 14]. There are a lot of limitations in experimental methods for studying the protein folding mechanism. They cannot provide detailed visual information about the transitional intermediates during protein unfolding from the nanosecond to the microsecond time scale [15]. They are also expensive in terms of money and labour. The currently available computational approaches can simulate the protein at a microsecond time scale in the presence of denaturants or temperature and can determine the exact unfolding steps in the form of complete trajectories saved at different snapshots. During simulation, several energy parameters can be analysed as well as the detailed insight molecular mechanism of unfolding can be investigated. These trajectories can be analysed using several software, and a lot of meaningful information can be extracted. With the support of graphical processing units (GPUs), currently, supercomputers can perform microsecond time simulation within a day and store petabytes of data. The simulation of a single virus is also possible [16]. However, the addition of solvent and other molecules in the simulation can increase the computational cost and complexity of the simulation result analysis [17–19].

2 Basic Concept of Protein Folding

Anfinsen's hypothesis shifted the concept of protein folding from the disulphide bridge protein folding theory towards the complete protein folding analysis through the eyes of computer scientists or polymer physicists in 1973 [20]. One had to cease thinking in terms of atomic coordinates to demonstrate the uniqueness (stability) of the native structure. It was stated that the necessary pieces of knowledge for folding must be present in the sequence, which was established as the Anfinsen thermodynamic hypothesis [21]. It essentially assumes that the sequence controls the interactions present in the native structure. This is the central idea behind the fascinating intersection of two major lines of research into the prediction of protein structure and the study of protein folding kinetics [22, 23]. Protein folding occurs through an enormous number of possible conformations that cannot be calculated through conventional chemical methods. Levinthal's paradox describes the astronomical number of local minima in the conformational space and the resulting inability to completely explore all conformational spaces. It has been established that even a straightforward explanation of protein folding based on the hydrophobic/hydrophilic model on a cubic lattice is nondeterministic polynomial (NP)-complete in this regard [24]. Overall, the link between sequence and structure and the elucidation of folding processes are challenging issues that are listed among the most significant scientific questions of the twenty-first century [25]. By virtue of their fundamental characteristics, Levinthal's paradox and Anfinsen's hypothesis appear at odds. To assure convergence towards the native state within a definite time, the folding process must first be constrained along a particular path (kinetic control). Conversely, the interim path (thermodynamic control) is comparatively irrelevant because it relies on the function, which is biased towards the final confirmation of the protein. Within the framework of the landscape theory of protein folding, in which both types of regulation are acknowledged, these contradictory criteria become consistent [26, 27]. According to the present theory, parallelization makes more sense early in the folding process and becomes more sequential in the latter stages [28, 29].

3 Chaotropic Agents and Their Mechanism of Action

The chaotropic agents (chaotropes) are chemical entities that disrupt the structure of biological macromolecules, such as nucleic acids and proteins, via the denaturation process. These molecules disrupt the non-covalent interactions such as van der Waals forces, hydrogen bonds, electrostatic interactions, and hydrophobic effects and increase the entropy of the system. The tertiary structure of well-folded biomolecules depends on these non-covalent forces; hence, increasing the concentration of chaotropes in the solution leads to the destabilization of protein followed by denaturation and reduced enzyme activity. The proper folding of a protein is depended on the hydrophobic interactions between the amino acids. Due to the

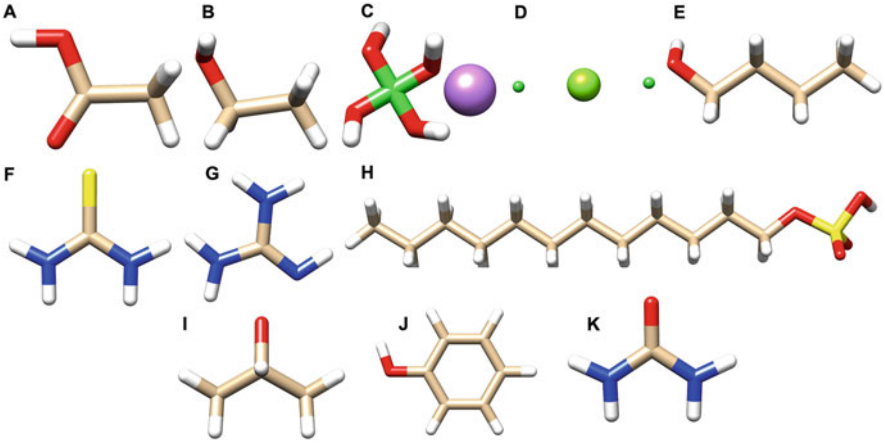

Fig. 1 Chemical structures of a few chaotropic agents. (**a**) lithium acetate, (**b**) ethanol, (**c**) lithium perchlorate, (**d**) magnesium chloride, (**e**) n-butanol, (**f**) thiourea, (**g**) guanidinium chloride, (**h**) sodium dodecyl sulphate, (**i**) 2-propanol, (**j**) phenol and (**k**) Urea

disordered water molecules, the chaotropic solutes reduce the net hydrophobic effects of the hydrophobic regions. This leads to the solubilization of the protein's hydrophobic regions via denaturation. This is also implicated in the case of hydrophobic regions of the lipid bilayers, where a high chaotropic concentration leads to cell lysis by disrupting membrane integrity [30].

The dissociation of chaotropes in solution results in different chaotropic effects. While the chaotropic solvents such as ethanol affect the non-covalent intramolecular forces, the chaotropic salts affect the charged interactions such as salt bridge, etc. A strong hydrogen bond network in proteins is observed in the non-polar medium; therefore, chaotropic salts that can increase the chemical polarity can affect the hydrogen bond network. This happens due to the smaller number of water molecules that can effectively solvate the ions. It leads to the ion–dipole interactions between the hydrogen bonding species and salts which are stronger and more favourable than normal hydrogen bonding [31, 32]. The common chaotropic agents are urea, guanidinium chloride (GdnHCl), thiourea, and sodium dodecyl sulphate (SDS). The chemical structure of a few chaotropic agents is shown in Fig. 1.

4 Molecular Dynamics (MD) Simulation

The molecular dynamics (MD) simulation method was introduced in the 1970s [33, 34]. Currently, with the improvement of computational power, it can be used to simulate from thousands of atoms to the complete virus, proteins, nucleic acids, nucleosomes [35, 36] or ribosomes [37, 38], etc., using the explicit water models. Today, simulations of ~50,000–100,000 atom size systems are in routine practice, and even simulations of more than 1,000,000 atoms are also possible when good computational facilities are available. This was made possible due to the improvement in the MD algorithms and new computing capabilities from the past few decades.

The input structure of any biomacromolecule can be obtained using computational modelling tools or experimental methods [39]. The simulated systems can be represented at different levels of time scale. The atomistic representation model is the best for the reproduction of actual systems. Although, in the case of long simulations or large biological systems, the coarse-grained representation is leading popularity [40]. There are many representation approaches, but the explicit solvent model is the simplest, most popular, and most effective [41–46]. However, increasing the system size in this model increases the size of simulated systems. This solvent model can achieve the solvation effects that happen in a real solvent, including those of entropic origin, like the hydrophobic effect. After building the complete system, using the deriving equations, the forces that act on each atom can be obtained using the force field. In the force field, the potential energy is inferred from the molecular structure [47–52]. The complex equations represent the force field terms which are easy to calculate. There are several simple molecular features that characterize the force field terms, such as bond angles and length, which are represented by springs, bonds rotations, and Lennard-Jones potential represented by periodic functions, electrostatic and van der Walls interaction calculation by Coulomb's law. These terms guarantee that force and energy calculations be very fast for large biological systems. Currently, the parameterization of the force field differs in various atomistic molecular simulations. There are several parameters in the force field which cannot be interchanged, and also, not all force fields allow to represent the all-molecule types though the simulation trajectories and analysis for all the force fields are similar [53, 54]. When the acting forces on each atom are calculated, Newton's classical law of motion is utilized for the acceleration and velocities calculation, including the update of the position of each atom. The MD system movement integration is done using numerical methods; therefore, to avoid instability, a time step shorter than the fastest movement in the molecules is used. This short-time integrator usually lies between 1 and 2 fs for the atomistic simulation and plays a crucial role in the overall simulation.

The long microsecond simulations hardly scratch the time scales for the biological systems and require iterating over the calculation cycle 10^9 times. The coarse-grained simulations are generally better with these limitations. They use a more simplified MD system and represent larger time steps for integration; hence, they can run the large-scale simulation of large biomacromolecules with good accuracy. The long simulations can run with several advantages that include fine-tuning several energetic parameters and parallelization of the simulation by using graphical processing units (GPUs) that can increase accuracy and improve the simulation speed. The current generation of computers can parallelize the process, which leads to faster MD simulation.

Several MD simulation software are available, and the most widely used are CHARMM [55], GROMACS [56], AMBER [57] and NAMD [58]. These software are well compatible with the messaging passing interface (MPI). Due to a large number of cores in the computers, the MPI can significantly increase the computation power and reduce the computational time. The MD simulation process can be divided into multiple CPU cores that can reduce computational time; this technique is known as spatial decomposition. The part of the complete system is used for the simulation in each processor. This division of MD simulation systems is based on the particle's position in space and not on the list of particles. The region of the space is dealt by each processor instead of the particles present in the MD simulation system. The processor communication is also reduced because only the neighbouring regions of the simulation share information among them [59]. Nowadays, GPUs are becoming the breakthrough in the case of MD simulation due to their ability to accelerate the simulation speed. The currently available MD simulation tools are compatible with GPUs, and even some MD simulation programs, such as ACEMD [60], are written to run on GPU systems. The combination of CPUs and GPUs is the default strategy in the case of atomistic simulations. Currently, high-performance computing (HPC) is the most popular among computational scientists, while GPU development is leading to the greater use of personal computers for atomistic simulations than HPC.

5 Application of MD Simulation in Investigating Protein Unfolding

As described earlier, MD simulation can mimic the in vivo conditions and can provide information on the real dynamics of the system, including effects of mutation in a protein [61–63], protein–ligand interactions [64–67] and protein unfolding [68–70]. The stepwise methodology of the MD simulation process is briefly described below [71, 72].

1. The biological macromolecules should be prepared. The structure may be modelled if an experimental structure is unavailable in the PDB (https://www.rcsb.org). All the hydrogen atoms should be added to the PDB structure.
2. The PDB file should be placed in a box, which can be cubic, dodecahedron, etc.
3. The explicit water molecules should be filled into the box.
4. The concentration of chaotropes should be calculated in the number and added to the simulation box by replacing the water molecules.

5. The MD systems should be neutralized by adding ions.
6. Energy minimization should be performed to remove the steric clashes of the systems caused by the addition of water and chaotropic agents.
7. The number of volumes and temperature (NVT) and the number of pressure and temperature (NPT) simulations should be run to fix the volume, pressure and temperature of the system. After this simulation, the quality of the system can be assessed by plotting all the graphs (pressure, volume, temperature, etc.).
8. Finally, the MD simulation should be run, and values should be saved at 1 to 2 fs time intervals.
9. Lastly, the obtained trajectories should be pre-processed by removing the periodic boundary condition (PBC) artifacts. Several results can be obtained in the form of various graphs, such as root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (Rg), solvent accessible surface area (SASA), principal component analysis (PCA), and secondary structure analysis.
10. The trajectories can also be visualized, and the unfolding of the biomacromolecules can be recorded in the form of a trajectory. These graphical and visual analyses can give a glimpse of the complete unfolding process of proteins.

These steps are generally used in all the MD simulation protocols to perform the unfolding analysis of a protein. The graphical user interface (GUI) simulation software such as Desmond and YASARA can be used for this process in a few steps, while the command lines tools such as AMBER and GROMACS complete it in many steps. The general methodology and concept are the same for all the software. Now we discuss two case studies of protein unfolding using urea and GdnHCl.

6 Case Studies

6.1 Urea-Induced Unfolding

We have reported the urea-induced unfolding of the *Acinetobacter baumannii* UDP-N-acetylglucosamine enolpyruvyl transferase (AbMurA) [69]. The structural and unfolding features of AbMurA were analysed using multiple spectroscopic methods, including circular dichroism and fluorescence spectroscopy [73]. The data showed the protein unfolds in a three-state manner with the presence of an unfolding intermediate at 3.5 M urea. The spectroscopic data was complemented using data from multiple 100 ns MD simulations [69]. To study the unfolding behaviour of the AbMurA enzyme, we created six systems where we placed the AbMurA in water, 3.5 M, and 8.0 M urea, and simulated at 300 and 400 K temperatures. In total, we created six systems (AbMurA$_{H2O}$, AbMurA$_{3.5}$ and AbMurA$_{8.0}$ at 300 and 400 K) and generated trajectories at 100 ns. The results were analysed in terms of RMSD,

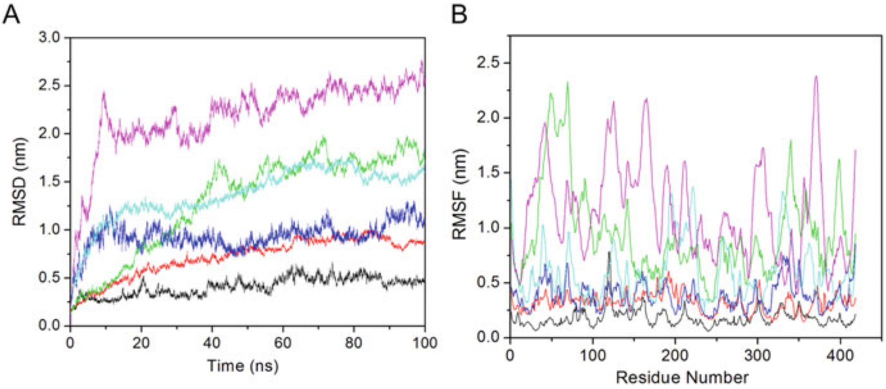

Fig. 2 (**a**) RMSD, (**b**) RMSF. The black, blue, red, green, cyan and magenta represent AbMurA$_{H2O}$ (300 K), AbMurA$_{H2O}$ (400 K), AbMurA$_{3.5}$ (300 K), AbMurA$_{3.5}$ (400 K), AbMurA$_{8.0}$ (300 K) and AbMurA$_{8.0}$ (400 K), respectively

RMSF, Rg, SASA, PCA, structural analysis, and secondary structure analysis. We briefly discuss the results; for a detailed analysis, readers can refer to the original article [69].

We first calculated the RMSD to study the detailed dynamics of the system. At 300 K, the average RMSD value for the AbMurA$_{H2O}$, AbMurA$_{3.5}$ and AbMurA$_{8.0}$ at 300 K were 0.41, 0.71, and 1.30 nm, respectively (Fig. 2a). Figure 2a shows that AbMurA$_{H2O}$ quickly achieved the equilibration state and showed a stable trajectory till 100 ns. The AbMurA$_{3.5}$ showed an increase in the RMSD value initially, while after 40 ns, it achieved the equilibration state. In the case of AbMurA$_{8.0}$, an abrupt pattern was observed till 65 ns and then attained the equilibration state. The average RMSD values represent that urea addition in the systems induces instability in the AbMurA enzyme. We then calculated the RMSD values of AbMurA$_{H2O}$, AbMurA$_{3.5}$ and AbMurA$_{8.0}$ at 400 K (Fig. 2a). 400 K temperature can immediately unfold the protein and provide information on the proper unfolding pathway in the presence of urea. In 400 K, AbMurA$_{H2O}$ attained the equilibration state after 20 ns, while the other two systems, AbMurA$_{3.5}$ and AbMurA$_{8.0}$, achieved the equilibration state after 40 ns and remained stable till 100 ns. The average RMSD values were 0.93, 1.36, and 2.18 nm for AbMurA$_{H2O}$, AbMurA$_{3.5}$ and AbMurA$_{8.0.}$ The RMSD result analysis represents that all the systems got the equilibration state and can be further used. It also showed that at 3.5 M concentration of urea, the AbMurA formed an intermediate state, while at 8.0 M of urea, it was completely unfolded.

The RMSF values for the systems were also calculated at 300 and 400 K (Fig. 2b). At 300 K, the RMSF values for the AbMurA$_{H2O}$ were stable, though a higher peak was observed between 115 and 125 residues (with RMSF value between 0.23 and 0.78 nm). When 3.5 M urea was added to the system, RMSF values of >0.5 nm were observed for all the systems. In the case of AbMurA$_{8.0}$, high RMSF values were observed, indicating that the addition of urea induces changes in the structural conformations followed by protein unfolding. At 400 K, an average

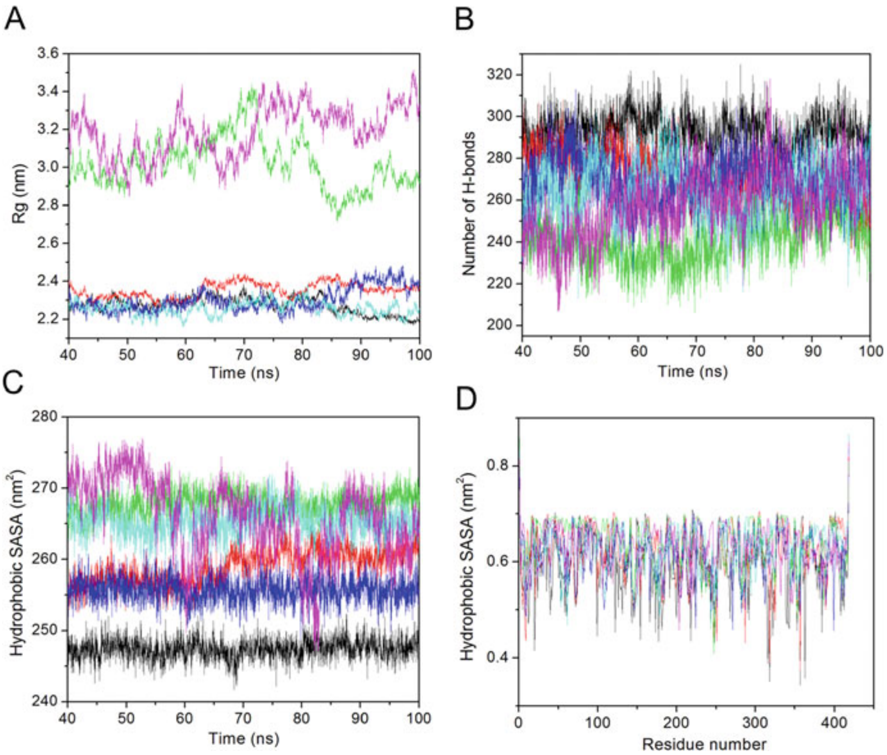

Fig. 3 (**a**) Radius of gyration. (**b**) Number of hydrogen bonds. (**c**) Solvent accessible surface area. (**d**) Solvent accessible surface area versus residues. The black, blue, red, green, cyan and magenta represent AbMurA$_{H2O}$ (300 K), AbMurA$_{H2O}$ (400 K), AbMurA$_{3.5}$ (300 K), AbMurA$_{3.5}$ (400 K), AbMurA$_{8.0}$ (300 K) and AbMurA$_{8.0}$ (400 K), respectively

fluctuation between 0.2 to 0.5 nm was observed for AbMurA$_{H2O}$. High RMSF values of >0.5 nm were observed for residues 35–49, 66–70, 323–352 and 411–148. AbMurA$_{3.5}$ showed RMSF values between 0.5 and 1.0 nm for all the systems. AbMurA$_{8.0}$ showed higher RMSF values for all the residues representing complete structure loss at 8.0 M urea. The overall RMSF analysis indicates that the addition of urea disrupts the original conformation of AbMurA.

We also calculated the Rg values for all the systems using the last 60 ns equilibrated trajectories. Compared to other systems, higher Rg values were observed for 8.0 M urea at 300 and 400 K. For other systems, the Rg values for AbMurA$_{3.5}$ were more than AbMurA$_{H2O}$ (Fig. 3a). The number of hydrogen bonds for all the systems was also calculated (Fig. 3b), which was in the order of AbMurA$_{H20}$ > AbMurA$_{3.5}$ > AbMurA$_{8.0}$. This suggests that the addition of urea leads to the loss of hydrogen bonds. The average number of hydrogen bonds was 270, 264, and 258 for AbMurA$_{H20}$, AbMurA$_{3.5}$, and AbMurA$_{8.0}$, respectively, at 400 K. The SASA values were also analysed (Fig. 3c), which closely agreed with the Rg data. Higher SASA values were observed for the AbMurA$_{3.5}$ and AbMurA$_{8.0}$

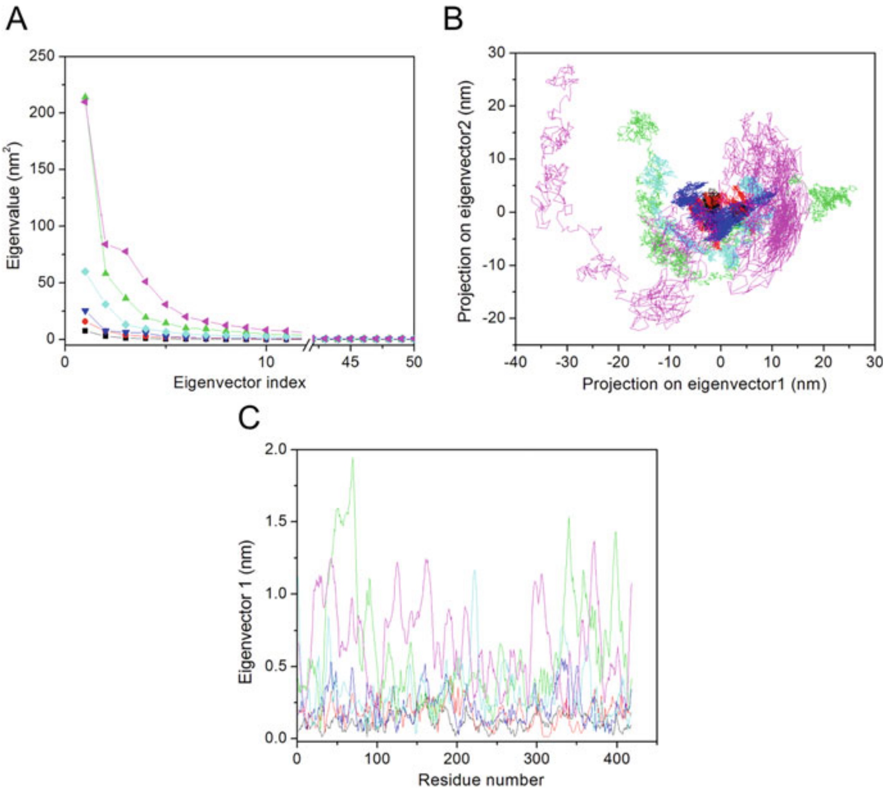

Fig. 4 (**a**) Eigenvalue versus eigenvector. (**b**) 2D project plot. (**c**) eigRMSF values. The black, blue, red, green, cyan and magenta represent AbMurA$_{H2O}$ (300 K), AbMurA$_{H2O}$ (400 K), AbMurA$_{3.5}$ (300 K), AbMurA$_{3.5}$ (400 K), AbMurA$_{8.0}$ (300 K) and AbMurA$_{8.0}$ (400 K), respectively

than AbMurA$_{H20}$, representing the unfolding of the protein. For residual SASA, we analysed the SASA value of tryptophan residue, which indicates that the addition of urea increases the exposure of tryptophan towards the solvent followed by unfolding (Fig. 3d). Collectively, all results suggest that the addition of the urea induces the unfolding of the AbMurA protein.

The PCA was carried out to analyse the correlated motions induced by the addition of urea (Fig. 4). Since the first few eigenvectors represent the overall dynamics of the system, hence first five eigenvectors were considered (Fig. 4a). AbMurA$_{H20}$ showed less correlated motions, while AbMurA$_{3.5}$ and AbMurA$_{8.0}$ showed higher correlated motions. The pattern was the same for both 300 and 400 K temperatures. PCA data also showed a partial unfolding of the protein at 3.5 M urea and complete unfolding at 8.0 M urea. The first two eigenvectors were then taken and plotted (Fig. 4b). The data showed a stable cluster for the AbMurA$_{H20}$ and dispersed clusters for AbMurA$_{3.5}$ and AbMurA$_{8.0}$. Lastly, the eigRMSF values (Fig. 4c) were analysed, which showed a similar pattern to the RMSF values. Higher

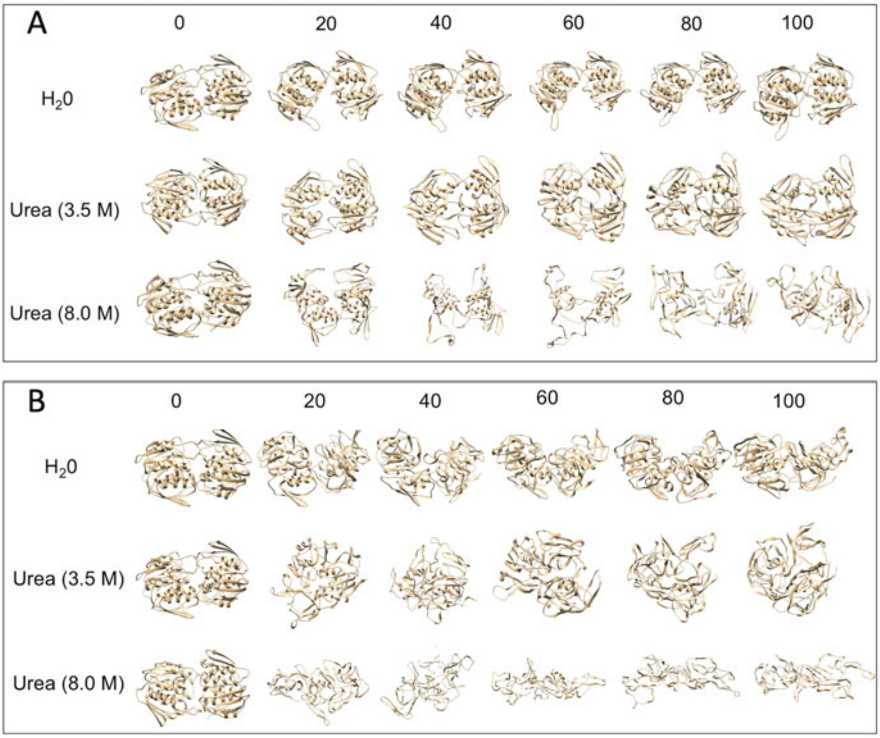

Fig. 5 Time-dependent secondary structural changes. Structural features obtained from the snapshots generated at 20 ns time intervals at (**a**) 300 K and (**b**) 400 K for AbMurA

residue fluctuations were observed in AbMurA$_{3.5}$ and AbMurA$_{8.0}$ systems, while lower fluctuations were observed for AbMurA$_{H20}$. The overall PCA results concluded that at 3.5 M of urea, the AbMurA formed an intermediate folding state, while complete unfolding was observed at 8.0 M urea.

The MD simulation can produce trajectories that can be visually analysed using any visualization software. We analysed the trajectories at 20 ns intervals to obtain a visual representation of the urea-induced unfolding at 300 and 400 K temperatures (Fig. 5). Firstly, we analysed the structural snapshots at 300 K for AbMurA$_{H2O}$, AbMurA$_{3.5}$ and AbMurA$_{8.0}$. It is evident from Fig. 5a that AbMurA$_{H2O}$ did not unfold till 100 ns while there were minor structural changes in the AbMurA$_{3.5}$ intermediate state. The AbMurA$_{8.0}$ started unfolding after 40 ns. It showed the disappearance of the stable secondary structures, such as alpha helices and beta sheets, and an increase in turns and loops. The data showed that 8.0 M urea induces the structural unfolding in the protein. We then analysed the structural changes at 400 K for AbMurA$_{H2O}$, AbMurA$_{3.5}$ and AbMurA$_{8.0}$. The structural snapshots at 20 ns time intervals are shown in Fig. 5b. The data shows that the presence of urea at 400 K temperature induces large structural changes in the protein. At 400 K, the intermediate state at 3.5 M urea also showed structural disruption, while major changes were observed in the presence of 8.0 M urea. $AbMurA_{8.0}$ showed total disruption of the structure after 40 ns. The data showed the presence of an intermediate state at 3.5 M urea while a complete structure disruption at 8.0 M urea.

The secondary structure analysis was carried out to analyse the secondary structure level changes with respect to time (Fig. 6). The coils, turns and bends were found to be increased at higher concentrations of urea while beta sheets and alpha helices disappeared. First, the secondary structural changes at 300 K were analysed (Fig. 6a). The AbMurA in water showed a stable secondary structure and no major changes throughout the simulation, while at 3.5 M urea, the AbMurA showed a few changes, such as an increase in coils, bends and turns but no major losses in the stable secondary structures. The AbMurA at 8.0 M urea showed much higher turns, coils and bends and loss of helices and sheets. From residues 1–130, we observed the loss of rigid structures and increased bends, turns and coils. The overall analysis showed that at 3.5 M of urea, AbMurA showed minor changes in the secondary structures, while at 8.0 M of urea, major structural changes occurred. The secondary structural changes were also analysed at 400 K for $AbMurA_{H2O}$, $AbMurA_{3.5}$ and $AbMurA_{8.0}$ (Fig. 6b). Here also, it was observed that $AbMurA_{H2O}$ showed stable structures with a few temperature-induced changes. The $AbMurA_{3.5}$ showed an increase in the coils, bends and turns and minor changes in helices and sheets. The $AbMurA_{8.0}$ system showed a much higher number of bends, turns and coils and the disappearance of sheets and helices. Only a few beta sheets were observed, while the alpha helices completely disappeared. The data indicated that in 8.0 M urea, the AbMurA completely lost secondary structures.

The combined spectroscopic and MD simulation data showed the structural characteristics of AbMurA in native ($AbMurA_{H2O}$), intermediate ($AbMurA_{3.5}$) and unfolded ($AbMurA_{8.0}$) states [69]. The data obtained from the MD simulation revealed the atomistic and structural basis of the unfolding of AbMurA, which was not possible using only spectroscopic methods.

6.2 GdnHCl-Induced Unfolding

GdnHCl is another chaotropic agent widely used for denaturation studies of proteins. It can also be added to the MD simulation box, and the structural changes can be captured at different time scales. We discuss a case study from the work of Syed et al. [74]. Firstly, they carried out the unfolding analysis using the series of in vitro experiments and then, for analysing the atomic level structural changes, they carried out the detailed MD simulation analysis. The authors described the folding pattern of the 196–443 residues of human integrin linked kinase (ILK) with 100 ns MD simulation in water, 2.0, 4.0, 6.0, and 8.0 M GdnHCl concentrations. We will discuss key findings from this study related to the MD simulation. They created a total of five MD systems and analysed parameters such as RMSD, RMSF, Rg, SASA, the number of hydrogen bonds, etc.

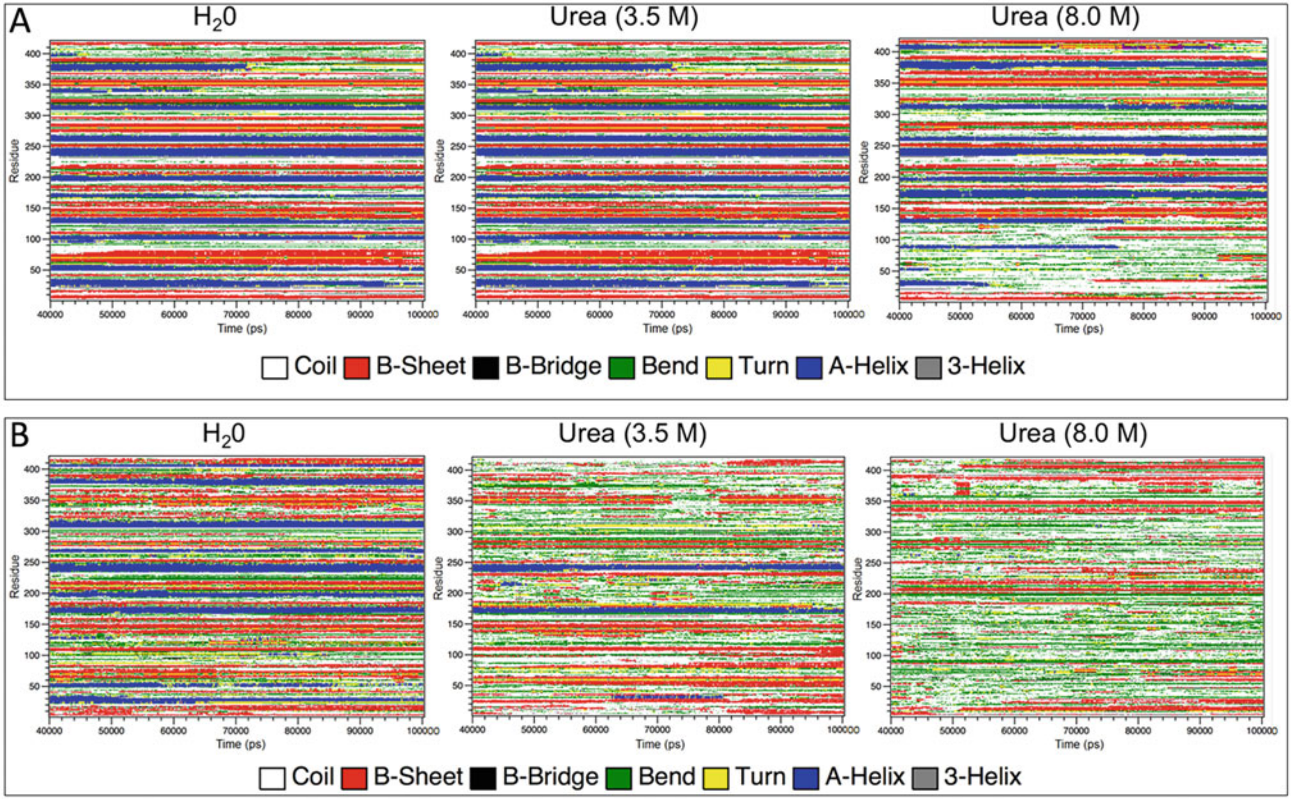

Fig. 6 Evolution of secondary structures of AbMurA at (**a**) 300 K and (**b**) 400 K

Firstly, they calculated the potential energy of the system, where they found that ILK_{H2O}, $ILK_{2.0}$, $ILK_{4.0}$, $ILK_{6.0}$ and $ILK_{8.0}$ showed −2.92196, −358,821, −411,597, −439,100, −1,044,520 kJ/mol energy, respectively. The potential energy of the systems represented that ILK_{H2O} is more stable than GdnHCl systems.

To find the deviation from the initial structure, the authors calculated the RMSD value for all the systems. The average RMSD for ILK_{H2O}, $ILK_{2.0}$, $ILK_{4.0}$, $ILK_{6.0}$ and $ILK_{8.0}$ were 0.33, 0.40, 0.34, 0.27 and 0.36 nm, respectively (Fig. 7a). The figure indicated that the $ILK_{2.0}$, $ILK_{4.0}$ and $ILK_{8.0}$ showed more deviation than ILK_{H2O}. The authors observed less RMSD value for $ILK_{6.0}$. They observed higher changes at 2.0 M GdnHCl concentration throughout the simulation. The $ILK_{4.0}$ and $ILK_{8.0}$ systems showed a little higher RMSD value than ILK in water, representing that at this GdnHCl concentration, partial conformational changes are occurring in the ILK protein. The RMSD analysis showed that all the systems were stable and generated trajectories that can be further utilized for other studies.

After RMSD analysis, the authors calculated the Rg and analysed it in detail. Rg is an important parameter to describe the unfolding pattern of a protein. The average Rg values for the ILK_{H2O}, $ILK_{2.0}$, $ILK_{4.0}$, $ILK_{6.0}$ and $ILK_{8.0}$ were 1.72, 1.72, 1.71, 1.76 and 1.76 nm, respectively. The Rg values were plotted with respect to the time (Fig. 7b) that showed that ILK is getting unfolded at 6.0 and 8.0 M concentrations of GdnHCl while conformation changes occur in the ILK at 4.0 M. The $ILK_{2.0}$ showed a similar Rg value as ILK in water. It indicates that $ILK_{6.0}$ and $ILK_{8.0}$ lost compactness and got unfolded.

To determine the GdnHCl-induced residue level changes, RMSF analysis was performed (Fig. 7c). It was seen that the addition of GdnHCl to the systems alters the original conformation of the protein and induces structural changes. Higher residual changes occurred between residues 221–230, 257–263, and 280–293, including the N- and C-terminals. It represents that GdnHCl disrupts the charge–charge interactions in the protein and induces global changes that lead to the unfolding of the ILK.

The SASA analysis was carried out to analyse the solvent accessible surface area changes induced by the GdnHCl. The average SASA values for ILK_{H2O}, $ILK_{2.0}$, $ILK_{4.0}$, $ILK_{6.0}$ and $ILK_{8.0}$ were 132.31, 132.51, 131.40, 133.04 and 134.09 nm^2, respectively (Fig. 7d). It was observed that $ILK_{6.0}$ and $ILK_{8.0}$ showed higher SASA values, indicating the unfolding of ILK. The $ILK_{2.0}$ and $ILK_{4.0}$ systems showed similar SASA values as the ILK in water. From the overall SASA analysis, it was observed that 6.0 and 8.0 M GdnHCl induces the unfolding in the ILK protein.

The folding of the protein strongly depends on the formation of hydrogen bonds. More number of hydrogen bonds in a protein represents a compact and well-folded structure, while a lesser number of hydrogen bonds represents a less compact and elongated structure. The authors plotted the number of hydrogen bonds with respect to time (Fig. 8). The average number of hydrogen bonds between ILK and water molecules were 420, 389, 346, 328 and 361 for ILK_{H2O}, $ILK_{2.0}$, $ILK_{4.0}$, $ILK_{6.0}$ and $ILK_{8.0}$, respectively (Fig. 8a). The hydrogen bonds between ILK and GdnHCl were also calculated (Fig. 8b). The average number of hydrogen bonds between ILK and GdnHCl was 18, 26, 41 and 30, respectively, for ILK_{H2O}, $ILK_{2.0}$, $ILK_{4.0}$, $ILK_{6.0}$ and $ILK_{8.0}$, respectively. The result indicates that adding GdnHCl decreases the ILK

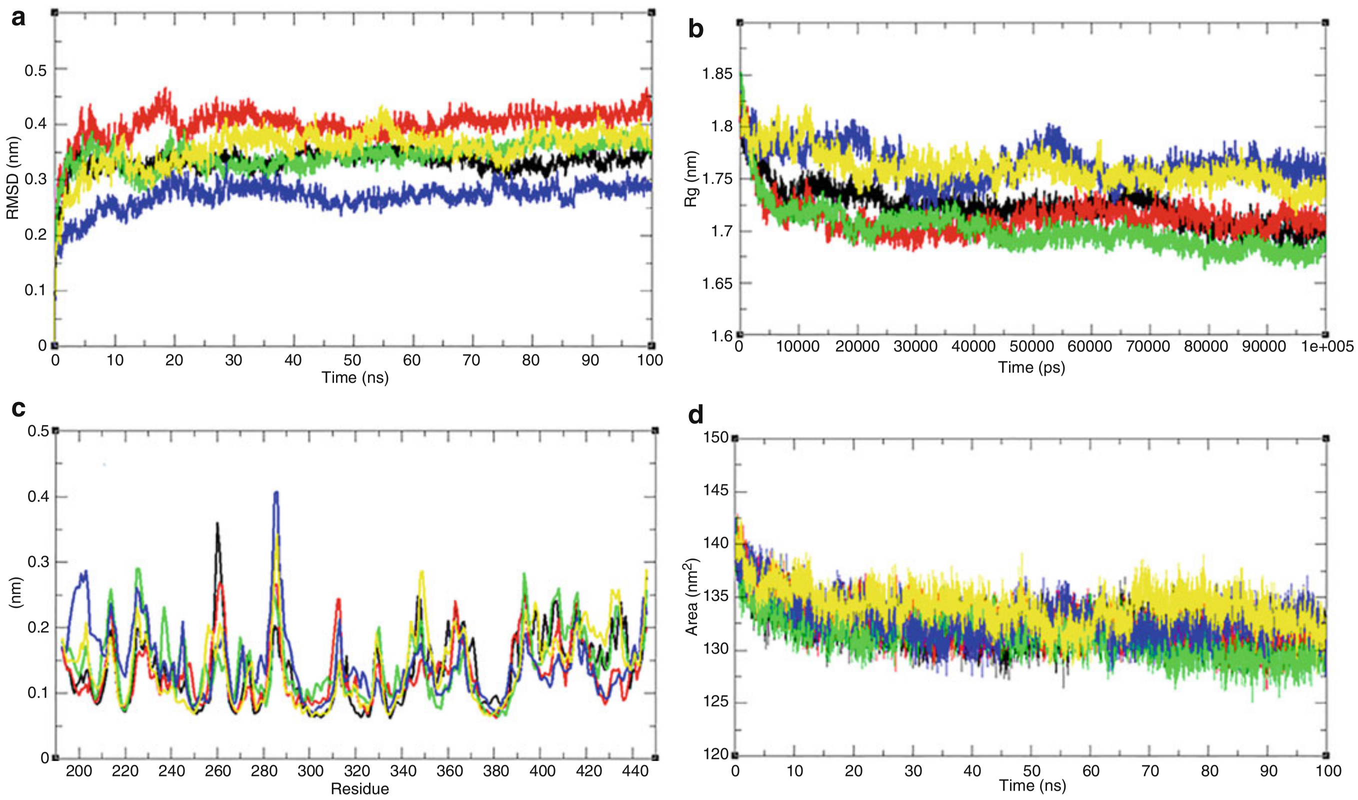

Fig. 7 (**a**) RMSD, (**b**) radius of gyration, (**c**) RMSF, and (**d**) SASA. The ILK$_{H2O}$, ILK$_{2.0}$, ILK$_{4.0}$, ILK$_{6.0}$ and ILK$_{8.0\ are}$ represented by black, red, green, blue and yellow colours

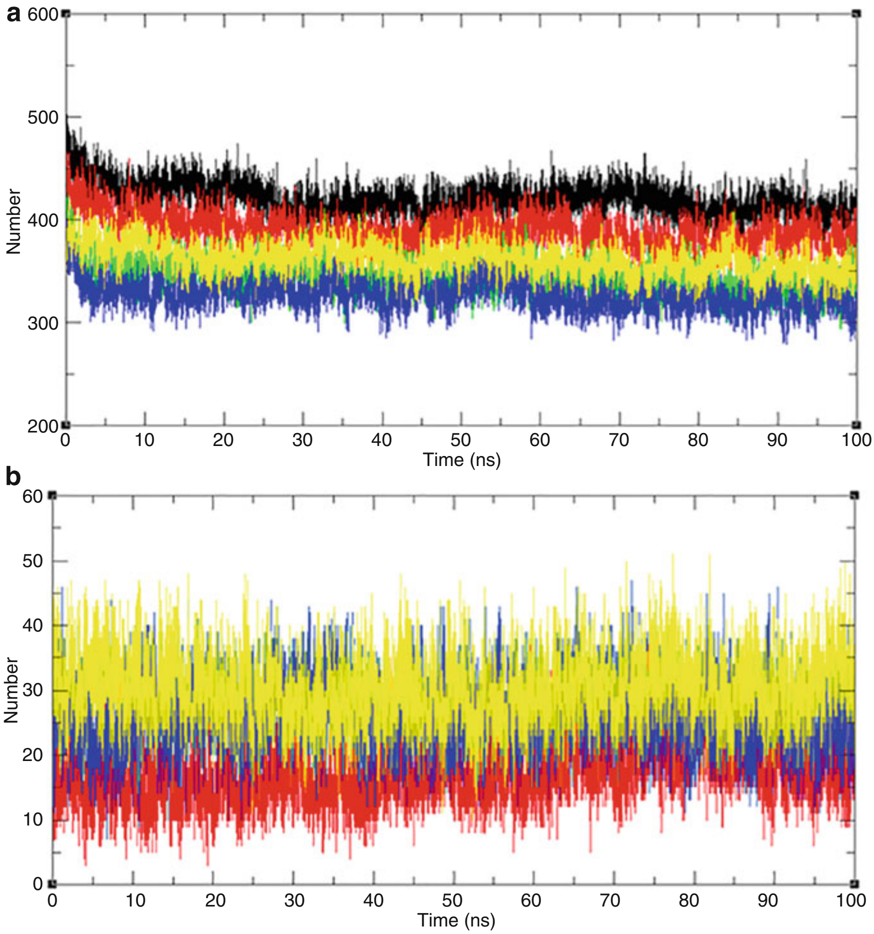

Fig. 8 Number of hydrogen bonds. (**a**) Intramolecular hydrogen bonds of ILK. (**b**) Hydrogen bonds between ILK and GdnHCl. The ILK_{H2O}, $ILK_{2.0}$, $ILK_{4.0}$, $ILK_{6.0}$ and $ILK_{8.0}$ are represented by black, red, green, blue and yellow colours

interaction with water while increasing the interaction with GdnHCl itself. A proper hydration state is required for the solubility of the protein; therefore, it represents that the addition of GdnHCl is disrupting the original conformation of the ILK and inducing the folding in the protein.

From the overall result, the authors concluded that ILK showed higher unfolding at 6.0 and 8.0 M GdnHCl concentrations, representing that the addition of chaotropic agents leads to the unfolding of ILK.

7 Conclusions

Chaotropic agents belong to several chemical families and can induce the denaturation of biomolecules. They follow different mechanisms to alter the structures and denature proteins. Several in vitro spectroscopic methods are available to analyse the effect of the chaotropic agents on proteins, but they cannot provide information on the atomic level changes in the protein structure with respect to time. MD simulation is emerging as an essential tool to track the structural changes and generate thousands of the conformations of a protein. It can also be used to visualize trajectories to analyse the detailed structural level changes. We discussed two case studies using urea and GdnHCl in MD simulation to study the unfolding of proteins in detail. The data showed that the MD simulation result agreed well with the spectroscopic findings and provided several additional atomistic information. Further improvements in the force field and algorithms may help gather precise conformational changes induced by chaotropic agents against the biological macromolecules.

References

1. A. Fershi, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W.H. Freeman, New York, 1999)
2. W. Kauzmann, Some factors in the interpretation of protein denaturation. Adv. Protein Chem. **14**, 1–63 (1959). https://doi.org/10.1016/S0065-3233(08)60608-7
3. H.S. Frank, F. Franks, Structural approach to the solvent power of water for hydrocarbons; urea as a structure breaker. J. Chem. Phys. **48**, 4746–4757 (1968). https://doi.org/10.1063/1.1668057
4. D.O.V. Alonso, K.A. Dill, Solvent denaturation and stabilization of globular proteins. Biochemistry **30**, 5974–5985 (1991). https://doi.org/10.1021/bi00238a023
5. A. Caflisch, M. Karplus, Molecular dynamics simulation of protein denaturation: solvation of the hydrophobic cores and secondary structure of barnase. Proc. Natl. Acad. Sci. U S A **91**, 1746–1750 (1994). https://doi.org/10.1073/pnas.91.5.1746
6. A. Wallqvist, D.G. Covell, D. Thirumalai, Hydrophobic interactions in aqueous urea solutions with implications for the mechanism of protein denaturation. J. Am. Chem. Soc. **120**, 427–428 (1998). https://doi.org/10.1021/ja972053v
7. R. Chitra, P.E. Smith, Preferential interactions of cosolvents with hydrophobic solutes. J. Phys. Chem. B **105**, 11513–11522 (2001). https://doi.org/10.1021/jp012354y
8. S. Shimizu, H.S. Chan, Origins of protein denatured state compactness and hydrophobic clustering in aqueous urea: inferences from nonpolar potentials of mean force. Proteins **49**, 560–566 (2002). https://doi.org/10.1002/prot.10263
9. B.J. Bennion, V. Daggett, The molecular basis for the chemical denaturation of proteins by urea. Proc. Natl. Acad. Sci. U S A **100**, 5142–5147 (2003). https://doi.org/10.1073/pnas.0930122100
10. S. Kunugi, N. Tanaka, Cold denaturation of proteins under high pressure. Biochim. Biophys. Acta **1595**, 329–344 (2002). https://doi.org/10.1016/s0167-4838(01)00354-5
11. M.I. Marqués, J.M. Borreguero, H.E. Stanley, N.V. Dokholyan, Possible mechanism for cold denaturation of proteins at high pressure. Phys. Rev. Lett. **91**, 138103 (2003). https://doi.org/10.1103/PhysRevLett.91.138103
12. J.S. Yang, W.W. Chen, J. Skolnick, E.I. Shakhnovich, All-atom ab initio folding of a diverse set of proteins. Structure **1993**(15), 53–63 (2007). https://doi.org/10.1016/j.str.2006.11.010

13. K.A. Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, The protein folding problem. Annu. Rev. Biophys. **37**, 289–316 (2008). https://doi.org/10.1146/annurev.biophys.37.092707.153558
14. V. Daggett, A. Fersht, The present view of the mechanism of protein folding. Nat. Rev. Mol. Cell Biol. **4**, 497–502 (2003). https://doi.org/10.1038/nrm1126
15. K. Prince, S. Sasidharan, N. Nag, T. Tripathi, P. Saudagar, Integration of spectroscopic and computational data to analyze protein structure, function, folding, and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 483–502. https://doi.org/10.1016/B978-0-323-99127-8.00018-0
16. P.L. Freddolino, A.S. Arkhipov, S.B. Larson, A. McPherson, K. Schulten, Molecular dynamics simulations of the complete satellite tobacco mosaic virus. Structure **14**, 437–449 (2006). https://doi.org/10.1016/j.str.2005.11.014
17. M. Levitt, R. Sharon, Accurate simulation of protein dynamics in solution. Proc. Natl. Acad. Sci. **85**, 7557–7561 (1988). https://doi.org/10.1073/pnas.85.20.7557
18. D.A.C. Beck, D.O.V. Alonso, V. Daggett, A microscopic view of peptide and protein solvation. Biophys. Chem. **100**, 221–237 (2003). https://doi.org/10.1016/s0301-4622(02)00283-1
19. V. Daggett, Protein folding−simulation. Chem. Rev. **106**, 1898–1916 (2006). https://doi.org/10.1021/cr0404242
20. C.B. Anfinsen, Principles that govern the folding of protein chains. Science **181**, 223–230 (1973). https://doi.org/10.1126/science.181.4096.223
21. T. Tripathi, Calculation of thermodynamic parameters of protein unfolding using far-ultraviolet circular dichroism. J. Protein. Proteomics **4**(2), 85–91 (2013)
22. B. Honig, Protein folding: from the levinthal paradox to structure prediction. J. Mol. Biol. **293**, 283–293 (1999). https://doi.org/10.1006/jmbi.1999.3006
23. D.B. Singh, T. Tripathi (eds.), *Frontiers in Protein Structure, Function, and Dynamics* (Springer, Singapore, 2020). https://doi.org/10.1007/978-981-15-5530-5
24. B. Berger, T. Leighton, Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. J. Comput. Biol. **5**, 27–40 (1998). https://doi.org/10.1089/cmb.1998.5.27
25. So much more to know. Science **309**, 78–102 (2005). https://doi.org/10.1126/science.309.5731.78b
26. J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins **21**, 167–195 (1995). https://doi.org/10.1002/prot.340210302
27. K.A. Dill, H.S. Chan, From Levinthal to pathways to funnels. Nat. Struct. Biol. **4**, 10–19 (1997). https://doi.org/10.1038/nsb0197-10
28. J. Schonbrun, K.A. Dill, Fast protein folding kinetics. Proc. Natl. Acad. Sci. U S A **100**, 12678–12682 (2003). https://doi.org/10.1073/pnas.1735417100
29. H. Kaya, H.S. Chan, Explicit-chain model of native-state hydrogen exchange: Implications for event ordering and cooperativity in protein folding. Proteins **58**, 31–44 (2005). https://doi.org/10.1002/prot.20286
30. P. Bhaganna, R.J.M. Volkers, A.N.W. Bell, K. Kluge, D.J. Timson, J.W. McGrath, H.J. Ruijssenaars, J.E. Hallsworth, Hydrophobic substances induce water stress in microbial cells. Microb. Biotechnol. **3**, 701–716 (2010). https://doi.org/10.1111/j.1751-7915.2010.00203.x
31. K.D. Collins, Charge density-dependent strength of hydration and biological structure. Biophys. J. **72**, 65–76 (1997)
32. G. Salvi, P. De Los Rios, M. Vendruscolo, Effective interactions between chaotropic agents and proteins. Proteins **61**, 492–499 (2005). https://doi.org/10.1002/prot.20626
33. J.A. McCammon, B.R. Gelin, M. Karplus, Dynamics of folded proteins. Nature **267**, 585–590 (1977). https://doi.org/10.1038/267585a0
34. A. Warshel, M. Levitt, Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J. Mol. Biol. **103**, 227–249 (1976). https://doi.org/10.1016/0022-2836(76)90311-9

35. D. Roccatano, A. Barthel, M. Zacharias, Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation. Biopolymers **85**, 407–421 (2007). https://doi.org/10.1002/bip.20690
36. S. Sharma, F. Ding, N.V. Dokholyan, Multiscale modeling of nucleosome dynamics. Biophys. J. **92**, 1457–1470 (2007). https://doi.org/10.1529/biophysj.106.094805
37. I. Tinoco, J.-D. Wen, Simulation and analysis of single-ribosome translation. Phys. Biol. **6**, 025006 (2009). https://doi.org/10.1088/1478-3975/6/2/025006
38. R. Brandman, Y. Brandman, V.S. Pande, A-site residues move independently from P-site residues in all-atom molecular dynamics simulations of the 70S bacterial ribosome. PLoS One **7**, e29377 (2012). https://doi.org/10.1371/journal.pone.0029377
39. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods* (Academic Press, London, 2022)
40. M. Orozco, L. Orellana, A. Hospital, A.N. Naganathan, A. Emperador, O. Carrillo, J.L. Gelpí, Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings. Adv. Protein Chem. Struct. Biol. **85**, 183–215 (2011). https://doi.org/10.1016/B978-0-12-386485-7.00005-3
41. T. Lazaridis, M. Karplus, Effective energy function for proteins in solution. Proteins **35**, 133–152 (1999). https://doi.org/10.1002/(sici)1097-0134(19990501)35:2<133::aid-prot1>3.0.co;2-n
42. B. Roux, T. Simonson, Implicit solvent models. Biophys. Chem. **78**, 1–20 (1999). https://doi.org/10.1016/s0301-4622(98)00226-9
43. U. Haberthür, A. Caflisch, FACTS: fast analytical continuum treatment of solvation. J. Comput. Chem. **29**, 701–715 (2008). https://doi.org/10.1002/jcc.20832
44. M. Orozco, F.J. Luque, Theoretical methods for the description of the solvent effect in biomolecular systems. Chem. Rev. **100**, 4187–4226 (2000). https://doi.org/10.1021/cr990052a
45. T. Luchko, S. Gusarov, D.R. Roe, C. Simmerling, D.A. Case, J. Tuszynski, A. Kovalenko, Three-dimensional molecular theory of solvation coupled with molecular dynamics in Amber. J. Chem. Theory Comput. **6**, 607–624 (2010). https://doi.org/10.1021/ct900460m
46. R. Anandakrishnan, A. Drozdetski, R.C. Walker, A.V. Onufriev, Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. Biophys. J. **108**, 1153–1164 (2015). https://doi.org/10.1016/j.bpj.2014.12.047
47. J. Hermans, H.J.C. Berendsen, W.F. Van Gunsteren, J.P.M. Postma, A consistent empirical potential for water–protein interactions. Biopolymers **23**, 1513–1518 (1984). https://doi.org/10.1002/bip.360230807
48. A.D. MacKerell Jr., J. Wiorkiewicz-Kuczera, M. Karplus, An all-atom empirical energy function for the simulation of nucleic acids. J. Am. Chem. Soc. **117**, 11946–11975 (1995). https://doi.org/10.1021/ja00153a017
49. K.-H. Ott, B. Meyer, Parametrization of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations. J. Comput. Chem. **17**, 1068–1084 (1996). https://doi.org/10.1002/(SICI)1096-987X(199606)17:8<1068::AID-JCC14>3.0.CO;2-A
50. A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B **102**, 3586–3616 (1998). https://doi.org/10.1021/jp973084f
51. W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. **117**, 5179–5197 (1995). https://doi.org/10.1021/ja00124a002
52. G.A. Kaminski, R.A. Friesner, J. Tirado-Rives, W.L. Jorgensen, Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate

quantum chemical calculations on peptides. J. Phys. Chem. B **105**, 6474–6487 (2001). https://doi.org/10.1021/jp003919d

53. M. Rueda, C. Ferrer-Costa, T. Meyer, A. Pérez, J. Camps, A. Hospital, J.L. Gelpí, M. Orozco, A consensus view of protein dynamics. Proc. Natl. Acad. Sci. U S A **104**, 796–801 (2007). https://doi.org/10.1073/pnas.0605534104
54. A. Perez, F. Lankas, F.J. Luque, M. Orozco, Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Res. **36**, 2379–2394 (2008). https://doi.org/10.1093/nar/gkn082
55. B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, M. Karplus, CHARMM: the biomolecular simulation program. J. Comput. Chem. **30**, 1545–1614 (2009). https://doi.org/10.1002/jcc.21287
56. M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX **1–2**, 19–25 (2015). https://doi.org/10.1016/j.softx.2015.06.001
57. D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, R.J. Woods, The Amber biomolecular simulation programs. J. Comput. Chem. **26**, 1668–1688 (2005). https://doi.org/10.1002/jcc.20290
58. M.T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L.V. Kalé, R.D. Skeel, K. Schulten, NAMD: a parallel, object-oriented molecular dynamics program. Int. J. Supercomput. Appl. High Perform. Comput. **10**, 251–268 (1996). https://doi.org/10.1177/109434209601000401
59. P. Larsson, B. Hess, E. Lindahl, Algorithm improvements for molecular dynamics simulations. WIREs Comput. Mol. Sci. **1**, 93–108 (2011). https://doi.org/10.1002/wcms.3
60. M.J. Harvey, G. Giupponi, G.D. Fabritiis, ACEMD: accelerating biomolecular dynamics in the microsecond time scale. J. Chem. Theory Comput. **5**, 1632–1639 (2009). https://doi.org/10.1021/ct9000685
61. H. Shukla, R. Shukla, A. Sonkar, T. Pandey, T. Tripathi, Distant Phe345 mutation compromises the stability and activity of mycobacterium tuberculosis isocitrate lyase by modulating its structural flexibility. Sci. Rep. **7**, 1058 (2017). https://doi.org/10.1038/s41598-017-01235-z
62. H. Shukla, R. Shukla, A. Sonkar, T. Tripathi, Alterations in conformational topology and interaction dynamics caused by L418A mutation leads to activity loss of mycobacterium tuberculosis isocitrate lyase. Biochem. Biophys. Res. Commun. **490**, 276–282 (2017). https://doi.org/10.1016/j.bbrc.2017.06.036
63. R. Shukla, H. Shukla, T. Tripathi, Activity loss by H46A mutation in mycobacterium tuberculosis isocitrate lyase is due to decrease in structural plasticity and collective motions of the active site. Tuberculosis **108**, 143–150 (2018). https://doi.org/10.1016/j.tube.2017.11.013
64. R. Shukla, T.R. Singh, Virtual screening, pharmacokinetics, molecular dynamics and binding free energy analysis for small natural molecules against cyclin-dependent kinase 5 for Alzheimer's disease. J. Biomol. Struct. Dyn. **38**, 248–262 (2020). https://doi.org/10.1080/07391102.2019.1571947
65. R. Shukla, T.R. Singh, High-throughput screening of natural compounds and inhibition of a major therapeutic target HsGSK-3β for Alzheimer's disease using computational approaches. J. Genet. Eng. Biotechnol. **19**, 61 (2021). https://doi.org/10.1186/s43141-021-00163-w
66. R. Shukla, P.B. Chetri, A. Sonkar, M.Y. Pakharukova, V.A. Mordvinov, T. Tripathi, Identification of novel natural inhibitors of opisthorchis felineus cytochrome P450 using structure-based screening and molecular dynamic simulation. J. Biomol. Struct. Dyn. **36**, 3541–3556 (2018). https://doi.org/10.1080/07391102.2017.1392897
67. R. Shukla, H. Shukla, T. Tripathi, Structural and energetic understanding of novel natural inhibitors of Mycobacterium tuberculosis malate synthase. J. Cell. Biochem. **120**(2), 2469–2482 (2019). https://doi.org/10.1002/jcb.27538

68. J. Kalita, R. Shukla, T. Tripathi, Structural basis of urea-induced unfolding of Fasciola gigantica glutathione S-transferase. J. Cell. Physiol. **234**, 4491–4503 (2019). https://doi.org/10.1002/jcp.27253
69. A. Sonkar, H. Shukla, R. Shukla, J. Kalita, T. Tripathi, Unfolding of acinetobacter baumannii MurA proceeds through a metastable intermediate: a combined spectroscopic and computational investigation. Int. J. Biol. Macromol. **126**, 941–951 (2019). https://doi.org/10.1016/j.ijbiomac.2018.12.124
70. P.B. Chetri, R. Shukla, J.M. Khan, A.K. Padhi, T. Tripathi, Unraveling the structural basis of urea-induced unfolding of Fasciola gigantica cytosolic malate dehydrogenase. J. Mol. Liq. **349**, 118170 (2022). https://doi.org/10.1016/j.molliq.2021.118170
71. R. Shukla, T. Tripathi, Molecular dynamics simulation in drug discovery: opportunities and challenges, in *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*, ed. by S.K. Singh, (Springer, Singapore, 2021), pp. 295–316. https://doi.org/10.1007/978-981-15-8936-2_12
72. R. Shukla, T. Tripathi, Molecular dynamics simulation of protein and protein–ligand complexes, in *Computer-Aided Drug Design—An Overview*, ed. by D.B. Singh, (Springer, Singapore, 2020), pp. 133–161. https://doi.org/10.1007/978-981-15-6815-2_7
73. P. Saudagar, T. Tripathi, *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics* (Academic Press, San Diego, 2023)
74. S.B. Syed, F.I. Khan, S.H. Khan, S. Srivastava, G.M. Hasan, K.A. Lobb, A. Islam, M.I. Hassan, F. Ahmad, Unravelling the unfolding mechanism of human integrin linked kinase by GdmCl-induced denaturation. Int. J. Biol. Macromol. **117**, 1252–1263 (2018). https://doi.org/10.1016/j.ijbiomac.2018.06.025

pH-Based Molecular Dynamics Simulation for Analysing Protein Structure and Folding

Santanu Sasidharan, Rohit Shukla, Timir Tripathi, and Prakash Saudagar

Abstract The structure and function of a protein are influenced by environmental factors like pH, temperature, salt concentrations, etc. The intrinsic dynamics of a protein in such environments involve temporal and spatial changes at the atomic level. These changes can be understood with the help of molecular dynamics (MD) simulations. This chapter concentrates on the MD simulation of proteins at constant pH, allowing researchers to bridge the gap. The constant pH approach accounts for the protonation states of the amino acid residues in a protein while receiving little or no inputs from the forcefields employed for the simulation. Once completed, the simulations provide valuable data on the folding process of a protein and the free energy of binding between the protein and other interacting molecules (another protein, ligand, DNA, etc.). The chapter introduces the concept of constant pH simulation and the effect of pH on each amino acid. The effect of the pH environment on the spatiotemporal arrangement of the proteins and the presence of intermediate states have been discussed. We then provide insights into the practical aspects of the all-atom simulation of the constant pH approach and the analysis of the trajectories using the MD simulation data. Case studies are provided

S. Sasidharan
Department of Biotechnology, National Institute of Technology Warangal, Warangal, India

Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada

R. Shukla
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

T. Tripathi
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

Regional Director’s Office, Indira Gandhi National Open University (IGNOU), Regional Centre Kohima, Kohima, India

P. Saudagar (✉)
Department of Biotechnology, National Institute of Technology Warangal, Warangal, India
e-mail: ps@nitw.ac.in

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_11

to help understand the nuances of different MD simulation systems and how they vary. In the end, we discuss the future directions to further research in this area and achieve MD simulation results with higher accuracy and reliability.

Keywords Constant pH simulations · Molecular dynamic simulations · Protein structure · Unfolding

1 Introduction

Protein folding is an important cellular event that governs several biological functions and regulations [1, 2]. Any protein must fold to its native state once it emerges from the ribosomes, and throughout its lifetime, it undergoes a series of folding and unfolding in the form of conformational fluctuations. The process of folding is intensely studied because of its importance and association with several diseases and disorders. The field of protein folding has passed over five decades, but we have not reached a general agreement on the folding pathways of proteins, and several questions remain unanswered. The folding of the proteins was assumed to be a straightforward biophysical event, which was later deferred, and scientists discovered that the event happens through a series of distinct intermediate states. Anfinsen later demonstrated that the proteins fold without external help [3]. At the same time, Levinthal ascertained that the undirected folding process follows a predetermined route to reach the native state and that the event cannot be a random process [4, 5]. The realisation that the folding event is not random and the randomness might lead to an ensemble of partially folded proteins led scientists to infer folding as a unique process through multiple unpredictable routes with several intermediate conformations. The thermodynamic hypothesis of Anfinsen further extended our understanding using the protein-funnel shape energy landscape, where the proteins are predicted to fold energetically downhill [3]. The protein energy funnel does not provide any realistic constraints for the real-time scenario, but it has been widely accepted that the protein energy decreases as the conformation reaches the native state and that there are several independent pathways leading to it [6].

Several factors influence the folding of a protein, such as temperature, pH, salts, macromolecular crowding, etc. Among these, pH has been of interest since cellular pH is highly susceptible to changes. Solution pH is often the most critical factor for the proper functioning of a protein and also catalysis. Both structure and function are strongly influenced by the solution pH due to the changes brought about by the protonation states of the side chains of the amino acid residues that make up the protein. The protonation states of the side-chain group are determined by the pH of the solution and the relative acidity of the group (measured by pK_a). The pK_a of any side-chain titratable group is influenced by the electrostatic environment, which in turn is determined by the conformation of the protein and the protonation states of titratable groups of the other side-chain amino acids. When there is a charge difference between the protonation states, the net result is a change in protein conformations. It is well established that there is a tight connection between protein conformations and the protein protonation state, which is decided by the solution pH. The use of molecular dynamics (MD) simulations to study this effect is of increasing interest among computational biologists [7, 8]. In this chapter, we discuss the importance of protein folding, the impact of pH on protein folding, and simulating proteins at varying pH conditions.

2 Protein Folding and the Intermediate States

The proteins have the ability to fold spontaneously, and the correct fold is necessary for functional activity. Several studies have determined that the information for the folding of a protein is specified in the linear amino acid sequence. While we know that each amino acid has a discrete set of backbone states, this does not limit the astronomical possibilities of folding a protein. One relief comes from the understanding that the entire conformational space of the folding states of a protein is not flat but a funnel hole, where the near-native conformations are present toward the bottom of the funnel. The downhill direction is obtained when the energy of the intermediate structures is lower than the previous structures, and finally, the structure with the lowest energy is deemed native. The force that drives the folding of the globular proteins is assumed to be the result of the burial of hydrophobic amino acids, i.e. to keep the hydrophobic side chains away from the water with minimal contact [9–11]. There is a loss of conformational entropy, thereby causing the collapse of the protein into the predefined three-dimensional (3D) structure. The tight packing of the non-polar amino acids results in increased van der Waals interactions and also reduces the unfavourable cavities. The intra-hydrogen bonds and the salt bridges formed in the protein largely compensate for the loss of interactions with water molecules surrounding the protein. Apart from these interactions, the polar amino acids also contribute to the stability of the protein by interacting with the solvent molecules and defining the specificities of the protein [12].

As discussed earlier, the folding of a protein is not a straightforward process. The polypeptide chain overcomes the Levinthal paradox to reach the required fold, and several models have been proposed to understand the phenomenon [13]. The unfolding and refolding of a protein under equilibrium conditions have been understood to be a two-way process where the significant population is either the native or unfolded state. The intermediate states that are observed are usually unstable and poorly populated when the conditions are at equilibrium. The two-state folding of a protein is favourable for small proteins, and in this case, the free energy of denaturation can also be determined [14]. The existence of intermediates has been shown through kinetic studies, even if small proteins have been described to have two-state folding. The intermediate states differ from protein to protein and even during folding and unfolding. It has been shown that protein may exhibit monophasic (two-state) unfolding but multiphasic refolding. Therefore, there are several ways a protein might unfold and refold, and to characterise the route, it is essential to obtain the structural conformations of the intermediates [15, 16]. The kinetic studies of folding/unfolding might reveal several intermediates, but their structural characterisation is complex because of their low population, transient accumulation, rapid process, and high cooperativity. The effect of external factors like pH and temperature causes the unfolding and refolding of the proteins, but as seen earlier, it is difficult to determine their structure [17]. It is at this point that in silico simulation tools come in handy. These methodological tools allow us to simulate the protein under various conditions and understand the folding/refolding process and the intermediate states [18].

3 Effect of pH on Amino Acids

The effect of pH on the amino acids in a protein sequence has been long studied. It is an important parameter since the intramolecular interactions and the fold of a protein are based on how the protein interacts with its environment. The pK_a of the amino acids is calculated at different pHs; accordingly, the titratable groups in the amino acids are protonated and deprotonated. The parts of the amino acids that can accept or release a proton are called the titratable groups. For example, aspartic acid has one side-chain carboxyl group, and therefore, it has one titratable side-chain group. In contrast, N-terminal lysine has two titratable groups, i.e. the N-terminal amino group and the side-chain amino group. On this basis, amino acids are classified into two categories: acids and bases. Amino acids, such as Asp and Glu, are acids, while His, Lys, and Arg are bases. The acidic amino acids are neutral in their protonated states, while basic amino acids are positively charged in their protonated state.

The pK_a value is the $-\log(K_a)$, and if we know the pK_a value of the titratable groups in a protein, we can predict the charge on the side chain of the amino acids present in the group. The p*K*a values of the titratable groups in water have been estimated and are given in Table 1. The Henderson Hasselback equation is usually used to determine the titration curves using the rearranged equation:

Table 1 The pK$_a$ values of amino acid residues in a polypeptide [19]

Titratable group	Estimated pK$_a$
N-terminal	8.0
C-terminal	3.0
Asp	4.0
Glu	4.3
Cys	8.7
Tyr	9.8
Ser	14.2
Thr	15
Arg	13
Lys	10.5

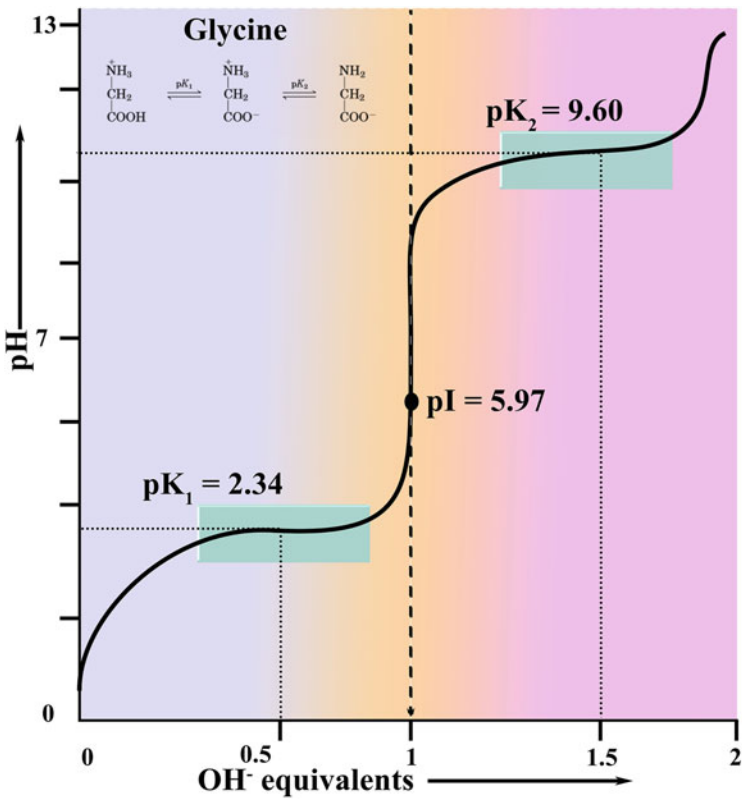

Fig. 1 Titration curve of glycine. The titration curve of 0.1 M glycine at 25 °C is shown. The different ionic species that are the key points in the titration curve are shown at the top. The teal boxes show the pK_1 and pK_2 of glycine and indicate the region with the highest buffering capacity

$$f_{\mathrm{HA}} = \frac{1}{10^{\mathrm{pH} - \mathrm{p}K_{\mathrm{a}}} + 1}$$

The titration curve can be obtained by plotting f_{HA} (where HA is the acid) versus pH. The titration curve of amino acid glycine is shown in Fig. 1.

As discussed earlier, the side chain of amino acid residues in a protein may have titratable groups. We limit our calculations and predictions, therefore, to the pK_{a} values of the side chains of the acidic and basic amino acids. The pK_{a} values of titratable groups are usually measured as a difference in the free energy of the neutral and the charged state of the titratable groups. Calculating the free energy difference between the states is possible, which involves three steps.

1. The desolvation energy associated with moving the charged and the neutral form of the titratable groups from the water to the interior position in the protein.

2. The interaction of the neutral and the charged titratable groups with the permanent dipole of the protein.
3. The pair-wise interaction between the titratable groups.

The pK_a calculations are important because the 3D structure of a protein is dependent on the pK_a values of the amino acid side chains. However, the 3D structures obtained through the X-ray crystallography are perturbed by the environment of the crystal, and therefore, the pK_a values are less accurate for the residues involved in the crystal contacts.

4 Simulating Proteins at Multiple pHs

To prepare the protein for pH-based simulations, approaches such as the titration of only acidic or basic amino acids or setting the pH of the system with the explicit solvent, etc. can be applied. The constant pH molecular dynamics (MD) simulation is a common method widely used. The addition of the ions (H^+ and OH^-) for determining the pH of the solvent in the simulation box, which can hinder the pK_a of the amino acids, is a complicated process and leads to various artefacts in the trajectory or simulation results. Therefore, it is an excellent way to consider only the titratable groups of the protein. The combination of different titratable groups at a particular pH can show the different pK_a values compared to its experimental value. To tackle this problem, several strategies have been used to generate an ensemble of structures at different pH [20, 21]. These structures have a fixed protonation state according to the hypothetical pH assumed for the study [22–25]. However, there are limitations to this method since the side-chain pK_a values of the positively and negatively charged residues shifts based on the surrounding electrostatic environment of the protein. Even if the pK_a of all the side chains at a particular pH is known, the conformational sample with all combinations needs to be explored. Also, the fixed protonation of the protein sometimes does not allow us to understand the pH-induced changes, such as the binding site catalysis mechanism and unfolding mechanism of the protein, because simultaneously, several titratable groups participate in the structural changes.

Another approach that can be applied to the simulation is the continuous protonation state along the continuous titration coordinate λ [26, 27]. In this approach, the protonation state can be changed at a given periodic interval based on the Monte Carlo Metropolis criterion [28–32]. The titratable groups and the titratable protons are defined explicitly, and the state list is defined for each residue. The simulation is then run with fixed protonation states for a definite time period, and the protonation states can change based on the Metropolis criterion again over time.

In this chapter, we will discuss the first approach, where the titratable groups have a fixed pK_a value depending upon the assumed pH in the starting structure. The structure of the protein can either be obtained from protein structure databases or can be modelled [33]. The structure can be used in a web-based server such as H++

(http://newbiophysics.cs.vt.edu/H++/) to set the pH. Various standalone tools, such as YASARA, GROMACS, and Desmond, can also set the protonation states. The YASARA and Desmond are GUI-based MD simulation software where the user can define the pH of the system, so it will automatically protonate the target structure based on that pH and then run the simulation. The GROMACS is the command line Linux-based widely used tool for molecular simulations, where also the user can set the pH. In the GROMACS, the user can assign the titratable residues for the protonation using the *gmx pdb2gmx* tool using the *-ter* option.

5 Case Study of *Leishmania donovani* Tyrosine Aminotransferase (LdTAT) Enzyme

We now describe, in detail, a case study by using the example of *Leishmania donovani* tyrosine aminotransferase (LdTAT) enzyme simulated at three different pH 2, 7, and 12 using GROMACS v5.1.4 [34–37]. The enzyme LdTAT was initially uploaded to the H++ server as described in the earlier section [38] to set the protonation state at a given pH [39, 40]. The authors used three different pH conditions for the input generation for the MD simulation. At pH 2, the side chains of the protein amino acids were highly protonated, while at pH 12, they were highly deprotonated. At pH 7, all amino acids remained neutral except for His, which was charged. The protein structure was then simulated using the Amber 99 forcefield [35]. A simulation box was constructed and filled with explicit water ($n = 24{,}692$) using the TIP4P water model. The system was then neutralised by adding Na^+ and Cl^- ions, followed by energy minimisation. Then 1 ns NVT (number of constant volume and temperature) simulation was run to set the volume and temperature of the simulation box. After this, the 1 ns NPT (number of constant pressure and temperature) was performed to fix the pressure of the system. Finally, all the equilibrated systems were utilised for the 100 ns detailed MD simulation. The authors calculated the root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (Rg), solvent accessible surface area (SASA), number of hydrogen bonds, and energy by using various GROMACS inbuilt utilities such as gmx rms, gmx rmsf, gmx gyration, gmx sasa, gmx h-bonds, and gmx energy, respectively. The principal component analysis (PCA) was carried out using the gmx covar tool, and then the 2D PCA analysis was carried out using the gmx anaeig tool.

The analysis of the MD simulation results from various angles was carried out to explore the pH-induced changes. The data analysis can provide detailed atomistic level alterations induced by pH. Therefore, it is crucial to analyse the simulation trajectory carefully. In the simulation of LdTAT at different pH, the first and foremost part was to evaluate the primary and secondary structures of the protein. LdTAT possesses an N-terminal and C-terminal domain which comprises 14 α-helixes and 9 β-sheets. It represents a proper folded and globular structure.

The N-terminal contains 9 positively and negatively charged residues, while the C-terminal contains 5 and 6 positively and negatively charged residues, respectively. The active site of the protein has Lys286, which is essential for co-factor binding. The detailed analysis showed that LdTAT has 52 negatively charged residues and 44 positively charged residues [41]. The authors calculated various structural parameters such as stability analysis (RMSD, RMSF, Rg), unfolding analysis (SASA), correlated motions analysis (PCA), and structural changes by using the time-dependent secondary structure calculation. They also calculated hydrogen bonds and different energetic terms and correlated them with the LdTAT stability.

5.1 Root Mean Square Deviation

The RMSD was calculated to predict the stability of the MD simulation as well as to evaluate the structure stability at different pHs. The RMSD was calculated by superposing the first frame to the corresponding time frame snapshot for the complete trajectory, as shown in Fig. 2a. At pH 2 and 12, the LdTAT enzyme attained stability after 20 ns, while at pH 7, the structure stabilised after 10 ns. The results showed that the enzyme remained stable throughout the simulation period, even at pH 2 and 12, and no unfolding was observed (Fig. 2a). At pH 2 and 7, the RMSD value was observed in a similar manner. The trajectory analysis revealed that the enzyme is stable at extreme pH conditions; therefore, no unfolding was observed. It should be noted that if the pH induces the unfolding in a protein structure, there would be large fluctuations in RMSD which would reflect in the other trajectory results.

5.2 Radius of Gyration

The Rg represents a protein's compactness or folding status, and perturbations in the Rg values refer to unfolding events in the structure. The higher and lower Rg values represent properly folded and unfolded structures. The Rg plot analysis of LdTAT did not show major changes between all three conditions. At all three pHs, the protein was compactly folded until 60 ns, and after 60 ns, a slight fold change was observed. The change in Rg after 60 ns might result from charged residues in the N and C terminals hindering the intramolecular interactions (Fig. 2b). The differences observed in Fig. 2b were insufficient for proposing the unfolding mechanism of the LdTAT at these pHs. Any large unfolding event would reflect in the Rg analysis with a steep increase in the plot curve with time. The Rg result also well agreed with the RMSD analysis and indicates that the protein is stable at both low and high pHs.

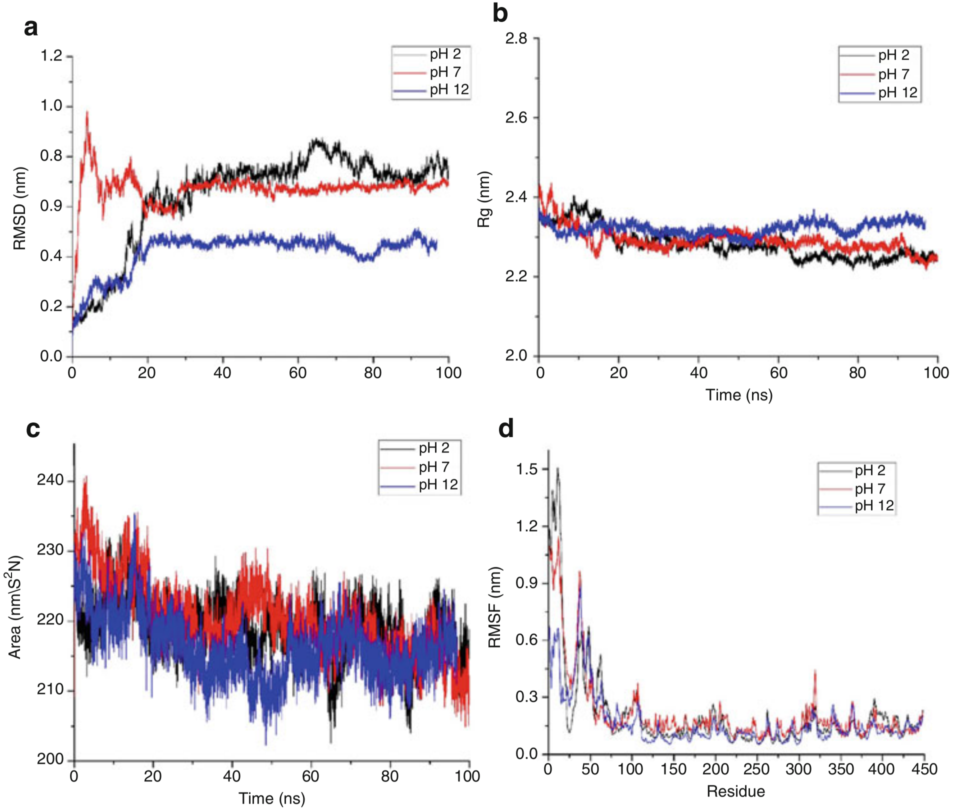

Fig. 2 Stability analysis of LdTAT at constant pH simulation at 310 K. (**a**) RMSD, (**b**) Rg, (**c**) SASA, (**d**) RMSF [34]. The black, red, and blue colour represent pH 2, 7, and 12, respectively

5.3 Solvent Accessible Surface Area

The SASA refers to the surface area of LdTAT that is exposed to the solvent during the simulation period. When a protein is folded in a particular conformation, the SASA values remain constant and do not vary. In contrast, the unfolding of a protein leads to larger surface areas being exposed to the solvent. The higher and lower SASA represents the folded and unfolded LdTAT structure. LdTAT enzyme at three different pH did not exhibit any large fluctuations in the SASA values, but at pH 7, the LdTAT folded more compactly than at pH 2 and 12 (Fig. 2c). The reason provided is that the LdTAT enzyme attains a properly folded structure at pH 7, while at pH 2 and 12, there might be minor changes in the intramolecular interactions resulting in higher SASA values. The result is in good agreement with the RMSD and Rg analysis, where the authors did not observe any major changes.

5.4 Root Mean Square Fluctuation

The RMSF represents the residue level deviation in the protein throughout the simulation time scale. This allows us to observe the structural changes in specific regions of the protein during the simulation period. It can also provide the atomic level changes if one analyses the single atom movement at different conditions. The RMSF is usually calculated by comparing the flexibility of the backbone chain per residue and the flexibility of the initial structure. LdTAT showed higher fluctuations in the N-terminal region at pH 2 and 7, while lower RMSF values were observed at pH 12. In the overall simulation, a lower RMSF value for the LdTAT at pH 12 while a higher RMSF value for pH 7 was observed (Fig. 2d). The RMSF analysis indicates that the LdTAT enzyme has a flexible N-terminal region that might have a potential functional consequence in the enzymatic activity. It was observed from the RMSF analysis that residues have different flexibility patterns for proper folding at different pHs. However, other analyses, such as RMSD, Rg, and SASA, did not show any major changes at different pH conditions.

5.5 Secondary Structure Analysis

One of the most critical analyses of the constant pH MD simulations is the secondary structure analysis that reveals the unfolding mechanism of the protein. The secondary structure calculation was carried out using the do_dssp tool with respect to time. The tool extracts the secondary structure of the protein at a particular time interval and provides a percentage of the overall secondary structure. LdTAT enzyme simulation at three different pHs showed a similar α-helix and β-sheet content of 26% and 12% at pH 2 and 12. While at neutral pH, the α-helix and β-sheet content

Table 2 Prediction of secondary structures

pH	Coils (%)	β-sheets (%)	β-bridges (%)	Bends (%)	Turns (%)	α-helices (%)	5-helices (%)	3-helices (%)
2	24	12	1	11	16	26	0	9
7	24	14	1	12	14	29	0	6
12	25	12	1	11	16	26	0	8

The percentage of α-helix, β-sheet, coil, turns, 5-helix, 3-helix, and other secondary structures predicted at pH of 2, 7, and 12 for a 100 ns MD run are given. The variations in α-helix, β-sheet, coil, and turn can be observed in the table

was higher at 29% and 14%, respectively. The pH-induced changes are shown in Table 2. It indicates that the molecular interactions are breaking at low and high pH, which leads to the loss of stable secondary structures. However, these changes were not enough to explore the unfolding mechanism of the LdTET enzyme.

5.6 Intramolecular Hydrogen Bonding and Internal Energy Analysis

Intramolecular hydrogen bonding is a vital parameter that determines the intactness and compactness of the protein. The number of intramolecular H bonds was calculated with respect to time for exploring the compactness of the LdTAT at different pH. In LdTAT, the intramolecular hydrogen bonding pattern was almost similar at all three pHs, which represents the stability of the enzyme. The authors predicted the average number of hydrogen bonds as 319 ± 9, 317 ± 11, and 315 ± 9 at pH 3, 7, and 12, respectively. The result corroborated the earlier results of Rg, RMSD, and SASA, where large fluctuations in the structure were not observed.

The authors then calculated the intramolecular energy to analyse the internal interaction changes in the protein at different pH conditions. This analysis can also reveal the interaction patterns and folding patterns of the enzyme at different pHs. They calculated two energetic terms: Columbic (electrostatic) and Lennard-Jones. They observed the Lennard-Jones interaction energies of $1.74 \times 10^5 \pm 21$ kJ/mol, $3.0 \times 10^5 \pm 11$ kJ/mol, and $1.74 \times 10^5 \pm 12$ kJ/mol at pH 2, 7, and 12, respectively. The result analysis showed that Lennard-Jones interaction energies were low at pH 2 and 12, which indicates the loss of intramolecular interactions in the LdTET enzyme. The authors observed a similar pattern in the case of secondary structure analysis, where they saw a loss in secondary structures at pH 2 and 12. The Coulombic energy was also calculated at pH 2, 7, and 12, which was $-1.37 \times 10^6 \pm 27$ kJ/mol, $-2.14 \times 10^6 \pm 40$ kJ/mol, and $-1.37 \times 10^6 \pm 39$ kJ/mol, respectively. The Coulombic energy change at pH 2 and 12 indicates the loss of salt bridge interactions. The overall result of the energy showed that at pH 2 and 12, LDTET is losing intramolecular interactions while these interactions are stable at neutral pH.

5.7 Principal Component Analysis

The principal component analysis gives the large-scale motions that are essential for protein dynamics. The first two principal components (PCs) were selected for the analysis and plotted in phase space, as shown in Fig. 3. Here the large and dispersed cluster represents the less folded protein, while the compact cluster represents the well-folded and compact protein. The 2D PCA of LdTAT at pH 2 showed a stable cluster compared to pH 7. The LdTAT showed a similar cluster at pH 7 and 12. The dispersed cluster at pH 2 might result from the large-scale motions observed in the N-terminal and the C-terminal in the RMSF trajectory results.

Few other in-depth analyses, like the distance between two titratable groups, salt bridge analysis, and intermolecular hydrogen bonds, can also be carried out in the case of protein–ligand interactions [42, 43]. However, the most critical analyses required for constant pH MD simulations are covered in this chapter through the above-described case study of the LdTAT enzyme.

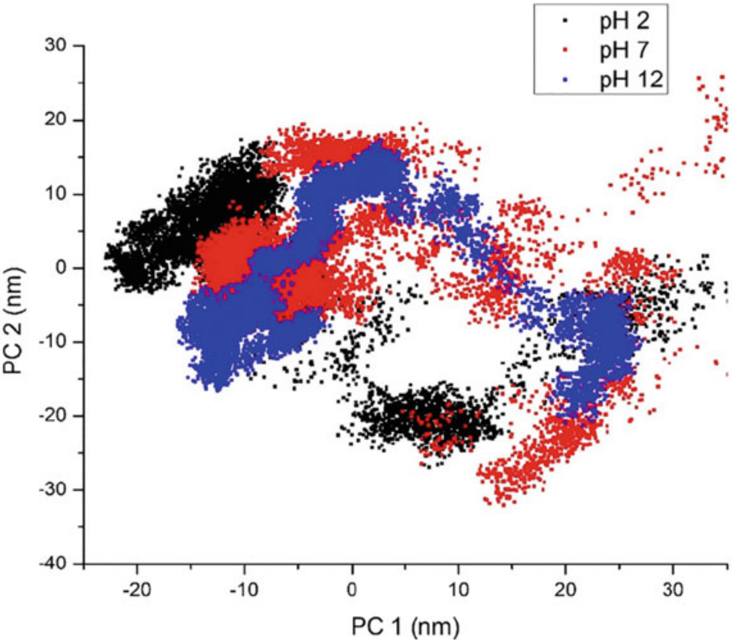

Fig. 3 Principal component analysis. The PC analysis of LdTAT at pH 2 (black dots), pH 7 (red), and pH 12 (blue) is shown. All pH exhibited similar clustering, and the area covered by the vectors was also the same, signifying that the structure remains relatively stable at all three pHs. The large-scale motions were observed only at pH 2, while the structures were relatively rigid at pH 7 and 12

6 Other Case Studies

Though we discussed the constant pH simulation of the LdTAT enzyme at three different pHs, several other studies are available for reference. Waseem et al. simulated irisin, a therapeutic protein involved in several diseases, at pH 2, 4, 6, 7.5, 10, and 12 [24]. They observed that apart from the stability of the protein at the isoelectric point, the protein exhibited higher stability at pH 4 and 6. Leone and Picone demonstrated the design of a pH-stable mutant of a sweet protein called MNEI [44]. Recently, a multi-spectroscopic approach by Yousuf et al. showed that the protein cyclin-dependent kinase 6 (CDK6) is stable between pH 7 and 8, and the tertiary structure remained intact over the complete alkaline range [25]. Another combined spectroscopic and MD simulation study revealed that the calcium/calmodulin-dependent protein kinase IV is stable over the pH range of 5–11.5 and the secondary and tertiary structures were also stable. While at pH 2 to 4.5, the study found a significant aggregation of the protein [23]. Hofer et al. studied the pH-induced unfolding of PhIp 6 pollen allergen protein from constant pH MD simulations. The study used extensive simulation data using the Markov state models and retrieved detailed thermodynamic and kinetic information at different pHs [45]. Another interesting study on constant pH simulations was conducted on the T7 RNA polymerase enzyme. The study concluded that the structural interaction of T7 RNA polymerase changes with pH, while the C-terminal end plays a vital role, and its inefficiency was recorded at lower pH [46]. Zhou et al. studied the pH-induced misfolding of prion protein and derived the unfolding mechanism using microsecond MD simulation analysis. The study used accelerated MD simulations clubbed with the Markov state model [47]. Khan et al. performed constant pH simulations of chitinase II isolated from *Thermomyces lanuginosus*. They observed the strong conformational dependence of chitinase II on the pH alteration [48]. Another interesting pH-induced conformational transition of prion protein was studied, and the effect of the protonation of the His residues was studied in detail [49]. The authors found that the protonation of His155 and His187 is crucial for the conformational rearrangement of the structure [49]. Apart from these constant pH simulation methods, several other methods of pH-based simulations for pK_a calculations have also been introduced [50–53].

7 Conclusions and Future Perspectives

Current approaches for constant pH MD simulation are powerful for addressing the proteins that have variable protonation states. The advances over the years have made the available theoretical data and the implementation of simulation techniques much easier for routine applications. However, several challenges remain in the accuracy and reliability of conformational sampling. Constant efforts are being made to solve these issues, and with time, we will be able to perform constant pH simulations with more accuracy and reliability and obtain better information from the simulation trajectories. Another vital advancement getting attention is the computation of titration curves from the simulation output [52]. It will provide in-depth insights for tracking the behaviour of individual residues in a protein as well as the sites that are involved in the functioning of the protein. The data obtained will help us understand the protein structure and function at different protonation states and solve the intrinsic dynamics of the biochemistry of proteins [18].

Another major problem that affects the constant pH simulation is the availability of the specific forcefields. The constant pH simulations currently model the interactions between the titratable residues with the help of physics-based potential, and there are no alternatives to the empirical and ad hoc descriptions. However, there are well-established equations, such as the Hill equation, that can be utilised considering the wide use of such approximations and the ease of employment of the equation. One can always use direct computational correlations, but the approach requires user-specific insights into the model being considered. Continuous efforts are being made to adapt additional forcefields in the perspective of constant pH MD simulations. These are straightforward and substantial, with the only barriers being in the initial topology and parameters. The future direction in this field should also be towards special sampling techniques and the incorporation of specific forcefields.

References

1. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Springer Nature, Singapore, 2020)
2. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods*, 1st edn. (Academic Press, Cambridge, MA, 2022)
3. C.B. Anfinsen, E. Haber, M. Sela, F. White Jr., The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U S A **47**(9), 1309 (1961)
4. C. Levinthal, Are there pathways for protein folding? J. Chim. Phys. **65**, 44–45 (1968)
5. C. Levinthal, How to fold graciously. Mossbauer Spectrosc. Biol. Syst. **67**, 22–24 (1969)
6. S.W. Englander, L. Mayne, The nature of protein folding pathways. Proc. Natl. Acad. Sci. **111**(45), 15873–15880 (2014)
7. R. Shukla, T. Tripathi, Molecular dynamics simulation of protein and protein-ligand complexes, in *Computer-Aided Drug Design*, ed. by D.B. Singh, (Springer Nature, Singapore, 2020), pp. 133–161
8. R. Shukla, T. Tripathi, Molecular dynamics simulation in drug discovery: opportunities and challenges, in *Innovations and Implementations of Drug Discovery Strategies in Rational Drug Design*, ed. by S.K. Singh, (Springer Nature, Singapore, 2021), pp. 295–316
9. L. Monticelli, S.K. Kandasamy, X. Periole, R.G. Larson, D.P. Tieleman, S.-J. Marrink, The MARTINI coarse-grained force field: extension to proteins. J. Chem. Theory Comput. **4**(5), 819–834 (2008)
10. V. Tozzini, Coarse-grained models for proteins. Curr. Opin. Struct. Biol. **15**(2), 144–150 (2005)
11. G.G. Maisuradze, P. Senet, C. Czaplewski, A. Liwo, H.A. Scheraga, Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field. Chem. A Eur. J. **114**(13), 4471–4485 (2010)

12. B. Kuhlman, P. Bradley, Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol. **20**(11), 681–697 (2019)
13. J. Yon, Protein folding: a perspective for biology, medicine and biotechnology. Braz. J. Med. Biol. Res. **34**, 419–435 (2001)
14. S.C. Harrison, R. Durbin, Is there a single pathway for the folding of a polypeptide chain? Proc. Natl. Acad. Sci. **82**(12), 4028–4030 (1985)
15. T. Tripathi, Calculation of thermodynamic parameters of protein unfolding using far-ultraviolet circular dichroism. J. Protein. Proteomics **4**(2), 85–91 (2013)
16. A. Sonkar, H. Shukla, R. Shukla, J. Kalita, T. Tripathi, Unfolding of Acinetobacter baumannii MurA proceeds through a metastable intermediate: a combined spectroscopic and computational investigation. Int. J. Biol. Macromol. **126**, 941–951 (2019)
17. S. Sasidharan, P. Saudagar, Biochemical and structural characterization of tyrosine aminotransferase suggests broad substrate specificity and a two-state folding mechanism in Leishmania donovani. FEBS Open Bio. **9**(10), 1769–1783 (2019)
18. P. Saudagar, T. Tripathi, *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, 1st edn. (Academic Press, San Diego, 2023)
19. J. Kyte, *Structure in Protein Chemistry* (Garland Publishing, New York, 1995)
20. E.P. O'Brien, B.R. Brooks, D. Thirumalai, Effects of pH on proteins: predictions for ensemble and single-molecule pulling experiments. J. Am. Chem. Soc. **134**(2), 979–987 (2012)
21. J. Khandogin, C.L. Brooks, Toward the accurate first-principles prediction of ionization equilibria in proteins. Biochemistry **45**(31), 9363–9373 (2006)
22. M.A. Haque, S. Zaidi, S. Ubaid-Ullah, A. Prakash, M.I. Hassan, A. Islam, J.K. Batra, F. Ahmad, In vitro and in silico studies of urea-induced denaturation of yeast iso-1-cytochrome c and its deletants at pH 6.0 and 25 C. J. Biomol. Struct. Dyn. **33**(7), 1493–1502 (2015)
23. H. Naz, M. Shahbaaz, K. Bisetty, A. Islam, F. Ahmad, M.I. Hassan, Effect of pH on the structure, function, and stability of human calcium/calmodulin-dependent protein kinase IV: combined spectroscopic and MD simulation studies. Biochem. Cell Biol. **94**(3), 221–228 (2016)
24. R. Waseem, A. Shamsi, M. Shahbaz, T. Khan, S.N. Kazim, F. Ahmad, M.I. Hassan, A. Islam, Effect of pH on the structure and stability of irisin, a multifunctional protein: multispectroscopic and molecular dynamics simulation approach. J. Mol. Struct. **1252**, 132141 (2022)
25. M. Yousuf, A. Shamsi, F. Anjum, A. Shafie, A. Islam, Q.M.R. Haque, A.M. Elasbali, D.K. Yadav, M.I. Hassan, Effect of pH on the structure and function of cyclin-dependent kinase 6. PLoS One **17**(2), e0263693 (2022)
26. M.S. Lee, F.R. Salsbury Jr., C.L. Brooks III, Constant-pH molecular dynamics using continuous titration coordinates. Proteins **56**(4), 738–752 (2004)
27. Y. Huang, R.C. Harris, J. Shen, Generalized born based continuous constant pH molecular dynamics in Amber: implementation, benchmarking and analysis. J. Chem. Inf. Model. **58**(7), 1372–1383 (2018)
28. A.M. Baptista, V.H. Teixeira, C.M. Soares, Constant-p H molecular dynamics using stochastic titration. J. Chem. Phys. **117**(9), 4184–4200 (2002)
29. J. Mongan, D.A. Case, J.A. McCammon, Constant pH molecular dynamics in generalized born implicit solvent. J. Comput. Chem. **25**(16), 2038–2048 (2004)
30. H.A. Stern, Molecular simulation with variable protonation states at constant pH. J. Chem. Phys. **126**(16), 04B627 (2007)
31. J.M. Swails, D.M. York, A.E. Roitberg, Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: implementation, testing, and validation. J. Chem. Theory Comput. **10**(3), 1341–1352 (2014)
32. J.M. Swails, A.E. Roitberg, Enhancing conformation and protonation state sampling of hen egg white lysozyme using pH replica exchange molecular dynamics. J. Chem. Theory Comput. **8**(11), 4393–4404 (2012)

33. S. Sasidharan, P. Saudagar, Prediction, validation, and analysis of protein structures: a beginner's guide, in *Advances in Protein Molecular and Structural Biology Methods*, (Elsevier, London, 2022), pp. 373–385
34. S. Sasidharan, P. Saudagar, Concerted motion of structure and active site charge is required for tyrosine aminotransferase activity in Leishmania parasite. Spectrochim. Acta A Mol. Biomol. Spectrosc. **232**, 118133 (2020)
35. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins **65**(3), 712–725 (2006)
36. S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. Van Der Spoel, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics **29**(7), 845–854 (2013)
37. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, GROMACS: fast, flexible, and free. J. Comput. Chem. **26**(16), 1701–1718 (2005)
38. J.C. Gordon, J.B. Myers, T. Folta, V. Shoja, L.S. Heath, A. Onufriev, H++: a server for estimating pK as and adding missing hydrogens to macromolecules. Nucleic Acids Res. **33**(Suppl_2), W368–W371 (2005)
39. E. Alexov, E.L. Mehler, N. Baker, A.M. Baptista, Y. Huang, F. Milletti, J. Erik Nielsen, D. Farrell, T. Carstensen, M.H. Olsson, Progress in the prediction of p*K*a values in proteins. Proteins **79**(12), 3260–3275 (2011)
40. C.A. Fuzo, L. Degrève, The pH dependence of flavivirus envelope protein structure: insights from molecular dynamics simulations. J. Biomol. Struct. Dyn. **32**(10), 1563–1574 (2014)
41. S. Sasidharan, P. Saudagar, Mapping N-and C-terminals of Leishmania donovani tyrosine aminotransferase by gene truncation strategy: a functional study using in vitro and in silico approaches. Sci. Rep. **10**(1), 1–15 (2020)
42. K. Prince, S. Sasidharan, N. Nag, T. Tripathi, P. Saudagar, Integration of spectroscopic and computational data to analyze protein structure, function, folding, and dynamics, in *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, ed. by P. Saudagar, T. Tripathi, (Academic Press, San Diego, 2023), pp. 483–502
43. A.K. Padhi, S.L. Rath, T. Tripathi, Accelerating COVID-19 research using molecular dynamics simulation. J. Phys. Chem. B **125**(32), 9078–9091 (2021)
44. S. Leone, D. Picone, Molecular dynamics driven design of pH-stabilized mutants of MNEI, a sweet protein. PLoS One **11**(6), e0158372 (2016)
45. F. Hofer, A.S. Kamenik, M.L. Fernández-Quintero, J. Kraml, K.R. Liedl, pH-induced local unfolding of the Phl p 6 pollen allergen from cpH-MD. Front. Mol. Biosci. **7**, 603644 (2021)
46. S. Borkotoky, C. Kumar Meena, G.M. Bhalerao, A. Murali, An in-silico glimpse into the pH dependent structural changes of T7 RNA polymerase: a protein with simplicity. Sci. Rep. **7**(1), 1–12 (2017)
47. S. Zhou, D. Shi, X. Liu, X. Yao, L.-T. Da, H. Liu, pH-induced misfolding mechanism of prion protein: insights from microsecond-accelerated molecular dynamics simulations. ACS Chem. Neurosci. **10**(6), 2718–2729 (2019)
48. F.I. Khan, K. Bisetty, D.-Q. Wei, M.I. Hassan, A pH based molecular dynamics simulations of chitinase II isolated from Thermomyces lanuginosus SSBP. Cogent Biol. **2**(1), 1168336 (2016)
49. E. Langella, R. Improta, V. Barone, Checking the pH-induced conformational transition of prion protein by molecular dynamics simulations: effect of protonation of histidine residues. Biophys. J. **87**(6), 3623–3632 (2004)

50. L. Nilsson, A. Karshikoff, Multiple pH regime molecular dynamics simulation for pK calculations. PLoS One **6**(5), e20116 (2011)
51. E. Socher, H. Sticht, Mimicking titration experiments with MD simulations: a protocol for the investigation of pH-dependent effects on proteins. Sci. Rep. **6**(1), 1–13 (2016)
52. B.K. Radak, C. Chipot, D. Suh, S. Jo, W. Jiang, J.C. Phillips, K. Schulten, B. Roux, Constant-pH molecular dynamics simulations for large biomolecular systems. J. Chem. Theory Comput. **13**(12), 5933–5944 (2017)
53. T. Mizukami, Y. Sakuma, K. Maki, Statistical mechanical model for pH-induced protein folding: application to apomyoglobin. J. Phys. Chem. B **120**(34), 8970–8986 (2016)

Molecular Dynamics Simulation to Study Thermal Unfolding in Proteins

Md Imtaiyaz Hassan, Mohd. Umair, Yash Mathur, Taj Mohammad, Afreen Khan, Md Nayab Sulaimani, Afsar Alam, and Asimul Islam

Abstract The proper folding of a protein is essential for its biological functions. Thermal denaturation of protein structure has been used as an essential tool to understand the unfolding mechanism and measure thermodynamic stability. New technologies have made it feasible to heat proteins using femtosecond laser technology and nanoparticle-targeting methods locally. It is crucial to comprehend how quickly proteins can unfold or lose their function at high temperatures. Protein folding and unfolding have been widely modelled using molecular dynamics (MD) simulations. MD simulations provide information about protein folding that is otherwise impractical through experimental approaches. Techniques like targeted molecular dynamics (TMD) simulations and acid-thermal denaturation correspond to varying degrees of success with experimental observations. These simulations, utilized in tandem with experiments, provide crucial information on the protein folding mechanism. Because of recent computer hardware and software improvements, it is now possible to include a broad range of temperature factors in thermal denaturation studies. In this chapter, we dwell on these details and discuss the thermal unfolding of proteins and their applications. Various computational methods and tools and their uses in protein folding/unfolding studies are described. We also cover an overview of limitations, significant contributions, and recent advancements in MD simulation approaches to study protein folding.

Keywords Protein unfolding · Thermal denaturation · Molecular dynamics simulation · Temperature-induced protein unfolding

M. I. Hassan (✉) · T. Mohammad · A. Khan · M. N. Sulaimani · A. Islam
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: mihassan@jmi.ac.in

M. Umair · Y. Mathur · A. Alam
Department Computer Science, Jamia Millia Islamia, New Delhi, India

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_12

1 Introduction

The main principle of structural biology is that the protein structure regulates its function [1, 2]. During thermal denaturation, a protein unfolds from its native structure upon heating [3]. The energy threshold must be crossed for a protein to reach the point at which its secondary and/or tertiary structures disassemble, leading to an unfolded state [4]. A range of external stimuli, such as strong acids or bases, organic solvents, and heat, can trigger protein denaturation [5, 6]. There are many experimental and theoretical research available to understand the mechanisms governing protein folding and unfolding [7–10]. The thermodynamic properties of unfolding, folding, and conformational changes have been studied using various experimental techniques [11]. However, conventional experimental and theoretical approaches face difficulties elucidating the specifics of the protein folding mechanism at the atomic level [12].

Over the past 20 years, the development of molecular dynamics (MD) simulations has provided deeper insights into the folding/unfolding process [13–15]. Theoretical scientists have worked with experimentalists to use computer modelling approaches to gather atomic-level data regarding the folding/unfolding mechanism [16]. The theoretical conclusions have a good agreement with the experimental data. To unfold a protein, simulations are performed at varying temperatures [17]. MD simulations are frequently used to analyse protein folding features that would be difficult to obtain experimentally [17–21]. Given the difficulties of studying protein stability and folding experimentally, one may opt to use computer simulations [22, 23]. For instance, an investigation of the physical and kinetic changes that occur throughout the simulation is possible because MD simulation integrates comprehensive information at the atomic level with high resolution with time [24]. Since the structure of the unfolded state is unknown, one can start with the native fold and track the feature that allows simulation software to understand protein folding [25–29]. The underlying presumption is that unfolding will mimic protein folding in its latter stages. Applying starting configurations for refolding research can also be done via unfolding simulations.

As computational power increases, we can expand the timescale that can be simulated, enabling simulations of protein denaturation to occur at much more realistic temperatures [30]. Replica-exchange molecular dynamics (REMD), a more sophisticated method, has improved protein folding sampling [31, 32]. It is believed that the free energy landscapes of protein folding in water are at least partially rough [33]. During routine MD simulations, protein systems can become caught in the local energy minima at room temperature. In REMD, several separate simulations are run at various temperatures, and attempts at exchanges are made following the Metropolis criterion, allowing for random travels in the temperature space and elements from existing energy traps [34]. REMD has been effectively used to study the folding of microproteins, helical peptides, three-strand beta-sheets, and hairpin structures [35].

MD simulations have emerged as a crucial tool for studying molecular-level chemical and biological processes [18, 19, 21, 36, 37]. They have been used to examine how proteins and peptides behave in specific environments using Newtonian mechanics and empirical obtained forcefield. This chapter provide a detailed overview to the available computational platforms and methods for studying thermal-induced protein unfolding. In addition, the contribution of MD simulations in illustrating the mechanism of unfolding kinetics is discussed.

2 Effect of Temperature on Protein Structure

Proteins are polymers, more precisely polypeptides created from amino acid sequences, which fold to form a functional three-dimensional (3D) structure. Amino acids undergo condensation processes, losing one water molecule at a time to form peptide bonds with one another to produce a linear chain of a protein. A protein's biological activity is attained by folding into one specific spatial conformation, which is mediated by various non-covalent interactions, including hydrogen bonds, van der Waals forces, ionic interactions, and hydrophobic packing (Fig. 1a). Determining the 3D structure of proteins is essential to comprehend their molecular functions [39–41].

The relationship between organisms and temperature has historically been one of the most active areas of comparative and environmental physiology research [42]. Most of the early studies on temperature–protein interactions concentrated on the influence of temperature on the catalytic rates, i.e. how enzymatic activity adjusts to temperature fluctuations and interspecific differences in the protein thermal

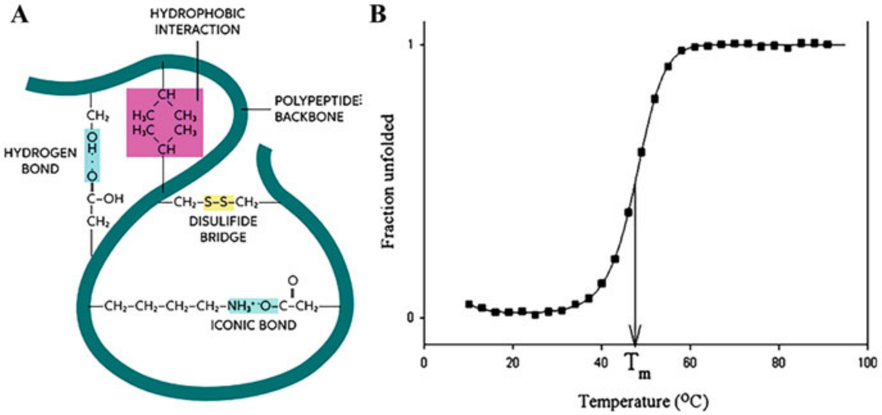

Fig. 1 (**a**) Forces involved in maintaining the 3D structure of a protein. (**b**) Different spectroscopic methods can monitor the thermal denaturation of proteins. The midpoint of this transition is the T_m, referred to as the melting point. The figure is adapted with permission from [38] (https://www.chegg.com/learn/chemistry/introduction-to-chemistry/tertiary-structure-of-protein)

stability [43]. These concepts have increasingly concentrated on two phenomena: adaptive changes in structural and kinetic properties across protein homologues and temperature impacts on protein expression, which may be the key factors in determining an organism's thermal optimum and distribution patterns. There is a requirement for a deeper understanding of the fundamental thermodynamic rules governing protein folding and assembly. Recent theoretical advances enabled the evaluation of the predicted impact of specific amino acid changes on thermal stability using genetically engineered proteins (Fig. 1a).

The free energy difference between the folded and unfolded forms of a protein is represented by the thermodynamic stability of the protein [44]. Since temperature is highly sensitive to the free energy difference, heating may cause unfolding or denaturation. With increasing temperature, molecular vibration increases, breaking the weak interactions and causing proteins to denature. Protein denaturation may cause a loss of natural state and function (Fig. 1b). Typically, soluble globular proteins have a free energy of stabilization of around 10 kJ/mol. When the large number of hydrogen bonds necessary for secondary structure stabilization and the stabilization of the inner core through hydrophobic interactions are considered, the free energy of stabilization appears as a minor difference between very high values [45].

A properly folded protein has a balance between a large number of weak intramolecular interactions (hydrophobic, van der Waals interactions, and electrostatic) and the interactions between the protein and solvent. Hence, the folding process depends on the solution in which the protein resides. These environmental conditions include temperature, salinity, pressure, solvent, etc. [46]. Hence, exposure to extreme conditions (such as heat or radiation, high salt concentrations, strong acids and bases, etc.) can induce a loss in protein structure, leading to denaturation. Although secondary and tertiary structures of a protein are changed during denaturation, the peptide bonds that hold the amino acids together in the core structure are unaffected. The structural levels of a protein determine its function. Therefore, once denatured, the protein can no longer perform its function. However, intrinsically disordered proteins are functionally active despite unfolding in their original state and tend to fold when they bind to a biological target [47].

3 Temperature-Induced Protein Unfolding

Temperature is a crucial and flexible parameter for proteins, as each protein behaves differently under high and low temperatures. Some proteins have high thermal stability, while others can denature or unfold at low-temperature conditions [48]. Many factors like temperature, pH, chemical denaturants, or mechanical stress negatively affect protein stability and can induce conformational changes with an adverse effect on its biological function [49].

It has been observed that proteins unfold at temperatures higher than the basal temperature of the organism it has evolved [50]. The tertiary structure of a protein, essential to its physiological functions, is kept stable by thermodynamic principles. Obtaining the thermodynamic characteristics of protein denaturation as a function of temperature is crucial for understanding the mechanics of protein folding and stability. Temperature also plays a critical role in the kinetics of proteins [51]. Performing MD simulations at different temperatures may help us understand the protein structure and functions to temperature. MD simulation has emerged as an important method for understanding the thermal denaturation of proteins [52].

4 MD Simulation to Understand Protein Denaturation

In 1977, a new era of protein biochemistry started with the MD simulation [53]. The precision and effectiveness of the application of MD simulation to proteins have been continuously improving, and the usage of MD simulation has broadened through the fields of chemistry, biochemistry, molecular biology, physics, and mathematics [54]. MD simulations are tremendously powerful for many reasons. In MD simulation, the motion of every atom at each point in time is captured, which is very difficult with any experimental techniques. The simulation conditions can also be chosen as required for the study, and one can compare the simulation results under different conditions to understand the effects of various molecular perturbations. The force fields significantly impact the outcomes of the MD simulation [55]. The quantitative and qualitative improvements in the force fields, like electrostatics potentials and dihedrals, have also improved the results of the MD simulation [55]. Many high-temperature simulation studies have been done to explore protein unfolding pathways, and comprehensive reviews exist on MD simulation protocols [56–59]. For example, MD simulations at 373 and 498 K of the engrailed homeodomain (En-HD), a three-helix bundle 61-residue protein, have been used to analyse a folding intermediate at the atomic level [60, 61].

4.1 Force Field in MD Simulations

The total potential energy of a system containing molecules being simulated in a solvent is usually depicted as a sum of intramolecular potentials (one for each molecule in the system) and intermolecular potentials. The intramolecular potentials usually involve a sum of covalent interactions describing how the energy varies with bond stretching, bond bending, and dihedral angle distortion. In contrast, the intermolecular potentials involve non-covalent interactions. Denaturation by temperature is a step in molecular simulation significantly impacted by the force field since it defines the energy of the bonds and angles that require denaturation.

A typical force field consists of covalent interactions (bond stretching, angle bending, torsional rotation) and non-covalent interactions (van der Waals forces, electrostatic forces, and hydrogen bonding).

$$U_{\text{total}} = E_{\text{stretching}} + E_{\text{bending}} + E_{\text{torsion}} + E_{\text{vdw}} + E_{\text{es}} + E_{\text{hb}}$$

where

$E_{\text{stretching}} = \frac{1}{2}\ \text{kB}(b - b_0)^2$ [where b is the new bond length and b_0 is the ideal bond length value, and kB is the stretching force constant].

$E_{\text{bending}} = \text{kA}(\theta - \theta_0)^2$ [where θ is the new bond angle and θ_0 is the ideal bond angle value, and kA is the bending force constant].

$E_{\text{torsion}} = \frac{V_{\text{n}}}{2}\ [1 + \cos(n\omega + \gamma)]$ [where V_{n} is the proportionality constant, ω is the angle, n is the time period, and γ is the cycle].

$E_{\text{vdw}} = 4\varepsilon\left[\left(\frac{\sigma_{\text{ij}}}{r_{\text{ij}}}\right)^{12} - 2\left(\frac{\sigma_{\text{ij}}}{r_{\text{ij}}}\right)^{6}\right]$ [where the equation models the Lennard-Jones potential with equilibrium distance σ_{ij}].

$E_{\text{es}} = \frac{1}{4\pi\varepsilon}\ \frac{q_{\text{i}}\ q_{\text{j}}}{r_{\text{ij}}}$ [Where the equation models the electrostatic potential between two charges particles q_{i} and q_{j}].

E_{hb} is the hydrogen bond energy obtained.

Thermodynamics studies suggest that in a system that contains a certain amount of energy and is in equilibrium, the energy is distributed in a simple manner. This distribution is calculated using the Maxwell–Boltzmann distribution. This implies that the probability of having the same velocity for every particle is minuscule in a complete system of particles. The velocities, on the contrary, follow a distribution assigned and proportional to their mass and the system's temperature. The speed distribution obeys the following relationship:

$$f(v) = 4\pi v^2 \left(\frac{m}{2\pi k_{\text{B}} T}\right)^{\frac{3}{2}} \text{e}^{\left(-\frac{mv^2}{2k_{\text{B}}T}\right)}$$

where m is the molecule's mass, k_{B} is the Boltzmann constant, v is the speed, and T is the absolute temperature. The temperature control methods can be divided into several categories:

4.2 Strong Coupling Methods

4.2.1 Velocity Rescaling

In such an experiment, the intention is to change the velocity at each step (or after a set number of steps) to obtain a particular desired temperature.

4.2.2 Velocity Reassignment

Randomized velocities are employed for all the velocities of the system, and reassigning occurs periodically to set the entire system to a particular desired temperature, as opposed to setting a particular velocity at a time (which in turn saves resources and time). Both velocity rescaling and reassignment do not generate an accurate canonical ensemble. The inconsistency arrives in the process since the kinetic energy does not fluctuate with the rescaling. For equilibrium dynamic studies, the methods mentioned above are not recommended; however, they are useful for system heating and cooling dynamic studies. Temperature reassignment is the better method because it avoids the undesired increment of already varied temperature spots. Although, both have their respective advantages and disadvantages.

4.3 Weak Coupling Methods

4.3.1 Berendsen Thermostat

At every simulation step, the Berendsen thermostat rescales the velocities of all particles to remove a predefined fraction of the difference from the predefined temperature [62]. The velocities are scaled at each step, such that the rate of change of temperature is proportional to the difference in temperature:

$$\frac{dT(t)}{dt} = \frac{1}{\tau} \left(T_0 - T\ (t)\right)$$

where τ is the coupling parameter that determines how tightly the bath and the system are coupled together, and T is the temperature.

Theoretically, the Berendsen method works similarly to how a simulation with a heat bath kept at a constant temperature might work. Since the strength of the coupling defines the rate of temperature equilibration with this thermostat, the Berendsen thermostat is a powerful algorithm. It allows the system to relax and obtain better results. For instance, this algorithm produces relatively good results when initiating a simulation study after energy minimization.

However, a disadvantage of the Berendsen method is that it cannot be mapped onto a specific thermodynamic ensemble. Statistical analysis suggests that the Berendsen thermostat produces a lower variance of energy distribution when compared to a true canonical ensemble. This is because it samples kinetic energies disproportionally and closer to T_0 than would be observed in the true Maxwell–Boltzmann distribution [63, 64]. Hence, this method is usually avoided for simulations that involve subsequent production. The weak coupling of this method can be comprehended as heat flows between the simulated systems. With the time constant τ_T, larger values suggest slower equilibration and weak coupling, while smaller values of τ_T mean tight coupling and relatively faster equilibration. The time constant for heat bath coupling for the system is measured in picoseconds (Table 1).

Table 1 Various temperature coupling functions in different software

S. no.	Thermostat/MD package	Functions		
		GROMACS	NAMD	AMBER
1	Velocity rescaling		reascaleFreq (steps)	
2	Velocity reassignment		reassignFreq (steps)	
3	Andersen	tcoupl = andersen		
4	Massive-Andersen	tcoupl = andersen-massive		Ntt = 2
5	Lowe-Andersen		loweAndersen on	
6	Berendsen	tcoupl = berendsen	tCouple on	Ntt = 1
7	Langevin		Langevin on	Ntt = 3
8	Bussi	tcoupl = V-rescale	stochRescale on	
9	Nosé–Hoover	tcoupl = nose-hoover		
10	Nosé–Hoover-chains	nh-chain-length (default 10)		

4.4 Stochastic Methods

Stochastic methods generally hint at the usage of randomly assigned parameters. Here, random velocity assignment takes place for a subset of atoms based on Maxwell–Boltzmann distributions for the target temperature. This stochasticity slows down the system's kinetics through the interference of motion.

4.4.1 Andersen Thermostat

As a typical stochastic method, the Andersen method controls the temperature of a system by assigning a subset of atoms with new velocities generated from the Maxwell–Boltzmann distribution for the target temperature randomly. The probability for a given particle to have its velocity reassigned at each step can be calculated as a fraction of time step $\mathit{\Delta} t$ and time constant, τ_{T}, $\frac{\mathit{\Delta} t}{\tau_{\mathrm{T}}}$. This implies that, on average, every atom experiences a random collision with a virtual particle every time step Δt [65]. A method derived from the Andersen thermostat algorithm, termed "massive Andersen thermostat", randomizes the velocity of every atom at every Δt, increasing the computation time and cost [63].

The Andersen method has been observed to sample canonical ensemble correctly; however, momentum is not conserved with this method. This is a feature of the Lowe-Andersen thermostat. Because of velocity randomization, some correlated motions are impaired, which slows down the system's kinetics. Hence, this method is not recommended while studying a system's kinetics or diffusion properties. This further applies to all stochastic methods. It should be noted that the steps involved in the randomization of velocities to distribution are important parameters to comprehend the speed of the collision and the rate at which the particles collide. An abnormal increase in collision rate, which means shorter steps in randomization, usually slows down the speed at which the molecules can confer to a better configuration, while an abnormal decrease in collision rate, which implies longer steps between randomization, means the canonical distribution of energies will be sampled slowly.

4.4.2 Lowe-Andersen Thermostat

A variant of the Andersen thermostat that conserves momentum is the Lowe-Andersen thermostat. This algorithm usually does not agitate the system dynamics more than the original Andersen method but enables the alleviation of the suppressed diffusion in the system [66].

4.4.3 Bussi's Stochastic Velocity Rescaling Thermostat

Bussi's stochastic velocity rescaling is an extension of the Berendsen thermostat method. This has been corrected for sampling the canonical distribution. A certain chosen random factor is used to rescale the respective velocities. While retaining the advantages of the Berendsen method, it produces a correct velocity distribution for the canonical ensemble. The ability to avoid oscillations that are observed in similar thermostats is done by converging the temperature deviations from the target via a first-order exponential decay. For most temperature-controlled MD simulations, this thermostat is an excellent choice [67].

4.4.4 Langevin Thermostat

Langevin equation is an equation of motion for a system experiencing a fluctuating force. The typical system where this equation can be implemented is a particle experiencing Brownian motion. The Langevin equation for a Brownian particle in a one-dimensional (1D) fluid bath is

$$mv^{\cdot}(t) + \zeta v(t) = f(t)$$

where m is the mass of the Brownian particle, $v(t) = x^{\cdot}(t)$ is the velocity of the Brownian particle, ζ is a coefficient describing friction between the particle and the bath, and $f(t)$ is a random force. Though it is random, we can make a couple of useful assumptions about the force, $f(t)$:

1. The probability of randomness in the force being calculated for the system is equally likely to cancel out since it can be equally likely to push in one direction as it is in the other, which implies:

$$\langle f(t) \rangle_f = 0$$

2. The random force of the system has no direct correlation with time but is associated with a strength factor g, which in turn does not change with time:

$$\langle f(t1) f(t2) \rangle_f = g\delta(t1 - t2)$$

It should be noted that the damping coefficient governs the friction in the system. A sudden increase in the coefficient value can result in the atoms experiencing an increased unnatural resistance and friction. On the contrary, if the coefficient value is decreased, the system has a high probability of fluctuating, and the desired temperature might not be achieved.

4.5 Extended System Dynamics

4.5.1 Nosé–Hoover Thermostat

The extended system method was introduced initially by Nose and subsequently developed by Hoover [59]. In this method, the heat bath is regarded as an integral part of the system, and a variable is introduced to the equations of motion, which are associated with the heat bath mass. The most prominent feature of this algorithm is that it enables the control of temperature without using stochastic functions that assign random numbers. Therefore, the correlated motions are not impaired, and this method better describes kinetics and diffusion properties. Adding the heat bath mass variable leads to heat dissipation since a second-order equation describes a time evolution. The exchange of heat in the system occurs in an oscillatory fashion, which implies that “heat bath mass” can be thought of as directly proportional to the frequency of temperature fluctuations that will occur because of the oscillations [68, 69]. The most prominent disadvantage of this algorithm is that it significantly affects the system’s distribution. The time constant parameter in this thermostat controls the period of temperature fluctuations at equilibrium.

4.5.2 Nosé–Hoover-Chains

This method is a modification of the Nosé–Hoover thermostat. It includes a chain of variables rather than using a single thermostat variable. [70]. Stochasticity is varied in the main method of the Nosé–Hoover thermostat, but for small or stiff systems, the algorithm cannot guarantee complete ergodicity. In contrast, chaining variables behave better for small or stiff cases, leading to comprehensive ergodicity and ensuring the entire system space is used. However, an infinite chain is required to adequately correct these issues, which increases the computational cost.

4.6 Analysis of MD Simulation Trajectories

Two essential questions that need to be asked before running any simulation are, first, what simulation process does one wants to use? And second, what result does one wants to obtain from that simulation? One should always have ideas about what type of data one wants to collect in their system. The first step towards data analysis or interpretation is gmx trjconv, a pre-processing tool to eliminate coordinates, adjust for any periodicity, or manually change the trajectory (time, unit, frame frequency, etc.). The protein might defuse through the unit cell and appear broken or jump across the other side of the box. Thus, this pre-processing tool is used to clear the periodic boundary condition (PBC), and the protein is placed centrally in the solvent box. After this process, the corrected trajectory is used to perform all analyses.

The trajectory file used for the analysis is generated in the MD run along with the . tpr file, which has all starting structure information, molecular topology, and the simulation parameters of the protein or compound. A .xvg file is generated, which is used to display the results (such as RMSD, RMSF, etc.) in graphical forms through the GRACE program in Linux/UNIX or GNUplot in Windows. The .xvg files are plain text files containing tabular data separated by tabulators and two types of comments with data labels. The MD simulation outcomes can be illustrated in several ways to provide insights into the structural changes that occurred by increasing temperatures or denaturing agents over a certain period. After simulation, trajectories are analysed to establish the role of temperature/denaturant to get atomistic insights.

4.7 Root Mean Square Deviation

Root mean square deviation (RMSD) is a standard measure of structural distance between atom coordinates. It is an average square root of all the C-alpha atom's distances. It is a numerical representation of the distance between two structures. It is used to study how a structure or a part of a structure behaves over time from the initial structure under certain conditions [71–75]. The variation of RMSD values over a period provides information on the structural changes in a protein [76]. The spatially equivalent structure shows a slight difference between the structures where the deviation is minimal, and greater RMSD values are shown by the more distantly related structures [77].

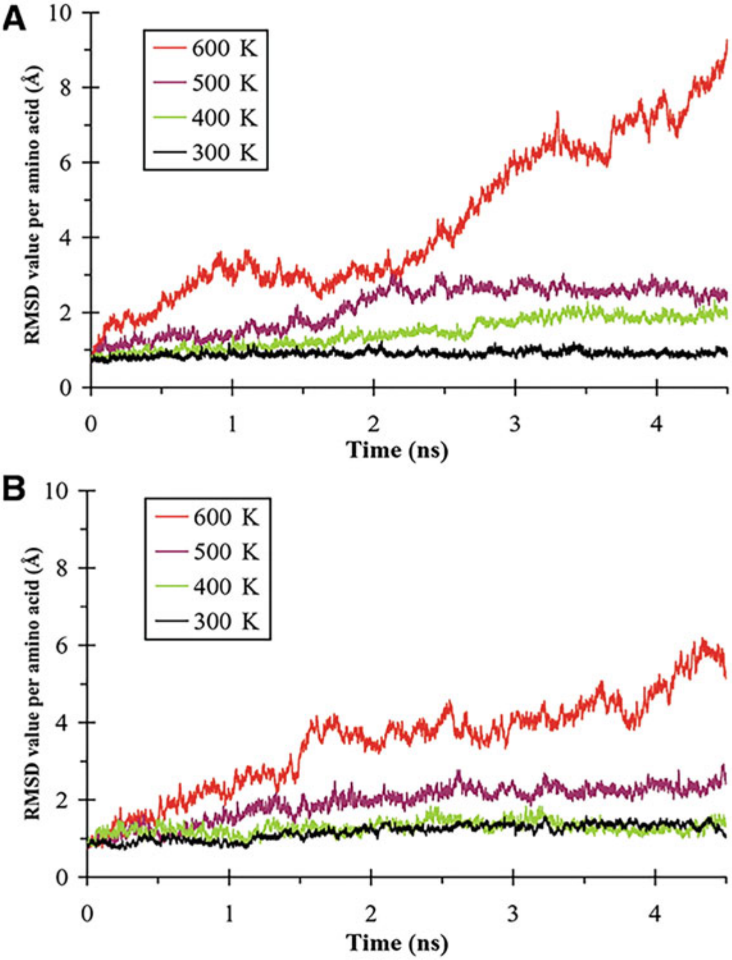

Fig. 2 RMSD of C-alpha atoms of (**a**) TRX II and (**b**) DTX at different temperatures as a function of time. The figure is adapted with permission from [78]

The RMSD analysis under high temperatures can reflect the protein backbone's average movement throughout the entire protein structure, as shown in Fig. 2. When the temperature of the system is low, i.e. below 300 K, there is little to no increment in the RMSD values. Still, as the temperature increases from 300 to 343 to 373 K, a significant rise in RMSD is observed [79], representing structural changes in the protein. Although depending upon the types of protein, this variation in RMSD value can also be different as thermophilic proteins are stable under much higher temperatures than mesophilic temperatures. The graph obtained by calculating RMSD from the MD simulation trajectories can help understand the changes between two structures, typically plotted versus time [80]. The flat curve indicates that the structure has equilibrated. Figure 2 shows the sensitivity of *Trichoderma reesei* xylanase II (TRX II), a mesophilic protein, and *Dictyoglomus thermophilum* xylanase (DTX), a thermophilic protein, under various temperatures. Initially, denaturation was not visible at lower temperatures (300 and 400 K), although there were slight structural changes in both proteins. As the temperature was increased to 500 K, the unfolding of mesophilic protein was evident compared to thermophilic. As the temperature reached 600 K, complete protein denaturation was observed, particularly in the mesophilic protein [78].

The RMSD of protein structure throughout the simulation, both on its ligand and unbound form, can also be calculated. The equation given below is used to calculate the RMSD value:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_\mathrm{i} - x_0)^2 + (y_\mathrm{i} - y_0)^2 + (z_\mathrm{i} - z_0)^2}$$

where N is the number of atoms where (x_i, y_i, z_i) is the coordinate of the structure whose RMSD is calculated and (x_0, y_0, z_0) is the coordinate of reference structure.

4.8 Root Mean Square Fluctuation

The root mean square fluctuation (RMSF) measures a particle's average change or deviation from its initial position over time under a specific condition [81]. It analyses the part of a structure which is deviating or fluctuating from its mean structure (Fig. 3). The RMSD measures the average change in the structure, while RMSF measures the average change in the particular residue or how much a particular residue has changed over time during the MD simulation [83]. It calculates the fluctuation of C-alpha atoms in an amino acid residue in the protein compared to the average structure throughout the simulation. An increased residual RMSF value indicates instability in the protein backbone or flexibility [84].

Unlike RMSD, RMSF is typically plotted against residue number and can help understand which amino acid residues are in dynamic motions [76]. The plot

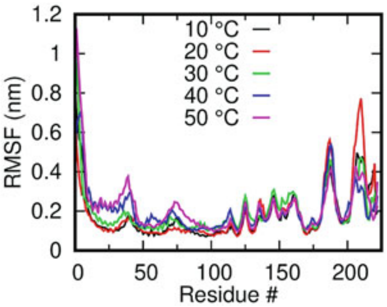

Fig. 3 Variation in the RMSF of M-protein of SARS-Cov 2 at different temperatures. The figure is adapted with permission from [82]

normally represents the residue that has gone through changes throughout the simulation cycle. It also helps to find the instability of the proteins under various conditions, including change in temperature, pH, etc., and assist in identifying local changes in the protein chain [85].

Figure 3 represents the effect of temperature on the RMSF of the M protein of SARS-Cov 2. Even though the overall structure appears stable, distinct peaks can be seen for residues 9–47 at 40 °C and 50 °C, the C-terminal loop around residues 180–190 and residues 203–220, which show higher flexibility between 20 °C and 40 °C [82]. This change in the flexibility of amino acid residues can cause protein instability and reduce or change the protein's function. The change in flexibility can also be used to study the inhibitor binding to the target. The RMSF values can be calculated using the equation:

$$\mathrm{RMSF} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}\left(x_{\mathrm{i}} - \bar{x}\right)^{2}}$$

Here t is the trajectory frame number, and $\bar{x}$ is the time-averaged position.

4.9 Hydrogen Bonding Analysis

Hydrogen bonding is a type of dipole–dipole interaction that forms between a hydrogen atom covalently bonded to an electronegative atom [86]. Hydrogen bonds in protein help stabilize its structure; for example, H-bond between the amide nitrogen and main chain stabilizes the secondary structure and is also linked to the compactness of the protein structure [87]. Hydrogen bonds are crucial in protein in maintaining the functional 3D conformation and proper binding with the substrate or ligand. The variation in the length of a particular intermolecular H-bond within a protein structure or intermolecular H-bond between two interacting proteins or H-bond involved in protein–ligand interaction. H-bonds can be measured throughout the MD simulation run for a specific time scale and temperature [88].

Hydrogen bonds are critical for the biological system as they stabilise the protein structure [89]. The trajectories from the MD simulation can be used to study the effects of temperature on the structure as temperature destabilizes and denatures the protein by disrupting hydrogen bonds [90–95]. Along with the number of H-bonds, the bond length also helps determine the overall strength of H-bonds [96]. The temperature increase not only disrupts the H-bonds but also increases the distance between the molecular chains, which in the process, does not allow the formation of new H-bonds.

Figure 4a indicates the impact of temperature on the H-bonds. As the temperature increases from 280 K, the number of H-bonds decreases in all systems. Fig. 4b shows the probability of the number of H-bonds at three different temperatures

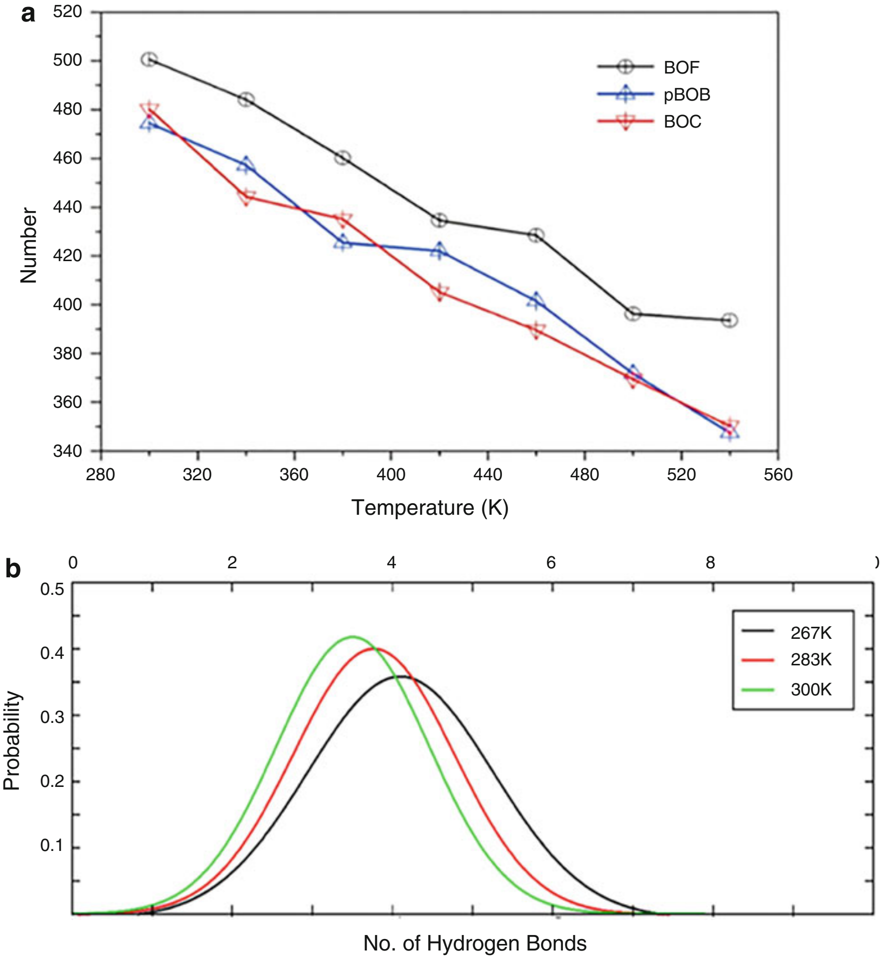

Fig. 4 (**a**) Number of H-bonds in different cross-linked systems under different temperatures ranging from 280 to 500 K. The figure was taken from [96]. (**b**) Distribution of H-bonds at 267, 283, and 300 K trajectories of short alpha-helix 2I9M peptides. The figure is adapted with permission from [97]

(267, 283, and 300 K) of a short alpha-helix 2I9M [98]. There is a probability of forming more than four H-bonds at 267 K, while less than four H-bonds as the temperature increases from 283 and 300 K.

4.10 Dihedral Angle Analysis

Determining dihedral angles (phi & psi angles) for all residues in a protein is important to analyse the mechanical importance of a particular residue in maintaining local conformations. Using the MD simulation trajectory of a protein, variation in the residual dihedral angle can be measured throughout the entire MD run or a specific time scale. The protein chain undergoes helix-coil transition with increasing temperature and through different transition states. During temperature-induced transition, there is a synchronous change in the dihedral angles along the helical chain, leading to the simultaneous breaking of helices [99].

4.11 Radius of Gyration

The radius of gyration (R_g) is used to measure the stability and structural flexibility of the protein in a biological environment [100]. It is one of the fundamental indicators of the overall size of a protein. It helps evaluate and verify protein structure compaction during the MD simulation [101]. A small R_g value indicates a rigid structure. The glass transition temperature is defined by the temperature at which the amorphous structure of the polymer changes from hard to soft. As the temperature increases, the stability of the structure decreases and the R_g value increases, implying the increase in flexibility of the structure [102]. Figure 5 shows plots of R_g versus temperature, showing that the increase in the R_g occurs with the temperature increase in all plots. Even though they all have different glass transition temperatures, their flexibility keeps increasing as the temperature increases [103].

The R_g is calculated using the following formula:

$$R_g = \sqrt{\frac{1}{M}\sum_{i=1}^{n} m_i (r_i - R)^2}$$

where M is the total mass of the protein and R is the centre of mass of the protein.

4.12 Protein Solvent Accessible Surface Area

Protein solvent accessible surface area (SASA) is considered one of the fundamental elements in the stability and folding of proteins [104, 105]. Various interactions between molecules and solvents depend on the secondary structural changes, and changes in the secondary structure can cause changes in functional properties. SASA specifies the area on the surface of biomolecules that can be used for interaction with

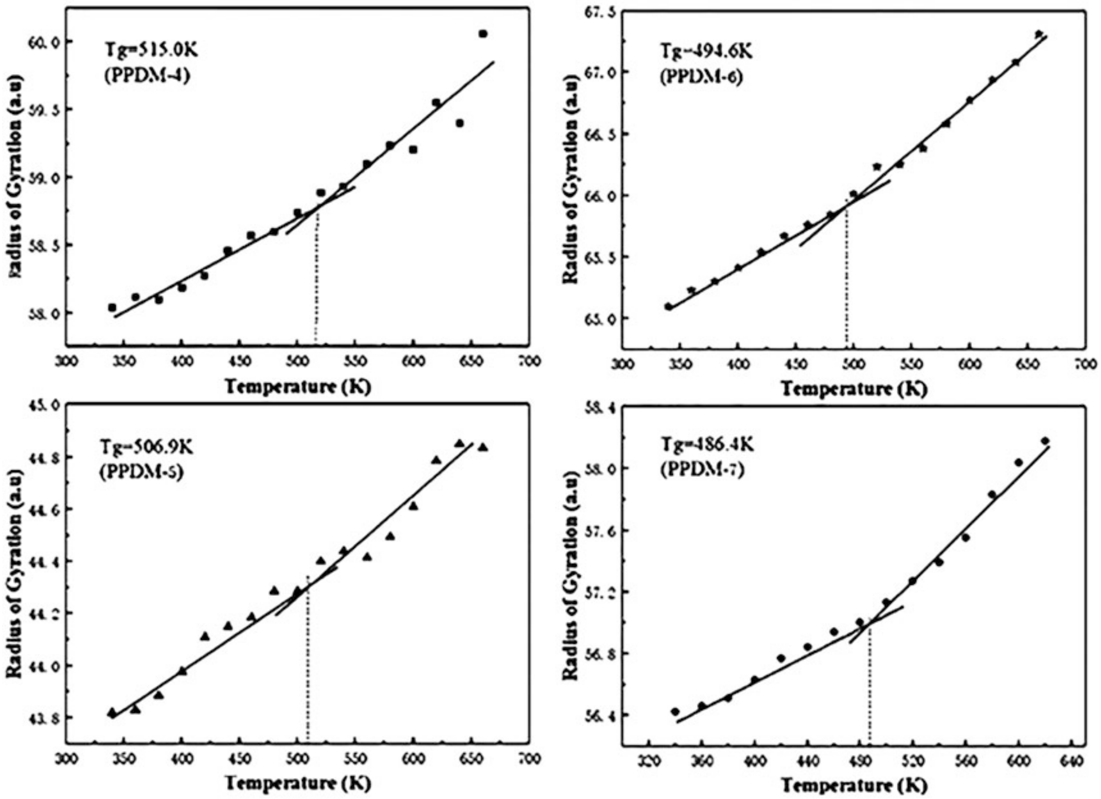

Fig. 5 Radius of gyration (R_g) versus temperature plots generated from the MD simulation trajectories of PPDM. (PPDM = p-phenylenediamine-alt-2,6-diformylmultiphenyl). The figure is adapted with permission from [103]

solvents and other molecules. A study of changes in SASA can help us understand the folding and unfolding of complexes [101]. Lower SASA means a more compact structure [106]. When exposed to thermal stress, protein tends to undergo structural changes, and this conformational change exposes the hydrophobic residues of water and solvents [107]. The SASA can be calculated using the following:

$$\mathrm{SASA} = A = \sum \left(\frac{R}{\sqrt{R^2 - Z_{\mathrm{i}}^2}} \times D \times L \right)$$

where A is surface area, R is the atom's radius, and L is the length of an arc drawn on a given section of i from the centre of the sphere [108].

4.13 Principal Component Analysis

Principal component analysis (PCA) provides information about the essential protein backbone motions in a complex system along the MD simulation trajectory [109] and helps understand the protein folding, loop movement, etc. It also helps us understand different conformations of a protein that is generated during the MD simulation. Interpreting these trajectories helps understand how a protein undergoes dynamics and performs a specific function in the biological environment [110]. Protein motion is mostly described by the eigenvalues, eigenvectors, and covariance matrix [111]. The covariance matrix can be calculated using the following:

$$C = < (q_{\mathrm{i}} - < q_{\mathrm{i}} >) \left(q_{\mathrm{j}} - < q_{\mathrm{j}} > \right) > (i, j = 1, 2, \ldots\ldots, 3N)$$

where i and j signify the position of the C-alpha atom and N signifies the number of C-alpha atoms.

Figure 6a shows the SASA plot of urea-induced denaturation studies performed and analysed to derive thermodynamic parameters associated with the stability of SphK1. Similarly, SASA can be studied at different temperatures; however, since there is an increase in temperature, there is a constant increase in the surface area of the protein, so the chances of interaction with other molecules also increase.

Sets of correlated observations (such as the movement of all atoms of a protein) are converted by PCA into linearly independent or correlated principal components. Mathematically, a new coordinate system is generated by transforming the data, in which each coordinate represents a different degree of variance. Figure 6b shows the dynamics of SphK1 that were computed from the backbone using the gmx cover module. The essential dynamics recognize significant average atomic motions of a protein molecule, showing the structures underlying the atomic fluctuations.

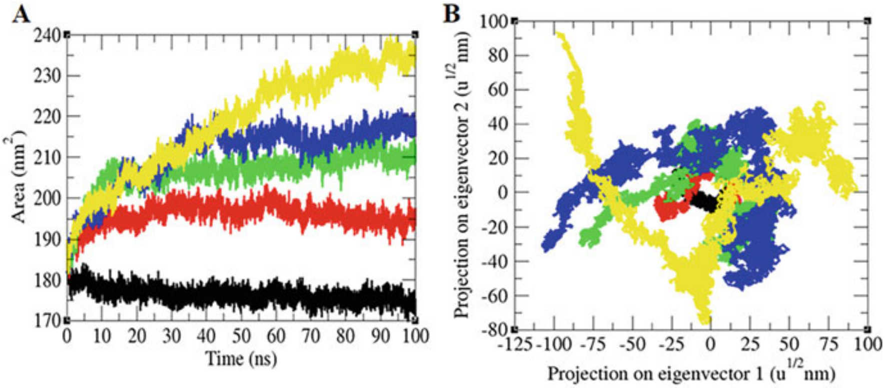

Fig. 6 (**a**) SASA of SphK1 at different urea concentrations. (**b**) The 2D projections of trajectories on eigenvectors showed different projections of SphK1 at different urea concentrations. The figure is adapted with permission from [26]

4.14 Free Energy Landscape Analysis

The free energy landscape (FEL) represents the dynamic and equilibrium properties of a protein. Understanding protein unfolding and folding can also be made using FEL [112]. FEL can be plotted in both 2D and 3D and represent the stability and conformational changes of a protein (ligand-bound/unbound) in terms of Gibbs free energy. Two data present protein stability and conformational changes: RMSD and R_g analysis from the trajectory of MD simulation of the protein system. They are correlated with the Gibbs free energy since the change in temperature affects the RMSD and R_g values. An increase in temperature causes an increase in RMSD and R_g values, which signifies a decrease in structural stability and a reduction in compactness and rigidity of the protein structure, thus implying that an increase in temperature will cause variation in the free energy landscape.

4.15 Dynamic Cross-Correlation Matrix

A correlation matrix represents the correlation motion of all the C-alpha atoms in a protein structure. The physical motion of atoms is studied using a dynamic cross-correlation matrix (DCCM). The simulation trajectories help investigate the dynamic changes of the system over time. By analysing them, one can study the degree to which the atoms move together. This is known as the dynamic correlation between all the atoms of the molecule [113]. DCCM can be calculated using the following:

$$C_{\mathrm{ij}} = \frac{\left(\Delta r_{\mathrm{i}} \cdot \Delta r_{\mathrm{j}}\right)}{\left(\left\langle \Delta r_{\mathrm{i}}^2 \right\rangle \left\langle \Delta r_{\mathrm{j}}^2 \right\rangle\right)^{\frac{1}{2}}}$$

Here Δr represents the average point movement of the atom.

4.16 Loss of Secondary Structures in High Temperatures

H-bonding majorly governs a protein's secondary structure; thus, it is responsible for its structural stability and is an essential indicator of the folding/unfolding of a protein. A rise in temperature causes loss in the native secondary structures. With the initial increase in temperature, proteins tend to retain their secondary structures due to increased flexibility and, thus, an increase in H-bond formation. However, as the temperature rises further, significant structural alterations occur, and the structural integrity is lost [114].

Figure 7 shows the secondary structure evolution at different temperatures for the enzyme Barnase. The data indicate that alpha-helices, beta-sheets, and loops are stable throughout the trajectories; however, as the temperature increases to 500 and 550 K, fluctuations are more pronounced, and the protein unfolds rapidly [115].

Figure 8 shows the snapshots of the thermal unfolding of Barnase at various temperatures and different simulation times. An immediate state is observed during unfolding at 600 K. The destruction of native secondary structures occurs instantly. Highly coiled protein can be seen in the early stages of simulation, and only a few secondary structures can be observed at the end of 1 ns.

4.17 Analysing the Unfolding of Human Prion Protein Under Low pH and High-Temperature Conditions

MD simulation studies have illustrated the unfolding kinetics of the human prion protein (HuPrP). Prion diseases are fatal neurodegenerative disorders caused by pathogenic prions in cattle and humans. The human prion protein domain MD simulations were performed for 10 ns at high temperatures (298 and 350 K) and low pH. The data suggested that heat and pH-induced unfolding of HuPrP follow different pathways. At neutral pH, the native structure was observed to be stable, while under acidic (weakly acidic and strongly acidic) environments, the structure started to unfold where only the core of the prion protein remained intact, which harboured disulphide bonds. The loss of helices and secondary structure changes were observed in both low-pH and high-temperature conditions [116].

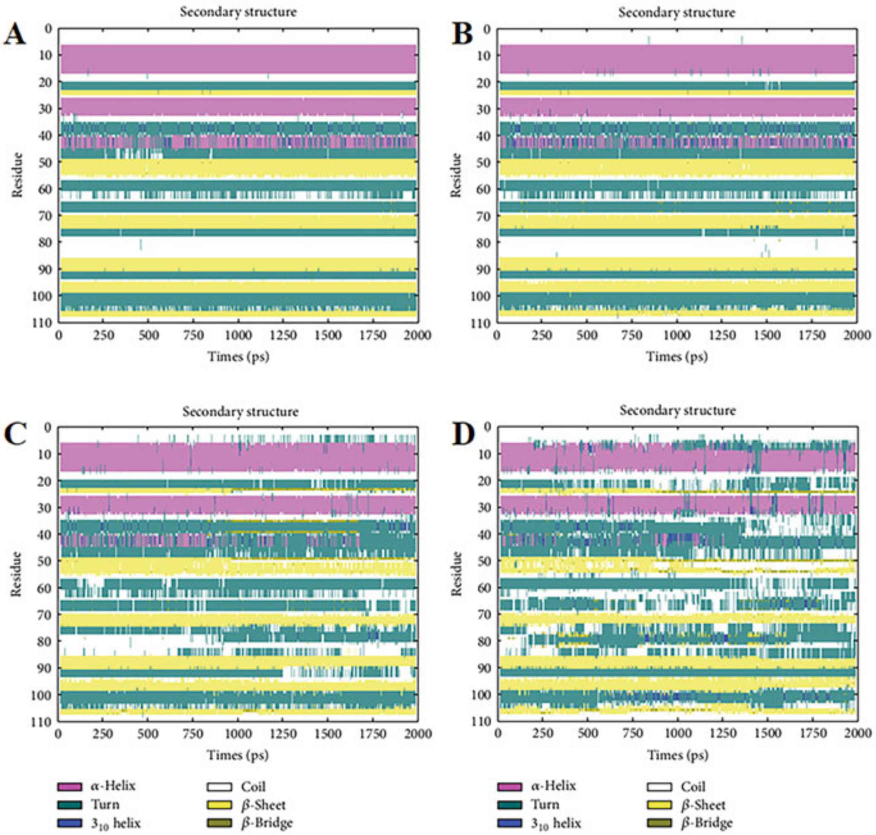

Fig. 7 Evolution of secondary structure in the unfolding trajectories at different temperatures for Barnase (**a**) 300 K, (**b**) 400 K, (**c**) 500 K, (**d**) 550 K. The figure is adapted with permission from [115]

5 Applications of MD Simulation in Understanding Biological Problems

Protein folding/unfolding has been investigated by both experiments and simulations [117]. However, advancements in computational techniques have made it easier to study the overall dynamics of proteins [118]. Due to the availability of high-end computer hardware, software, and algorithms, studying the processes involved in protein folding/unfolding has become feasible using MD simulation [119]. Let us understand with an example. Imagine an alien lands on Earth, hears about something called a "bicycle", and wants to know how it works, how to ride it, and how to fix it when it breaks; figuring this out would be challenging, given just a bicycle picture. Watching a movie about someone riding a bicycle would help.

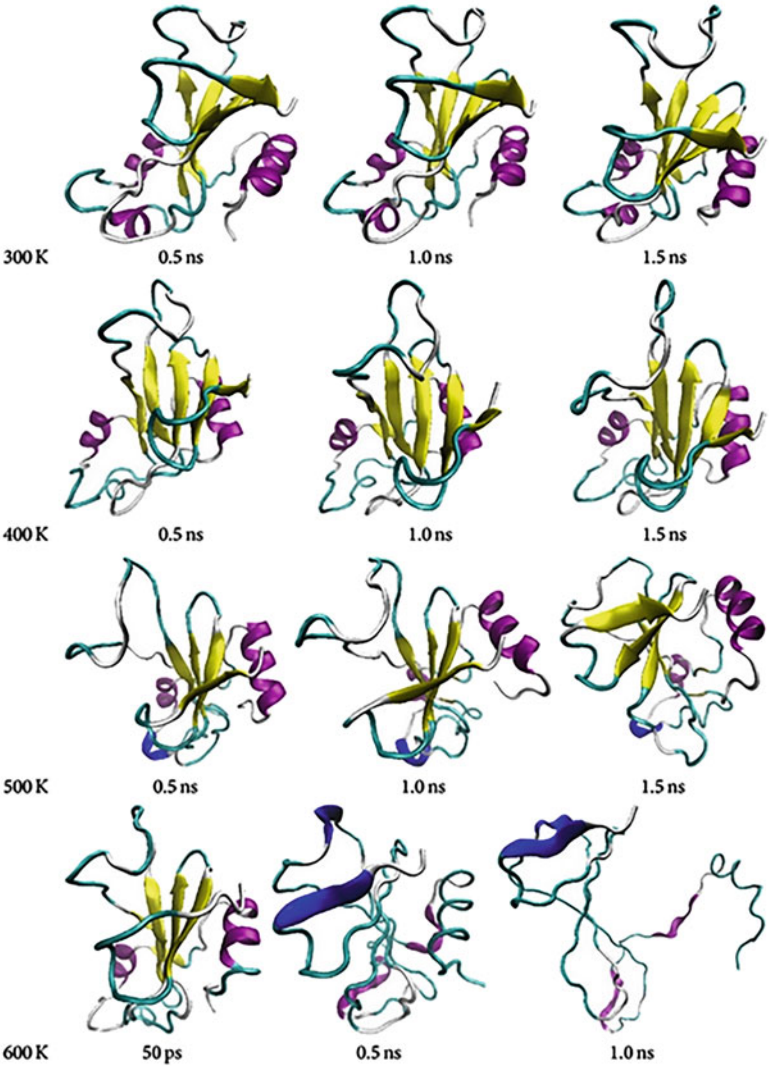

Fig. 8 Snapshots of the thermal unfolding of Barnase. The thermal unfolding of Barnase at various temperatures and different simulation times is shown. The figure is adapted with permission from [115]

Similarly, studying how a protein unfolds at high temperatures would be helpful when we capture the complete picture of events involved in protein unfolding.

There are many limitations with the experimental approaches to analysing protein folding events. They cannot provide an enhanced high-resolution description of the temporal process and the conformational changes with minute details. To overcome such limitations, researchers opt for computational techniques. The role of MD simulation has extended dramatically in structural biology in recent years [120]. All-atom simulation provides an insight into the atomic resolution of protein dynamic behaviour and non-equilibrium phenomenon like protein folding and unfolding. When these events are studied along with the wet-lab experiment, it is observed that simulation could provide intensified data on the system under study [121, 122]. Computational methods enable us to simulate the protein as a function of temperature or any other denaturant and generate massive data, allowing us to investigate and visualize the process of folding/unfolding from nanosecond to microsecond time scale [122]. MD simulation can also provide desirable information regarding the kinetics and thermodynamics of proteins [123].

6 Conclusion and Future Prospects

MD simulations are frequently used to obtain the atomic-level understanding of the protein folding process, which is otherwise challenging to get experimentally. However, to simulate the unfolding of a protein in an acceptable period of computer time, a substantial perturbation is necessary, which in turn adds inefficiencies to the data that cannot be avoided. More research into protein unfolding will be done as computing power increases. According to studies, simulations are significantly influenced by the kind and scale of perturbation employed to drive unfolding. A gap persists to be filled for a profound evaluation between simulations and experiments. The forthcoming investigational study of protein unfolding near the boiling point would bring critical insights into the validity of predicted unfolding rates.

Acknowledgement M.I.H. thanks the Council of Scientific and Industrial Research for financial support [Project No. 27(0368)/20/EMR-II].

References

1. J.C. Whisstock, A.M. Lesk, Prediction of protein function from protein sequence and structure. Q. Rev. Biophys. **36**(3), 307–340 (2003)
2. D.B. Singh, T. Tripathi, *Frontiers in Protein Structure, Function, and Dynamics* (Springer Nature, Singapore, 2020)
3. P. Davis, S. Williams, Protein modification by thermal processing. Allergy **53**, 102–105 (1998)
4. R. Jaenicke, Protein folding: local structures, domains, subunits, and assemblies. Biochemistry **30**(13), 3147–3161 (1991)
5. H. Wu, Studies on denaturation of proteins XIII. A theory of denaturation. Adv. Protein Chem. **46**, 6–26 (1995)
6. V.N. Uversky, N.V. Narizhneva, S.O. Kirschstein, S. Winter, G. Löber, Conformational transitions provoked by organic solvents in β-lactoglobulin: can a molten globule like intermediate be induced by the decrease in dielectric constant? Fold. Des. **2**(3), 163–172 (1997)

7. C.M. Dobson, M. Karplus, The fundamentals of protein folding: bringing together theory and experiment. Curr. Opin. Struct. Biol. **9**(1), 92–101 (1999)
8. D. Thirumalai, E.P. O'Brien, G. Morrison, C. Hyeon, Theoretical perspectives on protein folding. Annu. Rev. Biophys. **39**, 159–183 (2010)
9. P. Saudagar, T. Tripathi, *Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics*, 1st edn. (Academic Press, San Diego, 2023)
10. T. Tripathi, V.K. Dubey, *Advances in Protein Molecular and Structural Biology Methods*, 1st edn. (Academic Press, Cambridge, MA, 2022)
11. S.E. Radford, Protein folding: progress made and promises ahead. Trends Biochem. Sci. **25**(12), 611–618 (2000)
12. J. Yon, Protein folding: a perspective for biology, medicine and biotechnology. Braz. J. Med. Biol. Res. **34**, 419–435 (2001)
13. A. Gershenson, S. Gosavi, P. Faccioli, P.L. Wintrode, Successes and challenges in simulating the folding of large proteins. J. Biol. Chem. **295**(1), 15–33 (2020)
14. R. Lazim, D. Suh, S. Choi, Advances in molecular dynamics simulations and enhanced sampling methods for the study of protein systems. Int. J. Mol. Sci. **21**(17), 6339 (2020)
15. D. Bhowmik, S. Gao, M.T. Young, A. Ramanathan, Deep clustering of protein folding simulations. BMC Bioinformatics **19**(18), 47–58 (2018)
16. A.R. Fersht, V. Daggett, Protein folding and unfolding at atomic resolution. Cell **108**(4), 573–582 (2002)
17. V. Daggett, Protein folding-simulation. Chem. Rev. **106**(5), 1898–1916 (2006)
18. A. Prakash, D. Idrees, M.A. Haque, A. Islam, F. Ahmad, M.I. Hassan, GdmCl-induced unfolding studies of human carbonic anhydrase IX: a combined spectroscopic and MD simulation approach. J. Biomol. Struct. Dyn. **35**(6), 1295–1306 (2017)
19. H. Naz, M. Shahbaaz, M.A. Haque, K. Bisetty, A. Islam, F. Ahmad, M.I. Hassan, Urea-induced denaturation of human calcium/calmodulin-dependent protein kinase IV: a combined spectroscopic and MD simulation studies. J. Biomol. Struct. Dyn. **35**(3), 463–475 (2017)
20. D. Idrees, S. Rahman, M. Shahbaaz, M.A. Haque, A. Islam, F. Ahmad, M.I. Hassan, Estimation of thermodynamic stability of human carbonic anhydrase IX from urea-induced denaturation and MD simulation studies. Int. J. Biol. Macromol. **105**, 183–189 (2017)
21. D. Idrees, A. Prakash, M.A. Haque, A. Islam, F. Ahmad, M.I. Hassan, Spectroscopic and MD simulation studies on unfolding processes of mitochondrial carbonic anhydrase VA induced by urea. J. Biomol. Struct. Dyn. **34**(9), 1987–1997 (2016)
22. A.A.T. Naqvi, T. Mohammad, G.M. Hasan, M.I. Hassan, Advancements in docking and molecular dynamics simulations towards ligand-receptor interactions and structure-function relationships. Curr. Top. Med. Chem. **18**(20), 1755–1768 (2018)
23. A.A.T. Naqvi, U. Kiran, M.Z. Abdin, M.I. Hassan, Bioinformatic tools to understand structure and function of plant proteins, in *Transgenic Technology Based Value Addition in Plant Biotechnology*, (Academic Press, San Diego, 2020), pp. 69–93
24. R. Day, V. Daggett, All-atom simulations of protein folding and unfolding. Adv. Protein Chem. **66**, 373–403 (2003)
25. R. Day, V. Daggett, Ensemble versus single-molecule protein unfolding. Proc. Natl. Acad. Sci. **102**(38), 13445–13450 (2005)
26. F.I. Khan, P. Gupta, S. Roy, N. Azum, K.A. Alamry, A.M. Asiri, D. Lai, M.I. Hassan, Mechanistic insights into the urea-induced denaturation of human sphingosine kinase 1. Int. J. Biol. Macromol. **161**, 1496–1505 (2020)
27. F.I. Khan, K. Bisetty, S. Singh, K. Permaul, M.I. Hassan, Chitinase from thermomyces lanuginosus SSBP and its biotechnological applications. Extremophiles **19**(6), 1055–1066 (2015)
28. F.I. Khan, K. Bisetty, K.R. Gu, S. Singh, K. Permaul, M.I. Hassan, D.Q. Wei, Molecular dynamics simulation of chitinase I from thermomyces lanuginosus SSBP to ensure optimal activity. Mol. Simul. **43**(7), 480–490 (2017)

29. F.I. Khan, S. Ali, W. Chen, F. Anjum, A. Shafie, M.I. Hassan, D. Lai, High-resolution MD simulation studies to get mechanistic insights into the urea-induced denaturation of human sphingosine kinase 1. Curr. Top. Med. Chem. **21**(31), 2839–2850 (2021)
30. P.G. Wolynes, Evolution, energy landscapes and the paradoxes of protein folding. Biochimie **119**, 218–230 (2015)
31. X. Periole, A.E. Mark, Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. J. Chem. Phys. **126**(1), 01B601 (2007)
32. H. Lei, Y. Duan, Improved sampling methods for molecular simulation. Curr. Opin. Struct. Biol. **17**(2), 187–191 (2007)
33. K. Ostermeir, M. Zacharias, Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins. Biochim. Biophys. Acta **1834**(5), 847–853 (2013)
34. Y.M. Rhee, V.S. Pande, Multiplexed-replica exchange molecular dynamics method for protein folding simulation. Biophys. J. **84**(2), 775–786 (2003)
35. R. Zhou, Replica exchange molecular dynamics method for protein folding simulation, in *Protein Folding Protocols*, (Springer, Cham, 2007), pp. 205–223
36. R. Shukla, T. Tripathi, Molecular dynamics simulation of protein and protein-ligand complexes, in *Computer-Aided Drug Design*, ed. by D.B. Singh, (Springer Nature, Singapore, 2020), pp. 133–161
37. R. Shukla, T. Tripathi, Molecular dynamics simulation in drug discovery: opportunities and challenges, in *Innovations and Implementations of Drug Discovery Strategies in Rational Drug Design*, ed. by S.K. Singh, (Springer Nature, Singapore, 2021), pp. 295–316
38. S. Robic, Mathematics, thermodynamics, and modeling to address ten common misconceptions about protein structure, folding, and stability. CBE Life Sci. Educ. **9**(3), 189–195 (2010)
39. K. Anwer, R. Sonani, D. Madamwar, P. Singh, F. Khan, K. Bisetty, F. Ahmad, M.I. Hassan, Role of N-terminal residues on folding and stability of C-phycoerythrin: simulation and urea-induced denaturation studies. J. Biomol. Struct. Dyn. **33**(1), 121–133 (2015)
40. K. Anwer, S. Rahman, R.R. Sonani, F.I. Khan, A. Islam, D. Madamwar, F. Ahmad, M.I. Hassan, Probing pH sensitivity of αC-phycoerythrin and its natural truncant: a comparative study. Int. J. Biol. Macromol. **86**, 18–27 (2016)
41. K. Anwer, A. Parmar, S. Rahman, A. Kaushal, D. Madamwar, A. Islam, M.I. Hassan, F. Ahmad, Folding and stability studies on C-PE and its natural N-terminal truncant. Arch. Biochem. Biophys. **545**, 9–21 (2014)
42. G.N. Somero, Proteins and temperature. Annu. Rev. Physiol. **57**(1), 43–68 (1995)
43. V.Y. Alexandrov, Conformational flexibility of proteins, their resistance to proteinases and temperature conditions of life. Biosystems **3**(1), 9–19 (1969)
44. T. Tripathi, Calculation of thermodynamic parameters of protein unfolding using far-ultraviolet circular dichroism. J. Protein. Proteomics **4**(2), 85–91 (2013)
45. R. Jaenicke, Protein structure and function at low temperatures. Philos. Trans. R. Soc. Lond. B Biol. Sci. **326**(1237), 535–553 (1990)
46. S. Bondos, K. Matthews, *Protein Folding* (McGraw-Hill, New York, 2021)
47. H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. **6**(3), 197–208 (2005)
48. A.G. Rocco, L. Mollica, P. Ricchiuto, A.M. Baptista, E. Gianazza, I. Eberini, Characterization of the protein unfolding processes induced by urea and temperature. Biophys. J. **94**(6), 2241–2251 (2008)
49. A. Arsiccio, J.E. Shea, Pressure unfolding of proteins: new insights into the role of bound water. J. Phys. Chem. B **125**(30), 8431–8442 (2021)
50. L.J. Lapidus, Protein unfolding mechanisms and their effects on folding experiments. F1000Res **6**, 1723 (2017)
51. R. Day, B.J. Bennion, S. Ham, V. Daggett, Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. J. Mol. Biol. **322**(1), 189–203 (2002)

52. A. Caflisch, M. Karplus, Molecular dynamics simulation of protein denaturation: solvation of the hydrophobic cores and secondary structure of barnase. Proc. Natl. Acad. Sci. **91**(5), 1746–1750 (1994)
53. J.A. McCammon, B.R. Gelin, M. Karplus, Dynamics of folded proteins. Nature **267**(5612), 585–590 (1977)
54. M. Pechlaner, W.F. van Gunsteren, N. Hansen, L.J. Smith, Molecular dynamics simulation or structure refinement of proteins: are solvent molecules required? A case study using hen lysozyme. Eur. Biophys. J. **51**(3), 265–282 (2022)
55. K. Lindorff-Larsen, P. Maragakis, S. Piana, M.P. Eastwood, R.O. Dror, D.E. Shaw, Systematic validation of protein force fields against experimental data. PLoS One **7**(2), e32131 (2012)
56. V. Daggett, A. Fersht, The present view of the mechanism of protein folding. Nat. Rev. Mol. Cell Biol. **4**(6), 497–502 (2003)
57. A. Sonkar, D.L. Lyngdoh, R. Shukla, H. Shukla, T. Tripathi, S. Ahmed, Point mutation A394E in the central intrinsic disordered region of Rna14 leads to chromosomal instability in fission yeast. Int. J. Biol. Macromol. **119**, 785–791 (2018)
58. J. Kalita, R. Shukla, T. Tripathi, Structural basis of urea-induced unfolding of Fasciola gigantica glutathione S-transferase. J. Cell. Physiol. **234**(4), 4491–4503 (2019)
59. P.B. Chetri, R. Shukla, T. Tripathi, Identification and characterization of cytosolic malate dehydrogenase from the liver fluke Fasciola gigantica. Sci. Rep. **10**(1), 13372 (2020)
60. U. Mayor, C.M. Johnson, V. Daggett, A.R. Fersht, Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proc. Natl. Acad. Sci. **97**(25), 13518–13522 (2000)
61. U. Mayor, N.R. Guydosh, C.M. Johnson, J.G. Grossmann, S. Sato, G.S. Jas, S. Freund, D.O. Alonso, V. Daggett, A.R. Fersht, The complete folding pathway of a protein from nanoseconds to microseconds. Nature **421**(6925), 863–867 (2003)
62. H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, Molecular dynamics with coupling to an external bath. J. Chem. Phys. **81**, 3684–3690 (1984)
63. J.E. Basconi, M.R. Shirts, Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations. J. Chem. Theory Comput. **9**(7), 2887–2899 (2013)
64. M.R. Shirts, Simple quantitative tests to validate sampling from thermodynamic ensembles. J. Chem. Theory Comput. **9**(2), 909–926 (2013)
65. H.C. Andersen, Molecular dynamics simulations at constant pressure and/or temperature. J. Chem. Phys. **72**, 2384–2393 (1980)
66. E.A. Koopman, C.P. Lowe, Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations. J. Chem. Phys. **124**(20), 204103 (2006)
67. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. J. Chem. Phys. **126**(1), 014101 (2007)
68. S. Nosé, A molecular dynamics method for simulations in the canonical ensemble. Mol. Phys. **52**, 255–268 (1984)
69. W.G. Hoover, Canonical dynamics: equilibrium phase-space distributions. Phys. Rev. A Gen. Phys. **31**(3), 1695–1697 (1985)
70. M.L.K. Glenn, J. Martyna, Nosé–Hoover chains: the canonical ensemble via continuous dynamics. J. Chem. Phys. **97**, 2635–2643 (1992)
71. S. Roy, T. Mohammad, P. Gupta, R. Dahiya, S. Parveen, S. Luqman, G.M. Hasan, M.I. Hassan, Discovery of harmaline as a potent inhibitor of sphingosine kinase-1: a chemopreventive role in lung cancer. ACS Omega **5**(34), 21550–21560 (2020)
72. F. Naz, F.I. Khan, T. Mohammad, P. Khan, S. Manzoor, G.M. Hasan, K.A. Lobb, S. Luqman, A. Islam, F. Ahmad, M.I. Hassan, Investigation of molecular mechanism of recognition between citral and MARK4: a newer therapeutic approach to attenuate cancer cell progression. Int. J. Biol. Macromol. **107**(Pt B), 2580–2589 (2018)
73. T. Mohammad, S. Siddiqui, A. Shamsi, M.F. Alajmi, A. Hussain, A. Islam, F. Ahmad, M.I. Hassan, Virtual screening approach to identify high-affinity inhibitors of serum and

glucocorticoid-regulated kinase 1 among bioactive natural products: combined molecular docking and simulation studies. Molecules **25**(4), 823 (2020)

74. T. Mohammad, M. Amir, K. Prasad, S. Batra, V. Kumar, A. Hussain, M.T. Rehman, M.F. AlAjmi, M.I. Hassan, Impact of amino acid substitution in the kinase domain of Bruton tyrosine kinase and its association with X-linked agammaglobulinemia. Int. J. Biol. Macromol. **164**, 2399–2408 (2020)
75. S. Fatima, T. Mohammad, D.S. Jairajpuri, M.T. Rehman, A. Hussain, M. Samim, F.J. Ahmad, M.F. Alajmi, M.I. Hassan, Identification and evaluation of glutathione conjugate gamma-l-glutamyl-l-cysteine for improved drug delivery to the brain. J. Biomol. Struct. Dyn. **38**(12), 3610–3620 (2020)
76. V.N. Maiorov, G.M. Crippen, Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J. Mol. Biol. **235**(2), 625–634 (1994)
77. A. Bagaria, V. Jaravine, Y.J. Huang, G.T. Montelione, P. Güntert, Protein structure validation by generalized linear model root-mean-square deviation prediction. Protein Sci. **21**(2), 229–238 (2012)
78. M. Purmonen, J. Valjakka, K. Takkinen, T. Laitinen, J. Rouvinen, Molecular dynamics studies on the thermostability of family 11 xylanases. Protein Eng. Des. Sel. **20**(11), 551–559 (2007)
79. S. Fenwick, S.K. Vanga, A. DiNardo, J. Wang, V. Raghavan, A. Singh, Computational evaluation of the effect of processing on the trypsin and alpha-amylase inhibitor from Ragi (Eleusine coracana) seed. Eng. Rep. **1**(4), e12064 (2019)
80. K. Kobayashi, M.U. Salam, Comparing simulated and measured values using mean squared deviation and its components. Agron. J. **92**(2), 345–352 (2000)
81. A. Cooper, Thermodynamic fluctuations in protein molecules. Proc. Natl. Acad. Sci. **73**(8), 2740–2741 (1976)
82. S.L. Rath, M. Tripathy, N. Mandal, How does temperature affect the dynamics of SARS-CoV-2 M proteins? Insights from molecular dynamics simulations. J. Membr. Biol. **255**, 341–356 (2022)
83. I. Sarkar, A. Sen, In silico screening predicts common cold drug dextromethorphan along with prednisolone and dexamethasone can be effective against novel coronavirus disease (COVID-19). J. Biomol. Struct. Dyn. **40**(8), 3706–3710 (2022)
84. A. Kuzmanic, B. Zagrovic, Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. Biophys. J. **98**(5), 861–871 (2010)
85. N.N. Cob-Calan, L.A. Chi-Uluac, F. Ortiz-Chi, D. Cerqueda-García, G. Navarrete-Vázquez, E. Ruiz-Sánchez, E. Hernández-Núñez, Molecular docking and dynamics simulation of protein β-tubulin and antifungal cyclic lipopeptides. Molecules **24**(18), 3387 (2019)
86. J.A. Ippolito, R.S. Alexander, D.W. Christianson, Hydrogen bond stereochemistry in protein structure and function. J. Mol. Biol. **215**(3), 457–471 (1990)
87. W.B. Cardoso, S.A. Mendanha, Molecular dynamics simulation of docking structures of SARS-CoV-2 main protease and HIV protease inhibitors. J. Mol. Struct. **1225**, 129143 (2021)
88. M.S. Weiss, M. Brandl, J. Sühnel, D. Pal, R. Hilgenfeld, More hydrogen bonds for the (structural) biologist. Trends Biochem. Sci. **26**(9), 521–523 (2001)
89. Z. Bikadi, L. Demko, E. Hazai, Functional and structural characterization of a protein based on analysis of its hydrogen bonding network by hydrogen bonding plot. Arch. Biochem. Biophys. **461**(2), 225–234 (2007)
90. I. Habib, S. Khan, T. Mohammad, A. Hussain, M.F. Alajmi, T. Rehman, F. Anjum, M.I. Hassan, Impact of non-synonymous mutations on the structure and function of telomeric repeat binding factor 1. J. Biomol. Struct. Dyn. **40**(19), 9053–9066 (2022)
91. P. Gupta, T. Mohammad, P. Khan, M.F. Alajmi, A. Hussain, M.T. Rehman, M.I. Hassan, Evaluation of ellagic acid as an inhibitor of sphingosine kinase 1: a targeted approach towards anticancer therapy. Biomed. Pharmacother. **118**(109245), 25 (2019)
92. P. Gupta, T. Mohammad, R. Dahiya, S. Roy, O.M.A. Noman, M.F. Alajmi, A. Hussain, M.I. Hassan, Evaluation of binding and inhibition mechanism of dietary phytochemicals with sphingosine kinase 1: towards targeted anticancer therapy. Sci. Rep. **9**(1), 019–55199 (2019)

93. M. Amir, T. Mohammad, V. Kumar, M.F. Alajmi, M.T. Rehman, A. Hussain, P. Alam, R. Dohare, A. Islam, F. Ahmad, M.I. Hassan, Structural analysis and conformational dynamics of STN1 gene mutations involved in coat plus syndrome. Front. Mol. Biosci. **6**, 41 (2019)
94. M.F. AlAjmi, S. Khan, A. Choudhury, T. Mohammad, S. Noor, A. Hussain, W. Lu, M.S. Eapen, V. Chimankar, P.M. Hansbro, S.S. Sohal, A.M. Elasbali, M.I. Hassan, Impact of deleterious mutations on structure, function and stability of serum/glucocorticoid regulated kinase 1: a gene to diseases correlation. Front. Mol. Biosci. **8**, 780284 (2021)
95. M. Adnan, S. Koli, T. Mohammad, A.J. Siddiqui, M. Patel, N. Alshammari, F. Bardakci, A.M. Elasbali, M.I. Hassan, Searching for novel anaplastic lymphoma kinase inhibitors: structure-guided screening of natural compounds for a tyrosine kinase therapeutic target in cancers. OMICS **26**(8), 461–470 (2022)
96. W. Li, J. Ma, S. Wu, J. Zhang, J. Cheng, The effect of hydrogen bond on the thermal and mechanical properties of furan epoxy resins: molecular dynamics simulation study. Polym. Test. **101**, 107275 (2021)
97. Y. Gao, Y. Mei, J. Zhang, Treatment of hydrogen bonds in protein simulations, in *Advanced Materials for Renewable Hydrogen Production, Storage and Utilization*, (InTech, London, 2015), pp. 121–136
98. D. Pantoja-Uceda, M.T. Pastor, J. Salgado, A. Pineda-Lucena, E. Pérez-Payá, Design of a bivalent peptide with two independent elements of secondary structure able to fold autonomously. J. Pept. Sci. **14**(7), 845–854 (2008)
99. S. Zhang, N. Yuan, W. Li, C. Wang, F. Li, J. Xu, T. Suo, A close look at the conformational transitions of a helical polymer in its response to environmental stimuli. AIP Adv. **11**(8), 085107 (2021)
100. J.J. Tanner, Empirical power laws for the radii of gyration of protein oligomers. Acta Crystallogr. D Struct. Biol. **72**(10), 1119–1129 (2016)
101. S. Ghahremanian, M.M. Rashidi, K. Raeisi, D. Toghraie, Molecular dynamics simulation approach for discovering potential inhibitors against SARS-CoV-2: a structural review. J. Mol. Liq. **354**, 118901 (2022)
102. M.Y. Lobanov, N. Bogatyreva, O. Galzitskaya, Radius of gyration as an indicator of protein structure compactness. Mol. Biol. **42**(4), 623–628 (2008)
103. D. Li, H.-T. Li, H. Wu, Y. Wang, Using the group contribution method and molecular dynamics to predict the glass transition temperatures and mechanical properties of poly-(p-phenylenediamine-alt-2, 6-diformyl multiphenyl). J. Chem. Res. **45**(9-10), 823–830 (2021)
104. S. Ausaf Ali, I. Hassan, A. Islam, F. Ahmad, A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and unfolded states. Curr. Protein Pept. Sci. **15**(5), 456–476 (2014)
105. J.A. Marsh, S.A. Teichmann, Relative solvent accessible surface area predicts protein conformational changes upon binding. Structure **19**(6), 859–867 (2011)
106. M. Moret, G. Zebende, Amino acid hydrophobicity and accessible surface area. Phys. Rev. E **75**(1), 011920 (2007)
107. S. Sivakumar, M. Mohan, O. Franco, B. Thayumanavan, Inhibition of insect pest α-amylases by little and finger millet inhibitors. Pestic. Biochem. Physiol. **85**(3), 155–160 (2006)
108. B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. **55**(3), 379–400 (1971)
109. S.L. Rath, K. Kumar, Investigation of the effect of temperature on the structure of SARS-Cov-2 spike protein by molecular dynamics simulations. Front. Mol. Biosci. **7**, 583523 (2020)
110. G.G. Maisuradze, A. Liwo, H.A. Scheraga, Principal component analysis for protein folding dynamics. J. Mol. Biol. **385**(1), 312–329 (2009)
111. C.C. David, D.J. Jacobs, Principal component analysis: a method for determining the essential dynamics of proteins. Methods Mol. Biol. **1084**, 193–226 (2014)
112. G.G. Maisuradze, A. Liwo, H.A. Scheraga, Relation between free energy landscapes of proteins and dynamics. J. Chem. Theory Comput. **6**(2), 583–595 (2010)

113. B. Borges, G. Gallo, C. Coelho, N. Negri, F. Maiello, L. Hardy, M. Würtele, Dynamic cross correlation analysis of thermus thermophilus alkaline phosphatase and determinants of thermostability. Biochim. Biophys. Acta **1865**(7), 129895 (2021)
114. S. Kumar, P.A. Deshpande, Structural and thermodynamic analysis of factors governing the stability and thermal folding/unfolding of SazCA. PLoS One **16**(4), e0249866 (2021)
115. Z. Chen, Y. Fu, W. Xu, M. Li, Molecular dynamics simulation of barnase: contribution of noncovalent intramolecular interaction to thermostability. Math. Probl. Eng. **2013**, 504183 (2013)
116. W. Gu, T. Wang, J. Zhu, Y. Shi, H. Liu, Molecular dynamics simulation of the unfolding of the human prion protein domain under low pH and high temperature conditions. Biophys. Chem. **104**(1), 79–94 (2003)
117. E.R. Henry, R.B. Best, W.A. Eaton, Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations. Proc. Natl. Acad. Sci. U S A **110**(44), 17880–17885 (2013)
118. D.B. Singh, R.K. Pathak, *Bioinformatics: Methods and Applications* (Academic Press, San Diego, 2021)
119. S. Piana, J.L. Klepeis, D.E. Shaw, Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr. Opin. Struct. Biol. **24**, 98–105 (2014)
120. S.A. Hollingsworth, R.O. Dror, Molecular dynamics simulation for all. Neuron **99**(6), 1129–1143 (2018)
121. D.A. Beck, V. Daggett, Methods for molecular dynamics simulations of protein folding/unfolding in solution. Methods **34**(1), 112–120 (2004)
122. J. Ferina, V. Daggett, Visualizing protein folding and unfolding. J. Mol. Biol. **431**(8), 1540–1564 (2019)
123. L. Duan, X. Guo, Y. Cong, G. Feng, Y. Li, J.Z.H. Zhang, Accelerated molecular dynamics simulation for helical proteins folding in explicit water. Front. Chem. **7**, 540 (2019)

Principles, Methods, and Applications of Protein Folding Inside Cells

Subhashree Sahoo, Kummari Shivani, Amrita Arpita Padhy, Varsha Kumari, and Parul Mishra

Abstract The complex mechanism of protein folding in cells has intrigued the scientific community for decades. The physical and chemical forces that drive protein folding have been deciphered by intense experimental, theoretical, and computational methods. Although folding kinetics has been pursued for many proteins in vitro, the crowded cellular environment and the complex solution properties might differentially impact the folding process in vivo. Sampling the native conformation from thousands of folding intermediate states occurs within a timescale of milliseconds to seconds in cells, and replicating this dynamic and highly complex phenomenon under cell-free conditions is an extremely challenging task. The absence of critical regulatory parameters leads to protein misfolding and aggregation. Biophysical approaches like *in-cell* NMR spectroscopy, *in-cell* FRET, and FlAsH have facilitated studies focused on analysing protein folding in cells. In this chapter, we discuss the roles of various cellular factors in the protein folding process inside cells, methods to study the folding process in a dynamic cellular environment, and elaborate on the emerging applications of this knowledge to engineer proteins with native-like folds but novel properties of interest.

Keywords Protein folding · Conformation · Native · Unfolding · Misfolding · Chaperones · De novo protein design · Interactions · Misfolding diseases · *In-cell* NMR · FRET · Drug design

Subhashree Sahoo and Kummari Shivani contributed equally with all other contributors.

S. Sahoo · K. Shivani · A. A. Padhy · V. Kumari · P. Mishra (✉)
Department of Animal Biology, School of Life Sciences, University of Hyderabad, Hyderabad, India
e-mail: pmsl@uohyd.ac.in

P. Saudagar, T. Tripathi (eds.), *Protein Folding Dynamics and Stability*,
https://doi.org/10.1007/978-981-99-2079-2_13

1 Introduction

Proteins are highly sophisticated macromolecules that regulate all the major activities in a cell. The structural hierarchy observed in their native forms underscores the diversity and complexity of their functions. They exist either as monomers or multimers, which assemble with precision and integrity to define the characteristics of a living system. The intricate folding pathway traversed by every protein molecule to acquire the biologically active three-dimensional (3D) structure is tightly regulated to avoid misfolded moieties from being synthesized in the cells. Unravelling this mechanism has been one of the greatest challenges faced by scientists for the last 60 years. It was considered that along with the structural properties of the unique sequence, many other cellular components might also be critical in assisting the folding process in an extremely crowded cellular environment. Some of the early insights into the protein folding problem came from the pioneering work of Linus Pauling and his contemporaries in the early twentieth century. Pauling used the ideas of structural chemistry to unravel the chemical bonds that stabilize complex biopolymers like proteins. His research on haemoglobin hypothesized that its distinct structural features are the prerequisites of its native conformation for oxygen-binding functions. Pauling and his colleague, Alfred Mirsky, introduced the concept of hydrogen bonds and demonstrated it to be the most crucial form of interaction in proteins besides the peptide bonds between the amino acids. Further, Pauling and his colleagues modelled the first known structure of alpha keratin in which they detailed the interactions between the amino acids in its polypeptide chain as well as the distances between the repeating units to form helical conformation [1, 2]. The pioneering work of crystallographers like Astbury on globular and fibrous protein structures and contributions of Nobel laureates like Dorothy Hodgkin for insulin structure and Max Perutz for the atomic structure of haemoglobin have presented several fundamental concepts of protein structures to the scientific community.

While many studies performed until the early 1980s clarified that small intramolecular interaction within a unique primary sequence encodes secondary structures that adopt a defined 3D conformation, the prominent role of hydrophobic interactions distributed both locally and non-locally conveyed the essential role of side chains in protein folding. It is seen that if the distribution of hydrophobic and polar residues is conserved in a protein, but its primary sequence is altered, it still folds into its native structure [3]. Hence, two proteins with different hydrophobic contents tend to acquire different native structures. However, since the energetic barrier between the folded and denatured states is very low (1–5 kcal/mol), it is pertinent to think that all the intermolecular forces play a critical role in protein folding. Later in 1961, the classical experiment of Anfinsen projected the thermodynamic basis of protein to fold correctly. Following the refolding kinetics of a 124-amino acid enzyme ribonuclease A, the experiment demonstrated that folding into a biologically active protein is independent of any cellular factors that might affect kinetic parameters till a thermodynamically stable protein is obtained under a particular physiological condition [4]. This finding won him the Nobel Prize in 1972 and introduced the idea of in vitro protein folding. Although it appeared that the folding code for a protein was embedded in the underlying amino acid sequence and the cellular environment had little influence on the folding properties of a protein, it was imperative to decode the cellular properties that influence the stabilization of secondary and tertiary conformations [4].

A landmark finding stemmed from the observations of Cyrus Levinthal in 1969, who first noted that proteins could convert quickly to their native states in just a few microseconds, strongly contradicting that proteins perform a random search for all possible configurations. He suggested that there must be energy-driven or kinetically favoured folding pathways, and elucidating the physical mechanism behind protein folding could lead to fast algorithmic predictions of native structures from amino acid sequences. This concept of protein folding intrigued a large number of scientific studies, and many researchers began exploring the ‘folding intermediates’ and postulated many protein folding models to support their theories predicting folding pathways. The nucleation growth model [5], diffusion–collision–adhesion model [6], framework model [7], hydrophobic collapse model [8], jigsaw puzzle model [9], and nucleation-condensation model [10, 11] are major examples of protein folding models that well supported and took forward the idea of Levinthal by interpreting different routes of protein folding mechanism [5–12].

Recognizing the concept that an unfolded protein can explore ensembles of states through alternate folding pathways to achieve its biologically relevant configuration, the ‘new view,’ also known as the ‘energy landscape view’, was described in the 1990s. The view indicates protein folding as a funnel diagram, where the width of the funnel indicates all the possible conformations for a polypeptide chain, while the decreasing free energy function is represented by its depth. The native structures are presented at the lowest level (global minimum), with the denatured state for each protein shown at the wider top region. Hence, the top of the folding funnel represents the non-native states with high conformational entropy, and as the funnel gets narrower, conformational entropy decreases with the compact near-native states at the bottom. This concept of protein folding also helped to understand several other protein folding complexities. It denied the assumption that the energy landscape of a real protein is a perfect funnel and supported a somewhat rugged and bumpy shape of the funnel due to slow folding steps and kinetically distinguishable conformations. Being a form of microscopic view, the funnel indicates that each molecule follows a different route to its native folded structure, facing different obstacles on the energy landscape. It is possible that while one molecule starting from uphill may reach the bottom unhindered, another molecule may get kinetically trapped in a non-native conformation due to entropic barriers [13–15].

Computational analysis has accelerated the structural prediction of proteins from their amino acid sequences, with secondary structure prediction algorithms among the earliest [16]. It was also suggested that the protein folding rate depends on the topological characteristics of its native structure [17, 18]. The folding rate was higher in α-helices and turns than in β-sheets [18]. Other topological parameters like chain length, secondary structure content, contact distance of residues from the sequence, and overall contact distance also regulate the protein folding speed [19–22]. Besides the primary protein structure, other factors like solution properties, intermolecular interactions, protein modifications, and environmental perturbations can also disturb protein folding and unfolding equilibria. Interactions with chaperones assist the proteins in unfolding and refolding until their stable globular conformation is obtained, without which they are targeted for degradation. Failures of these regulated checkpoints yield misfolded proteins that can be physiologically fatal. Deeper insights into the fundamental principles of protein folding have not only helped to understand disease pathologies but have also been instrumental in the de novo design of proteins with novel functions. These avenues offer practical applications of our fundamental knowledge of protein folding to developing therapeutics and medical diagnostics. In this chapter, we discuss various techniques to study protein folding in cells with an emphasis on the cellular factors that affect the folding potentials of a protein. We further highlight the applications and methods to study protein folding in vivo.

2 Protein Folding in Cells

Protein folding is initiated at the N-terminus of a newly synthesized protein. The co-translational folding couples with the synthesis and vectorial folding of nascent polypeptides as they emerge from the ribosome exit tunnel. The small sections of the protein chain that are being synthesized constantly fold and unfold to enable correct intermolecular interactions until the full-length protein is released from the ribosome. The compactly folded intermediate often comprises native-like secondary structures called the 'molten-globule' state, which can rapidly transition to the unfolded form. With rapid equilibrium kinetics, the co-translationally folded proteins span a much smaller conformational space characterized by nested free energy landscapes [23]. Interestingly, the conformational entropy of full-length protein is closely related to the length of the emerging polypeptide chain. Hence, the rate of translation and rate of folding together dictate the number of unfolded states that can be sampled by a full-length polypeptide chain, with faster-translated chains having a higher risk of encountering misfolded intermediates [24]. While this may largely be true for most cellular proteins, O'Brien et al. demonstrate that protein segments that are prone to be misfolded tend to fold correctly if they have higher rates of codon translation [25], suggesting that a much more sophisticated orchestration of mechanisms underlie co-translational folding of proteins. The co-translational protein folding occurs in eukaryotes due to a slower translation rate, larger amino acid composition, and complexity of polypeptides. About one-third of the total proteins present in *E. coli* have been found to fold co-translationally. In *E. coli,* the average rate of protein synthesis is ~20 amino acids/s, whereas, in the case of eukaryotes, it is ~6 amino acids/s. While cytosolic proteins fold and function in the cytosol, secretory proteins need compartmentalization for modification and maturation. Ribosomes act as the site of the co-translational folding of proteins into 3D structures. Before emerging into the cytosol, the nascent polypeptide chain interacts with a number of ribosome-associated protein factors inside the ribosome exit tunnel. The dimensions of the ribosome exit tunnel and the slowly diffusing and semi-structured properties of water inside it facilitate the nascent polypeptide to fold into a compact structure. Based on the rRNAs and r-proteins, the tunnel comprises three regions, the upper region made by U2585 and A2062 from domain V of the 23 s rRNA, the central part includes uL4 and uL22 protein loops, and the lower region has the nucleotides of I and III domains of 23 s rRNA [26, 27]. These tunnel proteins provide electrostatic potential, which is essential for proper stability and conformation of the polypeptide.

Fluorescence resonance energy transfer (FRET) and biochemical assays based on site-specific cysteine tagging (PEGylation) of the nascent polypeptide chains revealed that the transmembrane segments could form secondary structures at the distal end of the tunnel. Cryo-electron microscopy (cryo-EM) of the ADRla zinc finger domain of ADR1 protein demonstrated its folding a few angstroms deep inside the peptide exit tunnel [28]. Another study showed that the folding of the N-terminal domain of Hem-K, a small five-helix protein domain, occurs ~33 Å away from the exit tunnel. The native full-length protein forms immediately after the domain emerges from the exit tunnel [29]. The volume confinement effect inside the deeper region of the tunnel has a compelling effect on folding and stabilizing the protein. Although tertiary structure formation is limited within the tunnel, the wide vestibule region located ~80 Å from the peptidyl transfer centre at the end of the tunnel supports the formation of tertiary structure. According to the force profile assay, small protein domains of molecular weight of 10 kDa containing alpha-helix or beta-sheet structures fold within the first 80 Å of the peptide exit tunnel. The space available for the polypeptide expands suddenly when adequate amino acids emerge from the exit vestibule, ultimately allowing the formation of the tertiary structure [26]. Many experimental studies using cysteine mapping PEGylation, FRET, and cryo-EM support the fact that the ribosome entropically stabilizes helix formation. The domains with beta-strand conformation and the proteins possessing repeat motifs can fold on the ribosomes sequentially [26, 30].

Soon after emerging from the ribosome exit tunnel, the secretory proteins are translocated across or inserted into the membrane of the endoplasmic reticulum (ER) with the help of a hydrophobic stretch of amino acids at their N-terminus recognized by the signal recognition protein (SRP), a cytosolic ribonucleoprotein, which then subsequently attaches to the ER membrane for translocating the nascent polypeptide into the ER. While the C-terminal segment of the nascent polypeptide chain is present in the tunnel of the 60 s ribosome, the N-terminal region is located in the protein conducting channel, which is a part of the translocon complex. Sec61α, β, γ, and TRAM are the four transmembrane proteins of the translocon complex that form an aqueous channel. The translocon is an excellent example of complex functional coordination to regulate protein folding in cells. Its association with the ribosome closes its cytosolic end and, at the same time, opens it at the ER lumen side to allow passage of the polypeptide traversing it. The lateral opening of the translocon also aids the translocation of membrane proteins and the insertion of their hydrophobic transmembrane domains into the lipid bilayer. The cytosolic regions of single-spanning membrane proteins are folded after the complete synthesis of the polypeptide, while for the multi-spanning membrane proteins, the cytosolic regions are folded co-translationally. The signal sequence of a protein defines its efficiency in translocating across the ER membrane [31, 32]. These proteins are further assisted by a host of ER-resident chaperones like BiP and Grp170 to acquire native conformations.

Interaction with SRP is not an essential requirement for proteins that are translocated across the ER membrane only after complete synthesis by the ribosome. Supported by a host of cytosolic chaperones, their co-chaperones, and the chaperonin complex, which assist the proteins to be unfolded or loosely folded, these intermediates are recognized by Sec63, a specific membrane receptor that further presents them to the translocon complex. Following the integration of the unfolded polypeptide chain into the translocon, the three covalent modifications, including signal peptide cleavage, N-terminal glycosylation, and disulphide bond formation, are critical for the folding of the protein in the ER lumen. While emerging from the exit tunnel, the nascent polypeptide chains interact with several ribosome-associated proteins such as peptide deformylase, methionine aminopeptidase, SRPs, and trigger factors, as well as with some molecular chaperones that aid the co-translational protein folding and their translocation to membrane compartments [27]. As evident from the earlier studies, co-translational protein folding can also be affected due to the presence of cofactors or ligands. The binding of ATP facilitates the folding of the N-terminal subdomain of human CFTR protein, which ultimately promotes the co-translational folding of other domains [23]. However, the proteins that cannot be properly folded to their native conformation are eventually degraded by the cytosolic 26S proteasome via the ER-associated degradation process [32]. This multistep process of protein folding progresses efficiently with the help of many cellular factors. The following section highlights their impact on the folding properties of proteins in cells.

3 Cellular Factors That Facilitate Protein Folding in Cells

3.1 Macromolecular Crowding and Compartmentalization

The cellular environment is over-occupied with diverse macromolecules such as proteins, carbohydrates, nucleic acids, ribosomes, etc. These biological molecules occupy 25–40% of most cellular compartments suggesting that these solution properties should be mimicked in the in vitro protein folding studies to accurately estimate the contribution of neighbouring solutes to the free energy of folding. Macromolecular crowding imposes excluded volume effects (hard interaction) and chemical interactions (soft interactions), which cause alteration in molecular diffusion, molecular collisions, protein folding, protein stability, protein–protein interaction, and enzyme kinetics. The hard interactions are entropic in nature, whereas soft interactions are mostly weak and enthalpic [33]. According to the statistical thermodynamic model of Zhou and Hall, high concentrations of larger solutes stabilize proteins, while high concentrations of smaller solutes facilitate destabilization of the same. On the contrary, medium-sized solutes with low concentrations stabilize proteins, whereas those with high concentrations destabilize them. Another model by Minton demonstrated that the excluded volume exerted due to stable inert macro solutes stabilizes the globular native folded protein [34].

A study on the effect of macromolecular crowding on the refolding of hen lysozyme shows that crowding positively affects the rate of formation of initial disulphide bonds. It increases the intrinsic properties of folding polypeptide chains and chaperone substrate interaction [35]. Supporting evidence exhibits that polyethylene glycol and ficoll act as crowding agents in vitro, which facilitate the refolding of denatured RNase A without forming any aggregate. Polyethylene glycol or ficoll of different molecular weights was added to the simple cell-free system, and the folding, compaction, and activity of the protein were observed using various biophysical approaches. CD and NMR spectra analysis revealed that the crowding agents, when present in the same volume as the intracellular milieu, enhanced the refolding of RNase A [36]. A large-scale analysis of *E. coli* cytoplasmic proteins was conducted under a cell-free translation condition to study the effect of crowders on protein folding and aggregation properties of these proteins. The study demonstrated variable effects of the crowding agents. While dextran inhibited the aggregation of positively charged proteins, it decreased the solubility of proteins, with aggregation-prone to the tertiary folds [37]. A combined in vitro and in silico analysis of apoflavodoxin was performed in the presence of macromolecular crowding agents. Far-UV CD data revealed that the addition of ficoll 70 enhances the formation of secondary structure as well as overall protein stability in a concentration-dependent manner [38]. ^{15}N relaxation dispersion technique was used to observe the difference in the protein folding and unfolding kinetics in the presence and absence of crowders. The unfolded proteins were found to be more compact in the presence of crowding agents. The ^{1}H-^{15}N correlation spectrum obtained shows that the folding rate of protein was increased by 80% at 20 °C in the presence of crowding agents, while at 30 °C, the rate of protein folding was increased by only 33%. This result indicates that higher temperature adversely affects protein folding [39]. Furthermore, analysing the impact of macromolecular crowding on reversed proteolysis unveiled that the crowding agents enhance the proteosynthesis of polypeptide products with the protein assembly into a coiled-coil structure [40]. On the other hand, studies on cytochrome c demonstrate that crowding agents did not affect the structural stability of the protein; instead, they enhanced its thermal stability. Moreover, the net effect of excluded volume on protein depends on several factors such as protein stability, conformation, crowders to protein size ratio, and geometry of crowding agents [41].

3.2 Inter- and Intramolecular Interactions in Proteins

Several noncovalent interactions, including hydrogen bond (H-bond), hydrophobic, coulombic, and van der Waals interactions, allow cooperative interactions in the amino acid residues along a polypeptide chain to form the native structure. Other interactions, such as C–H–O hydrogen bonding, C5 hydrogen bonding, chalcone bonding, and interactions involving aromatic rings, also contribute to the overall protein folding mechanism [42]. The H-bonds are formed between the hydrogen atoms and an electronegative donor atom to stabilize the α-helices and β-sheets. These bonds further provide directionality and rigidity to interactions. Proteins form approximately 1.1 H-bonds per residue during folding [43]. Interestingly, the activation energy of a hydrogen bond in proteins present in an aqueous solution is about 0.5–1.5 kcal/mol, which is much lower than its energy (~5–6 kcal/mol) in isolated form, suggesting that the water environment dramatically lowers the entropy of proteins which reduces the energetics of this bond within soluble proteins. Studies on 151 H-bond variants (Tyr-Phe, Thr-Val, and Ser-Ala variants) of polar side chains from 15 proteins revealed that H-bonds from peptide bonds make 65% of the H-bonds in a folded protein and contribute more towards protein folding and stability than the H-bonds made by -OH groups of Ser, Tyr, and Thr [44]. Interestingly, a nonpolar environment with a lower dielectric constant increases the strength of H-bonds by 1 kcal/mol [45]. For an aggregating peptide, H-bonding increases cooperatively when it coalesces with other aggregating species [46].

Apart from conventional H-bond, non-canonical H-bonds (C–H–O) involving carbon as hydrogen donors are also vital for protein stability, where C–H protons of the main chain serve as the most common C–H–O donors. C5 H-bonds found in β-sheets are formed when amide protons of β-sheets donate an intra-residue H-bond to its own carbonyl oxygen, and disturbance in this bonding changes the stability of β-sheets. The weak interaction occurs between adjacent carbonyl groups in the backbone due to the donation of lone pair (*n*) electron density from carbonyl oxygen into the π^* orbital of another carbonyl group is involved in stabilizing α-helix, 3_{10} helix, and polyproline II geometries. Cation-π, X–H–π, π–π, anion–π, sulphur–arene interaction and chalcogen bonding are some of the secondary interactions involving sidechain atoms that contribute to the overall energy of protein folding [42].

As a protein contains both hydrophobic and hydrophilic regions, understanding the exact mechanism of the hydrophobic effect in protein folding is critical. Thermodynamic analysis and comparison between cold and hot denaturation provide a better insight into the molecular determinants of the hydrophobic effect. Results have shown that yeast frataxin protein sampled different folding intermediates at low and high temperatures due to changes in the number of inter and intra-H-bonds between water and the protein surface. A thermally denatured state shows only compact secondary structures with reduced H-bonds between the water and protein surface [47]. Hydrophobicity indicates the reduced unfavourable interactions between the hydrophobic residues and water molecules, which converts the molecule into a more condensed structure [48]. A model proposed to depict the role of hydrophobicity in the initiation and propagation of protein folding illustrated that the nonpolar residues come in contact with each other due to the negative free energy of hydrophobic interactions and form hydrophobic pockets at the initiation site of protein folding [49].

Van der Waals and electrostatic interactions also contribute to the folding and structural stability of a protein. Geometric parameters such as distance between sidechain groups help speculate the van der Waals interactions [50]. Electrostatic interactions regulate the thermodynamics and kinetic properties, such as the binding and folding of proteins, as they are affected by the non-homogenous medium surrounding the protein charges [51]. Like hydrophobic interactions between non-polar residues, electrostatic interactions between polar charged residues also support protein folding stability. Electrostatic interactions are temperature-dependent, and an increase in temperature favours the contribution of electrostatic interactions to protein folding [51]. A study on the electrostatic effect using mutations in the residues of three proteins showed a decrease in repulsive electrostatic interaction due to the reduction in enthalpy. While electrostatic interactions in the unfolded proteins are more prone to ion shielding, in the folded state, this interaction is less dependent on the concentration of ions [52].

3.3 Post-Translational Modifications

Post-translational modifications (PTMs) following protein biosynthesis refer to the chemical changes occurring in a protein due to proteolytic cleavage and covalent attachment of small chemical moieties to specific amino acid residues. PTMs have a vital role in innumerable biological processes, such as modulation of protein structure and dynamics, protein folding, protein–protein interactions, protein solubility, protein localization, enzyme conformation, and activity. The most common PTMs include attachment or removal of various modifying groups, disulphide bond formation, and defined cleavage of precursor proteins. Even though all amino acid residues of a polypeptide chain can undergo PTMs, the side chains containing strong or weak nucleophiles are the most commonly affected sites. As documented in previous studies, around 300–400 PTMs have been identified as of date. Phosphorylation, acetylation, methylation, ubiquitination, glycosylation, nitration, SUMOylation, sulfation, palmitoylation, and myristoylation are characterized as the major PTMs. More than 140 chemical moieties are involved in different PTMs [53, 54]. The functional maturation of amyloid precursor protein (APP) demonstrates the various post-translational modifications that APP undergoes during its movement through the secretory pathway. The signal peptide that directs it into the ER is cleaved before it enters the ER, where N-glycosylation of Asn residues further matures it to reach the Golgi complex. In ER, it again undergoes O-glycosylation and sulfation of Try residues. Aberrant glycosylation of APP reduces its solubility, while the addition of sialic acid to its oligosaccharide chain makes APP more soluble [55].

Most of the secretory and type I membrane proteins contain cleavable signal peptides near the N-terminal hydrophobic domain that is recognized by the SRP for targeting into ER. The cleavable signal peptides consist of a basic N-terminal domain, a middle hydrophobic H-domain in the core of the lipid bilayer, and a polar C-terminal domain that contains the signal peptide recognized by the signal peptidase. The cleavage of signal peptide mostly occurs co-translationally, while for some other proteins, such as HIV envelope glycoprotein, this event is considered late post-translational. The peptide cleavage for hemagglutinin, preprolactin, and tyrosinase occurs after the synthesis of 120 amino acids of the polypeptide chains. Mutations in the signal peptide sequence have been found to inhibit its binding with SRP, translocation, or cleavage, which eventually is responsible for a number of diseases such as Ehlers-Danlos syndrome, autosomal dominant familial isolated hypoparathyroidism, factor X syndrome, etc. [31, 56].

N-linked glycosylation is considered the most complex and ubiquitous modification process required for proper folding and the quality control of proteins in the ER. While in the case of prokaryotes, glycans are attached after protein folding, N-linked glycosylation occurs before folding in eukaryotes for generating diverse proteins. Glycans are bulky hydrophilic polymers that play a significant role in increasing the solubility and stability of protein against proteolysis. This process refers to the transfer of the oligosaccharide portion of lipid-linked oligosaccharide by oligosaccharyltransferase onto the Asn residue of the Asn-X-Thr/Ser consensus sequence residing in a polypeptide chain. The oligosaccharyltransferase catalyses the transfer of the lipid-linked oligosaccharide, which is composed of three glucose, nine mannose, and two N-acetyl glucosamine, to the Asn residue. The covalent binding of glycans to proteins improves the kinetics and thermodynamics of protein folding, stability of protein, and immune recognition [31, 56, 57]. However, according to a recent study, N-linked glycosylation does not have a significant effect on modulating protein conformation, although it enhances protein stability [58].

Disulphide bond formation also stabilizes folding intermediates or native states of secretory or membrane-bound proteins. Oxidizing reagents or enzymes induce disulphide linkage between two cysteines. The thiol-disulphide oxidoreductase of the protein disulphide isomerase (PDI) catalyses disulphide bond formation, reduction, and isomerization [59]. PDI acts as a chaperone to recognize folded or unfolded protein conformations, though it possesses a higher affinity for misfolded proteins via hydrophobic interaction. It positively regulates the degradation of misfolded protein via the ER-associated degradation pathway [60, 61].

3.4 Chaperones

Molecular chaperones are conserved multidomain proteins that play a fundamental role in proteostasis (Fig. 1 and Table 1). Most of the chaperones involved in protein quality control use the ATP cycle for the folding or unfolding of non-native polypeptides. Chaperones like Hsp70 sequentially interact with nascent polypeptide chains on the ribosome and assist their proper folding. This process restricts the premature folding or misfolding of the growing polypeptide until all the amino acids are synthesized, and the protein is folded into a stable, compact structure. While

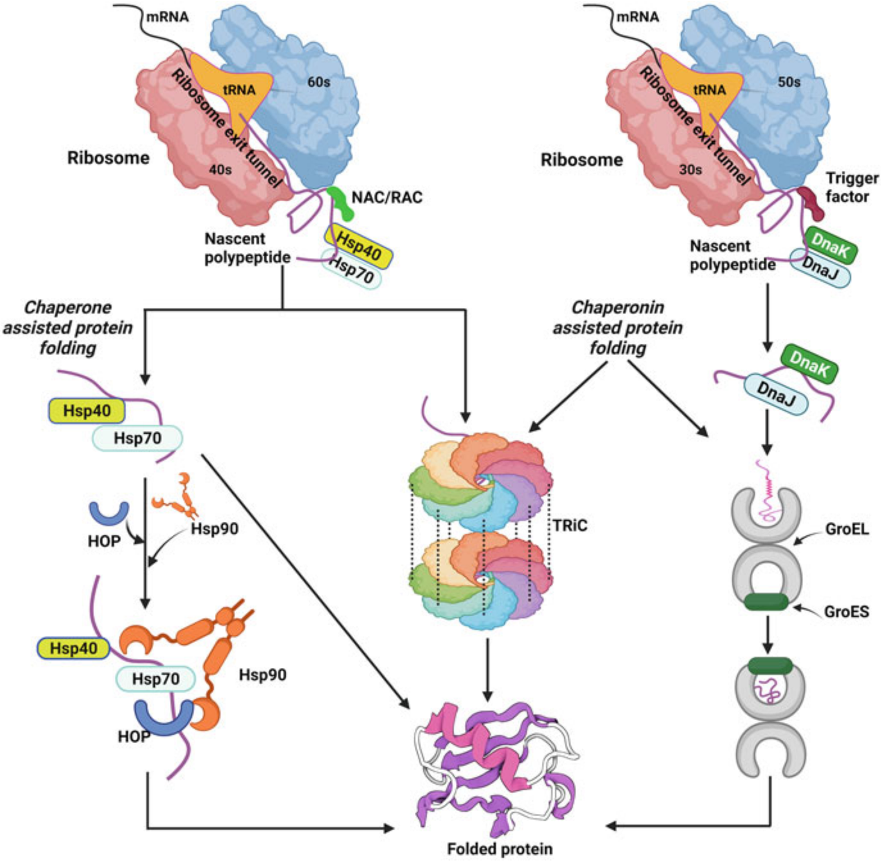

Fig. 1 Schematic representation of the role of chaperones and chaperonins in the protein folding pathway in cells. As the mRNA template translates into a nascent polypeptide, it progresses to acquire the native conformation. On emerging from the ribosome exit tunnel, it is recognized by the trigger factor (in prokaryotes) and the ribosome-associated complex (RAC) or nascent polypeptide-associated complex (NAC) (in eukaryotes). The nascent polypeptide either binds co-translationally to Hsp40 and Hsp70, which assist it in acquiring the native fold, or this complex is further associated with Hsp90 and its cochaperone-like HOP. Protein folding in bacteria is assisted by DnaK and DnaJ (bacterial homologs of Hsc70 and Hsp40), which channel the nascent polypeptide into GroEL/ES chaperonins. Tric is the chaperonin of eukaryotes with a cage-like structure to capture the unfolded forms of cellular proteins

Hsp70 acts at the early stages of protein folding, Hsp90 stabilizes the metastable conformation of over 200 cellular clients until they are appropriately localized with their ligands. Chaperones like Hsp110 and Hsp70 disaggregate misfolded conformers; further, they can also target them for degradation [66, 67].

Table 1 Molecular chaperones inside cells

Sl. No.	Chaperones		Co-chaperones	Organisms	Compartment	Functions	References
1.	Hsp70		Hsp40, GrpE	Prokaryotes, eukaryotes	Cytosol, nucleus, ER, mitochondria	Protein folding and degradation, protein transport to ER and mitochondria, mRNA translation	[62, 63]
2.	Hsp90		Hop, Hip, Hsp70, Immunophillins, Grp78	Prokaryotes, eukaryotes	Cytosol, nucleus, ER, mitochondria	Folding and assembly of secretory proteins, stress response, cell signalling, tumour suppression	[64]
1.3.	Hsp60 (Chaperonins)	Group I (GroES/GroEL)	Hsp10	Prokaryotes	Cytosol, mitochondria, chloroplast	Folding and stabilization of over-produced proteins	[65]
		Group II (TriC, CCT)	Prefoldin/GimC	Eukaryotes	Cytosol	In vivo refolding of unfolded polypeptides, cell cycle, cytoskeletal regulation	

3.4.1 Hsp70 Chaperone

Hsp70, a 70 kDa protein, is considered the most abundant chaperone present in all cellular compartments. The dynamic interactions between its N-terminal ATPase domain and C-terminal substrate-binding domain, assistance by its co-chaperones and nucleotide exchange factors, together regulate the activity of this chaperone. While in the ATP-bound state, the peptide interaction with Hsp70 is dynamic; stable interactions between Hsp70 and unfolded proteins occur in the ADP-bound and apo state of the chaperone. The hydrophobic binding motifs within the core of the folding intermediate bind the chaperone, which assists it in achieving a native-like fold [68]. Hsp70 disassembles clathrin-coated vesicles after clathrin-mediated endocytosis and helps Hsp100 ATPase in disaggregating large aggregates to inhibit cellular toxicity [67]. Interestingly, the activity of many eukaryotic regulatory proteins is governed by their transient interactions with Hsp70 [69].

The co-chaperones of the J-domain family and the substrate proteins enhance the ATPase activity of Hsp70. There are two alternative mechanisms via which Hsp70 supports the substrate proteins to obtain their native folded form. Firstly, its iterative association and dissociation from the substrate protein in a cyclic manner help to maintain a low concentration of free unfolded proteins, allowing them to fold properly and avoid aggregation. Secondly, this cyclic function enables the unfolding of misfolded proteins such as misfolded β-sheets and facilitates their folding into a proper conformation [69]. In the ATP-bound state, the α-helical lid of the substrate-binding domain opens, leading to rapid binding and unbinding of substrate proteins. In contrast, in the ADP-bound state, the lid closes, and stable interactions occur between the chaperone and substrate proteins and direct them for folding [70]. A direct interaction between Hsp70 and its substrates is demonstrated by Lu et al., using molecular simulation to remodel the energy landscape of the Hsp70-mediated protein folding. The interaction of Hsp70 with an unfolded protein reduces the probability of the protein to sample numerous folding intermediates and persists till that protein reaches its near-native conformation [71].

3.4.2 Chaperonins

Chaperonins are a class of molecular chaperones that are involved in the folding of proteins of molecular weight up to 60 kDa by enclosing them inside their double-ringed cavity (Table 1). Their structural differences and dependence on co-chaperones help them capture the substrate proteins in the hydrophobic interior of their cavity [70, 72].

A well-studied example of chaperonins is provided by the GroEL/GroES complex found in the bacterial cytosol. The cylindrical cavity of GroEL is capped by its cofactor GroES and captures the unfolded protein within itself. The partially folded or unfolded protein is protected in the interior cavity of this chaperonin until it acquires favourable folds and is subsequently released into the cytosol by removing the GroES cap by ATP hydrolysis. Interestingly the folding cage of this protein complex not only modifies the unfolded substrate but is also seen to undergo stoichiometric changes from asymmetric to symmetric conformations in the presence of foldable substrate proteins [73]. Following this complex folding mechanism, several models have been proposed to unravel the protein folding mechanism inside the chaperonin cavity. The **passive cage model,** or **Anfinsen's model,** states that protein folding is a slow but spontaneous process. The GRoEL/GRoES chaperonin system provides an infinite dilution environment for proper folding [72, 74]. The **confinement model** explains that the confined space inside the GroEL cage aids protein folding. According to the **iterative annealing model**, the repetitive binding of unfolded denatured protein to GroES and release from GroEL coupled with ATP hydrolysis leads to their folding conformation. The incompletely folded proteins follow further iterations until they reach their proper folding state. The **tethering model** demonstrates that the encapsulated unfolded proteins interact with the hydrophobic residues of the chaperonin cavity and form tether intermediates. This weak hydrophobic interaction allows substrate proteins to undergo conformational changes that ultimately accelerate folding. However, it has been recently found that the denatured protein is not completely enclosed by the chaperonin cavity; instead, it interacts with the hydrophobic residues present at the interface of GroEL [72].

Another ATP-dependent chaperonin contains tailless complex polypeptide 1 (CCT), also known as tailless complex polypeptide 1 ring complex (TRiC), which is also an important regulator of protein homeostasis in cells. This Group II chaperonin and an archaebacterial thermosome are both made up of eight subunits. These subunits have a specific orientation within the ring structure to recognize unfolded proteins. The apical domains of these chaperonins contain finger-like protrusions that act as the lid of the cavity. The opening and closing of this lid in an ATP-dependent manner are similar to that of the GroEL-GroES system. As the ATP reaction cycle, in this case, is much slower than the group I chaperonins, it provides a longer duration for protein folding. TRIC binds to approximately 10% of the cytosolic proteins, including cyclin, tubulin, and other cell cycle regulatory proteins, and assists in their folding [70, 75].

3.4.3 Hsp90 Chaperone

Hsp90, a 90 kDa protein, imparts stability to diverse proteins under cell stress conditions and is aptly recognized as a master regulator in cells. While some proteins like Ste11 bind to Hsp90 to acquire its native conformation, others like src-kinase require the support of Hsp90 while it is transported to the cell membrane, suggesting that functions of Hsp90 in the cell are unique for each client. Its ability to activate and stabilize a subset of clients that cause neurodegeneration or cancer makes Hsp90 a therapeutic target for many diseases. It is a homodimeric protein that assists in protein folding by hydrolysing ATP and binding to over 20 co-chaperones. Asymmetric binding to ATP and co-chaperones by the monomers is implicated in allowing the association of Hsp90 simultaneously to many cellular proteins [76, 77].

Contrary to chaperonins or Hsp70, Hsp90 does not possess a defined substrate-binding region. Large-scale mutagenesis studies have successfully demonstrated specific substrate-binding sites that might be available only during conformational rearrangements tightly regulated by ATP or specific co-chaperone binding [78, 79]. Hsp90 acts downstream of Hsp70 in the protein folding pathway. It provides a larger and more extended surface for substrate binding. The continuous switch between the apo state and ATP-bound structure of Hsp90 is extremely important for its substrate maturation functions. Further, each monomer of Hsp90 comprises an intrinsically disordered stretch of amino acids at its C-terminus. Following the analysis of evolutionarily conserved regions of Hsp90, it has been demonstrated that the charge properties of this disordered structure not only enhance the solubility of Hsp90 independently but also of the complex of Hsp90 with its aggregation-prone clients [80]. Some co-chaperones like Sti1/Hop form a bridge between Hsp70 and Hsp90 by simultaneously binding the C-terminal end of both proteins. However, Hsp90, unlike Hsp70, does not block protein folding. The substrate reaches its folded conformation while still bound to Hsp90 [81]. Interestingly, PTMs like phosphorylation, methylation, acetylation, SUMOylation, *O*-GlcNAcylation, and ubiquitination influence the functions of Hsp90. One of the early reports showed that hyperphosphorylation of this chaperone by casein kinase 2 prevents the maturation of pp60v-src [82], and since then, binding of many substrates and even co-chaperones have been reported to be impacted by structural modifications of Hsp90 [83].

3.5 Solution Properties

Solution properties like physiological ion concentrations, temperature, or the presence of small molecules affect the folding of proteins inside cells. Protein folding occurs in different cellular compartments with different pHs, where cytosol and ER have neutral pH while Golgi is basic with a pH of 5. These pH values are maintained by passive or active proton efflux systems, and the isoelectric points of proteins might have even evolved to utilize these cellular conditions for function with proteases exemplifying proteins that are functional only at the acidic pH of lysosomes. Notably, the averaged pI of proteins in any subcellular compartment is largely different from the pH of that compartment, but the pKa of histidine residues positively correlates with subcellular pH suggesting their critical role in protein stability [84]. The pH of subcellular compartments is tightly coupled to the stability and function of proteins, and even small changes in pH can have drastic physiological consequences [85]. The variation of pH and stability is shown by a bell-shaped curve with maximum stability observed at its pH optima. Studies have shown that proteins are thermodynamically more stable either near-neutral pH or near their isoelectric points. Changes in the protonation status of ionizable groups of some amino acid side chains cause conformational changes. While acidic pH protonates the amino acids, basic pH deprotonates them, thereby altering the interactions between positively and negatively charged groups. As the pH changes, these amino acid side chains either get protonated or deprotonated, changing their ability to form hydrogen bonds. Lower pH results in protonation of the amino acid side chains implicating that the pKa values of these ionizable groups are important in the folding kinetics and can be changed by local changes in the protein's microenvironment. The carboxyl and the phenolic groups remain uncharged when protonated, whereas the nitrogen groups become charged upon protonation. Hence, the alteration in the electrostatic interactions indicates the relation between pH and protein stability. Although the solubility of proteins is a multi-dimensional property, changes in pH play a major role in regulating the stability of their folded conformations. Interestingly, for PrP, conformational transitions from misfolded to folded forms are reversible with changing pH [86]. Considering the complexity of the folding pathway, it seems challenging to probe pH-dependent adaptation in sequence and function of proteins by computational modelling, and more sophisticated techniques should be devised to study the same.

The temperature has also been a prominent factor affecting the folding properties of proteins both in vitro and in vivo. Organisms that dwell at higher temperatures have proteins with higher melting temperatures compared to their homologs in mesophilic organisms. This further implies that proteins that are more stable at higher temperatures have more stable folding intermediates with stronger intramolecular bonds. The process of protein folding is affected at both high and low temperatures. In both cases, the protein unfolds as the hydrogen bonds, disulphide bonds, hydrophobic interactions, and the van der Waals forces are disrupted, whereas the primary structure of the protein remains intact. The hydrophobic bonds primarily affect the stability of the proteins both towards the entropy and enthalpy of the folded conformers; hence temperature alterations will change these parameters, and subsequently, the Gibbs free energy ($\Delta G = \Delta H - T\Delta S$) of folded or unfolded forms. Interestingly, osmotic balancing agents like glycerol have been shown to reduce the thermal denaturation of many cellular proteins [87]. Besides cellular chaperones, these chemical chaperones also assist in protein folding by providing a suitable microenvironment.

Osmolytes are small-sized low molecular weight substances that naturally occur inside the cells. Molecules such as sorbitol, arginine, urea, sucrose, trimethylamine-N-oxide (TMAO), and trehalose are organic osmolytes, whereas ions such as K^+ and Na^+ are inorganic osmolytes. They play a vital role in either inducing protein aggregation or inhibiting the process of protein aggregation. The same osmolyte shows distinctive effects on the aggregation of different proteins, which is entirely dependent on the structural properties of the proteins [88]. In vitro analysis has shown that at higher concentrations, polyols tend to be removed from the vicinity of proteins causing them to form a more compact structure with enhanced stability towards denaturation [89]. Different osmolytes act differently on proteins. Arginine has been the most extensively studied osmolyte. It acts both as a stabilizing and destabilizing agent. Arginine suppresses protein aggregation by keeping the charge constant for the guanidino group at both neutral and alkaline pH [90]. Trehalose is a non-reducing sugar of glucose. It differentially suppresses the aggregation of both

Aβ40 and Aβ42 peptides in Alzheimer's disease, being less efficient in preventing the aggregation of Aβ42 [91]. This is probably because Aβ42 is more hydrophobic than Aβ40 peptide and, consequently, has higher entropy gain during the association of molecules and the aggregation process. Trehalose is unable to compensate for the free energy change in aggregating Aβ42. Taurine, a free amino acid found abundantly in mammalian cells, also serves as an osmolyte and forms favourable interactions with the denatured or unfolded states, further stabilizing them. It is seen to delay the fibrillation of glucagon but promote the aggregation of β-amyloids [88].

4 Protein Misfolding Diseases

Protein misfolding and deposition are a hallmark of an extended series of heterogeneous diseases mentioned in Table 2. Many types of aggregates are characterized by an increased content of β-sheet structures that finally accumulate in cells as fibrillar species. These toxic conformations get deposited in the tissues as well as propagate to neighbouring sites, resulting in several serious diseases. Protein aggregates either accumulate at the site of protein production, leading to a set of localized amyloidosis like Alzheimer's disease (AD, in CNS) and Type 2 diabetes (Pancreas), or they can be transmitted to multiple tissues and organs, yielding systemic amyloidoses like prion diseases (comprising Creutzfeldt-Jakob disease and Fatal Familial Insomnia) [93]. Elevated levels of serum amyloid A or unstable light chain and the presence of genetic mutants of the transthyretin (TTR) protein are a few folding-associated aberrations associated with non-neuropathic amyloidosis. Changes in folding properties of amyloid-β or tau protein in AD and α-synuclein aggregates forming Lewy bodies in Parkinson's disease (PD) cause neuropathic amyloidosis [92]. Expansion of the CAG triplet repeat in a gene results in a misfolded, pathogenic protein causing a neurodegenerative condition linked to polyglutamine diseases, including Huntington's disease and various spinocerebellar ataxias and atrophies [95]. In a few cases, mutations in a specific lysosomal enzyme cause its misfolding in the ER rather than affecting the functionality of the enzyme. Thus, the inability of the mutant enzyme to follow its native conformation results in its inappropriate trafficking to the lysosomes, as seen in lysosomal storage disorders, including Fabry's disease, Gaucher's disease, and Niemann Pick's disease [97]. Other diseases, including certain types of cancers, sickle cell anaemia, cystic fibrosis, phenylketonuria, and atherosclerosis, are also included in the extensive set of protein misfolding diseases, emphasizing the need to understand the complex roles of protein quality control systems in different organelles as they are of valuable significance in comprehending the fate of a protein.

Table 2 Protein misfolding diseases

Sl. No.	Group of diseases	Name of the disease	Misfolded protein(s)	References
1.	Neuropathic amyloidoses	Alzheimer's disease	Amyloid β, tau	[92, 93]
2.		Parkinson's disease	α-Synuclein	
3.		Amyotrophic lateral sclerosis	SOD1, FUS, TDP-43	
4.	Non-neuropathic amyloidoses	Localized AL amyloidosis	Locally secreted monoclonal immunoglobulin light chain	
5.		Systematic AL amyloidosis	Circulating monoclonal immunoglobulin light chain	
6.		AA amyloidosis	Serum amyloid A	
7.		ATTR amyloidosis	Transthyretin	
8.		Type 2 diabetes	Islet amyloid polypeptide	
9.	Prion diseases	Creutzfeldt-Jakob disease	PrP (prion protein)	[94]
10.		Familial insomnia	PrP (prion protein)	
11.	Polyglutamine diseases	Spinocerebellar ataxia	Ataxin	[95, 96]
12.		Huntington's disease	Huntingtin	
13.		Spinobulbar muscular atrophy	Androgen receptor	
14.	Lysosomal storage disorders	Fabry's disease	Alpha-galactosidase	[97]
15.		Gaucher's disease	Beta-glucosidase	
16.		Niemann-Pick type C disease	NPC1	
17.	Other diseases	Cystic fibrosis	Cystic fibrosis transmembrane regulator	[98]
18.		Phenylketonuria	Phenylalanine hydroxylase	[99]
19.		Sickle cell anaemia	Haemoglobin S	[100]
20.		Nephrogenic diabetes insipidus	Aquaporin-2/vasopressin	[101]
21.		Desminopathy	Desmin and beta-crystalline	[102]
22.		Cancer	P53, non-receptor tyrosine kinase	[103, 104]
23.		Marfan's syndrome	Fibrillin	[105]
24.		Scurvy	Collagen	[106]
25.		Atherosclerosis	Modified LDL	[107]
26.		Retinitis pigmentosa	Rhodopsin	[108]
27.		α-1-antitrypsin deficiency	α1 antitrypsin	[109]
28.		Emphysema, COPD	α1 antitrypsin	[110]

5 Biophysical Methods to Study Protein Folding in Cells

As our knowledge about the effects of the cellular environment on protein folding expands, developing methods that can allow us to analyse the folding process in living cells can enable us to exploit this information for engineering proteins with desirable features. Early studies employing mass spectrometry or FRET analysis of target proteins in denaturant-resistant bacterial cells helped to estimate equilibrium constants and differences in the stability of protein of interest (POI) both in situ and in dilute solutions used for in vitro studies [111, 112]. Recent advances involving isotopic labelling or microscopic analysis of fluorescently tagged POI have achieved appreciable success in investigating the *in-cell* folding processes.

5.1 In-Cell *NMR Spectroscopy*

In-cell NMR spectroscopy is a tool for characterizing protein conformers under physiological conditions inside living cells [113]. It provides atomic-level resolution of structural changes associated with the change in solution properties or protein–protein interactions for a target protein. It uses multi-dimensional NMR in conjunction with isotope-labelled proteins, where the chemical shift is used to study protein folding dynamics, intrinsically disordered proteins, and post-translational modification of proteins inside the cells [114].

The prokaryotic POI can be overexpressed in bacteria, while eukaryotic proteins are induced in yeast or insect cells and labelled with NMR-active isotopes to detect chemical shifts. Inducing the target protein reduces the background noise and improves the signal intensities, although overexpression might shift the equilibrium of monomeric proteins and favour their aggregation/oligomerization. Further, this technique does not involve the purification of target proteins which is essentially required in the conventional solution-NMR. However, it has been shown that some eukaryotic proteins can be purified from bacteria and subsequently either microinjected in *Xenopus* oocytes or covalently linked to cell-penetrating peptides [115], which enables the proteins to enter the cells directly. While high molecular weight proteins can also be targeted using *in-cell* NMR, proteins with a large number of intermolecular interactions have lower tumbling and lower rotational correlation times, which impedes the spectral output (Fig. 2). The NMR-active isotopes such as ^{15}N, ^{13}C, and ^{19}F are widely used for labelling target proteins inside cells. However, there are a few challenges associated with these isotopes. For instance, the ^{19}F labelling necessitates the addition of an unnatural amino acid, and this addition impairs the biological functions of the proteins [116] while labelling a protein with ^{13}C gives a low signal-to-noise ratio in its spectra due to the high abundance of carbon within the cell [117]. To overcome these drawbacks, selective labelling of methyl groups of alanine and methionine is effective. ^{15}N labelling also proved effective because nitrogen is present in very small amounts among the other cellular

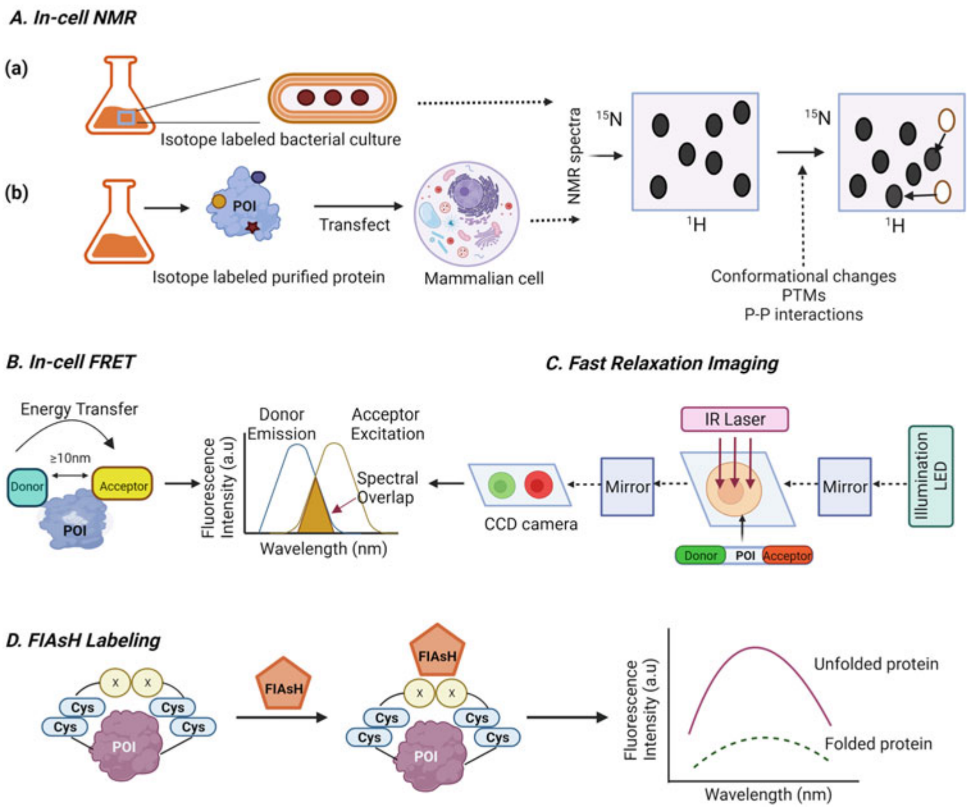

Fig. 2 Schematic representation of biophysical methods to study protein folding inside cells. (**a**) *In-cell* NMR: where the isotope-labelled POI is directly analysed in bacterial cells (*a*) or purified from bacteria and transfected to eukaryotic cells (*b*) for visualizing the chemical shift in atoms upon conformational changes, protein–protein interaction, or PTMs of proteins. (**b**) *In-cell* FRET: where changes in fluorescence signal of acceptor fluorophore are recorded when the acceptor at one end of the POI comes within a 10 nm distance of donor fluorophore. The spectral overlap indicates the folding of POI. (**c**) Fast Relaxation Imaging: involves imaging a FRET-labelled protein as cells get exposed to rapid temperature fluctuations. The shift in folding equilibrium is observed via fluorescence microscopy. (**d**) Cytosolic proteins can be labelled with FLAsH, which reacts with the C-C-X-X-C-C motif to investigate the stability of proteins inside living cells. Unfolded proteins show relatively higher fluorescence than proteins in their native state

components. To date, most *in-cell* NMR experiments have been carried out by labelling proteins with ^{15}N.

The intrinsically disordered proteins (IDPs) such as α-synuclein and tau have been widely studied using *in-cell* NMR. 2D ^{1}H-^{15}N correlation NMR spectra were used to compare the conformation of α-synuclein across different cell lines, like neuronal B65 cells, SK-N-SH cells, and RCSN-3 cells. The study revealed that the monomeric disordered conformation of α-synuclein was consistent in varied intracellular conditions. The study also showed that the N-terminus of α-synuclein is acetylated in cells and its N and C-terminus transiently engage with cytoplasmic components and/or membranes, but they do not interact stably with cellular membranes [118]. The effect of oxidative stress on α-synuclein was identified by injecting ^{15}N-labelled, N-terminally acetylated, and methionine (Met1, Met5, Met116, and Met127) oxidized α-synuclein into non-neuronal and neuronal cells. Time-resolved *in-cell* NMR revealed that C-terminal methionine oxidation preferentially inhibited Tyr125 phosphorylation, implying that changes in the cellular environment can affect post-translational modifications of α-synuclein, and this, in turn, controls its conformational landscape [119]. *In-cell* NMR was done on isotope-enriched tau in HEK-293T cells, which showed the interaction of tau with microtubules (MT) primarily at its MT-binding repeats. Interestingly, when phosphorylated tau was introduced into HEK-293T cells, disease-associated phosphorylation of tau was promptly removed, revealing a possible cellular protective mechanism under stressful conditions [120]. *In-cell* NMR has also shown the folding pathway of wild-type and mutant superoxide dismutase (SOD1) proteins [121]. Wild-type SOD1 can bind zinc on its own to stabilize the native structure of the monomer, but it requires the copper chaperone for SOD1 (Ccs) for copper insertion in its active site and the formation of disulphide bonds. Some fALS-linked mutations impede zinc binding and cause SOD1 to unfold irreversibly, creating cytotoxic aggregates. Ccs's SOD-like domain works as a molecular chaperone, stabilizing mutant SOD1 and permitting zinc binding and protein maturation [122]. *In-cell* NMR has characterized the mutants as unstructured species of SOD1, and their abundance can be prevented by Ccs.

5.2 In-Cell *FRET*

The Förster (or fluorescence) resonance energy transfer (FRET) has been successfully used to investigate the protein dynamics, folding kinetics, and structural changes in living cells. It can either be based on the energy transfer between two fluorophores where the energy is transmitted non-radiatively from the excited donor to the acceptor via a long-range intermolecular dipole-dipole coupling or on a split-reporters where the binding of two parts of a reporter protein is coupled to the folding of POI (Fig. 2) [123, 124]. FRET signals are obtained when the fluorescent donor and acceptor are situated within the Förster radius (around 3–6 nm). In this method, either the POI is genetically tagged with a fluorescent probe at both N- and

C-terminus or can be fluorescently labelled with exogenous fluorescent probes. The exogenous probes like organic fluorophores can modify proteins expressed with unnatural amino acids, which can easily react with organic probes *in-cell* or in vitro, and then the modified protein be delivered into living cells [125] by microinjecting it into living cells [126]. A single-molecule FRET (smFRET) analysis of conformational changes in Raf kinase was carried out using the fluorescent fusion protein GFP-Raf-YFP expressed in HeLa cells. In live cells, alternative laser excitation (ALEX) was utilized to evaluate the native state of Raf while it was experiencing native interactions with other intrinsic proteins and the reaction network of the signal transduction pathway. For cytosolic Raf, three conformational states, including the inactive, closed-form, the active open form, and the inactive fully-open form, were seen, which underwent spontaneous changes between conformational states when the epidermal growth factor (EGF) was stimulated. Interestingly, the S621A mutation in Raf causes the conformational state distribution to shift to an inactive fully open state [125]. This study suggests that smFRET can be used to detect conformational changes in other cytosolic proteins in living cells as a result of intracellular interactions.

5.3 Fast Relaxation Imaging (FREI)

FREI is used to study the fast macromolecular dynamics inside the cells under temperature fluctuation. A modest temperature up or down jump is applied to a cell, and then a FRET-labelled protein is tracked while screening the response of the whole cell with the help of an epifluorescent microscope. Using FREI, one can study protein folding kinetics and protein–protein interactions inside the cells. With a diffraction-limited spatial resolution, FREI can be utilized to analyse protein folding in a variety of cells [127]. The POI should be labelled with two fluorescent probes to monitor FRET changes where the AcGFP acts as a donor and the mCherry probe is commonly used as an acceptor (Fig. 2). While selecting a FRET pair, it is essential to consider that they retain their fluorescent properties between temperatures ranging from 20 °C to 50 °C. Dhar et al. investigated the thermodynamics and kinetics of various cellular compartments. Each cellular compartment is a unique microenvironment that affects the stability and function of interacting macromolecules by modulating the energy landscape. FRET-PGK (FRET-phosphoglycerate kinase) was introduced into the cell nucleus and ER using localization markers. After analysis, it was found that PGK-FRET was more stable in the nucleus than in the ER [128]. Effects of temperature fluctuations on stability have also been explored using FREI for proteins stabilized on hydrogels that are extensively used in drug delivery. The microenvironment experienced by proteins in hydrogels enables many interactions crucial for function [129].

5.4 *FlAsH as an* In-Cell *Protein Folding Probe*

For proteins in which large tags might interfere with folding kinetics, protein stability *in-cell* is investigated by employing a dye system of 4′,5′-bis(1,3,2-dithioarsolan-2-yl) fluorescein (FlAsH), a fluorescein analogue containing two arsen oxides. To study the stability of the mammalian cellular retinoic acid-binding protein 1 (CRABP1) in vivo, the Cys-Cys-X-X-Cys-Cys motif of the wild-type and mutant protein was labelled with FlAsH. The treatment of the protein with urea allows it to unfold, and the denatured protein was found to have a higher fluorescence intensity signal than the folded form (Fig. 2). The time course of FlAsH fluorescence demonstrated that the mutant CRABP1 has a relatively higher signal intensity than the wild-type CRABP 1 [112]. The same group of scientists then successfully made fusion proteins by attaching the Htt exon1 with varying lengths of poly Q tracts and the tetra Cys-CRABP1. The result obtained from this study corroborated the time-dependent increase in fluorescence intensity with an increase in poly Q length [130].

6 Applications of Protein Folding in Cells

The principles of protein architecture and interactions have paved the way for de novo protein design and protein engineering with unlimited applications in biomedicine. Initiated by optimal backbone structure and sequence for targeted functions, novel proteins have revolutionized synthetic biology research. These novel proteins or peptides can serve diverse pharmacological benefits and can be inhibitors of pathogenic infections, immune modulators, or self-assembling biomaterials, to name a few.

6.1 *De Novo Protein Design*

Engineering proteins with desirable changes in function or de novo protein design are scientific accolades achievable only by advancements in understanding in vivo protein folding. Some recombinant proteins like insulin and growth hormones are pharmacologically valuable but tend to aggregate during their overexpression in host cells. Modifying these proteins, based on our knowledge of the contributions of each amino acid to native-like folding, to retain their function with enhanced stability has facilitated their utility vastly [131]. Besides focussed modulation of features, directed evolution utilizing large-scale mutant libraries and extensive screening of variants with desired properties has generated proteins with improved enzymatic activities and stabilities [132]. The systematic approach of assessing the impact of each amino acid on protein properties has helped broaden the substrate specificities and acquire novel functions.

De novo protein design aims to create proteins with functions that are not found in naturally occurring proteins. It is facilitated by understanding the physicochemical basis of protein folding and sample sequence space which has been avoided by evolutionary forces shaping the structure and function of known proteins. Since de novo design is conducted computationally, assessing the free energy of the system where it can be expressed is challenging. Nonetheless, it reduces the cost of manufacturing and experimentally testing each computationally designed variant. The concepts of protein folding that underlie this technique include ensuring the burial of hydrophobic residues in the core of the protein, which is packed with polar groups that form intra-chain hydrogen bonds so that the free energy barrier of this conformation cannot be overcome by the unfolding of this protein for polar residues to interact with water. Apart from the parameters of the core, the interaction of side chains of backbone amino acids with their neighbouring atoms and their torsional effects impacts the free energy values of folding. Converging upon energy functions derived from hydrogen bonds, van der Waal forces, steric interactions, electrostatic interactions, and torsional energies of the main chain and side chain, multiple rounds of optimization are required. Generally observed scaffolds of cellular proteins like alpha-beta folds, repeat units as in symmetrical TIM barrel proteins, and parallel helical bundles have been designed successfully and demonstrate high stability experimentally [133]. Proteins can be selected by a local conception approach in which a protein with known structure and functional properties is selected as a scaffold or a global conception approach that entails designing a structure by analogy with one of the protein data bank's classic folds. To locate those that fold into a certain three-dimensional form, genetic approaches can be used to screen a large number of randomized sequences [131]. Many folds found in naturally occurring proteins have also been repurposed for different functions, including the reduction of viral infections, which is a highly sought-after proposition [134]. Peptides (18–47 aa) with enhanced stability to thermal and chemical denaturation, including both D and L amino acids, have also been designed [135]. An interesting study by Lisa et al. tested de novo designed inhibitors with scaffolds of ACE2 helix to bind to the receptor-binding domain of SARS-CoV-2 and found them to have binding affinities in the picomolar range [136].

Helices, helix bundles, ß-hairpins, and ß-sheets are among the secondary and supersecondary structures that have been synthesized and structurally characterized. They have made significant contributions to our understanding of protein secondary structures. A substantial variety of parallel or antiparallel helix bundles have also been constructed. Some exhibited molten globule-like qualities, resulting in the desired fold but little stability. These motifs can be used to design proteins [131]. In the fields of catalysis, metal ion, and heme-binding, the designed polypeptides showed complete functionality. Recently eight-stranded transmembrane β-barrel proteins have been designed using geometric models and Rosetta protein structure simulations. These transmembrane proteins show little homology with the existing transmembrane proteins, which can fold and assemble into both detergent micelles and lipid bilayers [137]. In silico simulation methods for structure prediction, such as trRosetta and AlphaFold, have been employed to generate 2000 new proteins using random amino acid sequences. These newly formed proteins have been found to be quite different from the naturally occurring proteins of the same length in the context of overall sequence and structure [138].

6.2 Drug Design

Proteins and peptides serve a plethora of useful functions in biotechnology as biocatalysts, biosensors, signalling molecules, and high-affinity effectors like antibodies. Emerging research on the sequence-structure-function relation of protein and its interaction patterns has led to the development of computational methods that have been used to create novel proteins and peptides for therapeutic targets against various diseases, as listed below [139].

6.2.1 For Cancer

Inhibitions of signal transduction pathways and angiogenesis, along with induction of apoptosis in tumours, are required in therapeutic peptides and proteins used for cancer treatment. Overcoming the challenges associated with peptide cell penetration and stability, Aftabizadeh et al. designed anti-tumour peptides against acetylated signal transducer and activator of transcription 3 (STAT3), which functions by disrupting STAT3 dimerization and activation [140]. Induction of apoptosis in various cancer cell lines was also achieved by VDAC1-based peptides, which inhibited tumour development by competing with VDAC1 interacting proteins such as Bcl-2, Bcl-xL, and HK [141]. Hao et al. employed a combination of phage display technology and computational methods to identify peptides with a strong binding affinity towards cysteine-rich intestinal protein 1 (CRIP1), a breast cancer biomarker, facilitating early detection of cancer [142]. Similarly, to modulate protein interactions, Istivan et al. used the resonant recognition model (RRM) to design a short bioactive peptide with antitumour/cytotoxic activity against the myxoma virus. This model uses electromagnetic radiation of a defined frequency to identify amino acids essential for protein's activity or stability based on the distribution of electron-free energy along with the amino acids. The computationally designed peptide from this study proved to be an effective candidate for cancer therapy [143].

6.2.2 For Human Immunodeficiency Virus

The advantages associated with peptide therapeutics, including high potency, specific and efficient binding affinity towards target domains, and low drug resistance with minimal side effects, have contributed to the development of peptide therapeutics for human immunodeficiency virus (HIV) treatment [144]. These peptides are targeted against viral or host proteins or interacting partners, which are essential for virus replication. In order to increase immunogenic response and achieve conformational stability against conserved HIV epitope, 4E10, Correia et al. devised a computational method for transplanting 4E10 into scaffold proteins by side-chain grafting and Rosetta [145]. This procedure generated epitopes that bind to a monoclonal antibody (mAb) 4E10 more strongly than 4E10 alone and were observed to block HIV neutralization by HIV-positive sera. Another approach to computational design includes using the de novo design framework tools such as WISDOM [146], which provides a design template for the generation of novel peptides with target protein affinity in a user-friendly way. The structure of the C^{14}-linked peptide in association with the hydrophobic core of gp41 (transmembrane subunit of HIV-1 envelope protein complex) was used as a template to build HIV-1 cell-cell fusion inhibitors by Bellows and colleagues [147]. Recently, structure-based studies led to the design of E1P47-derivative peptides with the ability to inhibit HIV infection in colorectal tissue explants [148]. To overcome the limitations of low anti-HIV activity and a genetic barrier to induce drug resistance associated with peptide drug enfuvirtide (T-20), T-20-based lipopeptide (LP-40) and LP-80 were designed with high potency in vitro and with promising therapeutic efficacy [149].

6.2.3 For Alzheimer's Disease

Contributing to developing effective AD therapies, peptide-based drugs offer greater specificity and efficacy along with the potential of bio-inspired peptides [150] instead of alternate amyloid reduction therapies [151]. Sievers et al. used computationally driven design to anticipate and experimentally validate peptide inhibitors of fibril formation by the tau protein linked to Alzheimer's disease, as well as amyloid that promotes HIV transmission. Briefly, they designed a tight interface between the peptide and the end of the steric-zipper motif present in the amyloid-forming proteins [152], which resulted in the inhibition of fibril elongation. Screening through commercially available non-natural peptides, which maximizes hydrogen bonding and hydrophobic interactions, led to the design of candidate peptide inhibitors for amyloid formation [153]. Targeting the propensity of wild-type Aβ peptides to form oligomers and fibrils, Rajadas et al. designed a mutant Aβ peptide with two-point mutations known to promote β–strand character. Incubation of the two peptides in solution led to stabilization of the wild-type Aβ peptide, suggesting inhibition of Aβ aggregation [154]. In a recent study, peptide binders of two key amyloid segments (KLVFFA and GGVVIA) of Aβ42 were designed using RosettaDesign, which exhibited inhibition of Aβ fibril formation. Modification of the peptide to β-conformation and its targeted binding to the C-terminus of the protein provided a significant increase in selective inhibition of Aβ42 aggregate [155]. The assembly of Aβ monomers into fibrils was obstructed by a peptide comprising only D-amino acids to recognize the core hydrophobic region of amyloid [156]. Further, an in vivo study conducted with DesBP, a rationally designed bicyclic peptide, in the *C. elegans* model of Aβ42 showed modulations in morphology and inhibition of the Aβ-associated toxicity upon binding to Aβ peptide [157].

7 Conclusions

Pathways of protein folding, unfolding, misfolding, and aggregation are dynamically balanced in a cellular environment. Advances in experimental methods and computational analysis have helped explore the complex interplay of cellular factors that regulate protein folding. Molecular chaperones have been identified as major regulators of protein folding under physiological conditions, and effective strategies to upregulate chaperone behaviour can help combat the rising debilitating diseases. Techniques such as chemical exchange saturation transfer-magnetic resonance imaging (CEST-MRI) and smFRET have made it possible to estimate the global status of in vivo protein folding but hold paucity in precision studies of individual proteins and detection of membrane proteins, respectively. The insights into the folding mechanism of membrane proteins inside the cell have been made accessible by recently developed membrane mimics like nanodiscs and cell-unroofing techniques like non-canonical amino acid energy transfer in combination with Anap-cyclen Cu^{2+} resonance energy transfer (ACCuRET). Both ACCuRET and *in-cell* NMR approaches provide information on the conformational dynamics of cytosolic and membrane proteins. Computational platforms are continuously being modified to obtain accurate energy functions for both native and folding intermediates. The robust protein designs hence derived might form the basis for protein folding experiments in the near future. Efforts in the direction of investigating protein folding properties at atomic levels and employing these principles in the de novo design of therapeutic proteins can be immensely effective in mitigating non-curable diseases.

Acknowledgements This work is supported by SERB (ECR/2017/003431), MoE-STARS/STARS-1/PID (STARS1/634), and, IoE (UoH-IoE-RC5-22-011) grants to P.M. Authors acknowledge funding support from DBT-SAHAJ/BUILDER (BT/INF/22/SP41176/2020) grant. P.M. duly acknowledges the support of the Ramalingaswami Fellowship, Department of Biotechnology, GOI. A.A.P is a recipient of the PMRF, GOI.

References

1. L. Pauling, R.B. Corey, H.R. Branson, The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U S A **37**(4), 205–211 (1951)
2. L. Pauling, R.B. Corey, Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc. Natl. Acad. Sci. U S A **37**(5), 235–240 (1951)
3. M.H. Cordes, A.R. Davidson, R.T. Sauer, Sequence space, folding and protein design. Curr. Opin. Struct. Biol. **6**(1), 3–10 (1996)
4. C.B. Anfinsen, E. Haber, M. Sela, F.H. White Jr., The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U S A **47**(9), 1309–1314 (1961)
5. D.B. Wetlaufer, Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. U S A **70**(3), 697–701 (1973)
6. M. Karplus, D.L. Weaver, Protein-folding dynamics. Nature **260**(5550), 404–406 (1976)

7. R.L. Baldwin, How does protein folding get started? Trends Biochem. Sci. **14**(7), 291–294 (1989)
8. K.A. Dill, Theory for the folding and stability of globular proteins. Biochemistry **24**, 1501–1509 (1985)
9. S.C. Harrison, R. Durbin, Is there a single pathway for the folding of a polypeptide chain? Proc. Natl. Acad. Sci. U S A **82**, 4028–4030 (1985)
10. A.R. Fersht, V. Daggett, Protein folding and unfolding at atomic resolution. Cell **108**, 573–582 (2002)
11. A.R. Fersht, Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc. Natl. Acad. Sci. U S A **92**, 10869–10873 (1995)
12. A.R. Fersht, Nucleation mechanisms in protein folding. Curr. Opin. Struct. Biol. **7**, 3–9 (1997)
13. K.A. Dill, H.S. Chan, From levinthal to pathways to funnels. Nat. Struct. Biol. **4**, 10–19 (1997)
14. K.A. Dill, Polymer principles and protein folding. Protein Sci. **8**, 1166–1180 (1999)
15. D. Hamada, S. Segawa, Y. Goto, Non-native a-helical intermediate in the refolding of b-lactoglobulin, a predominantly b-sheet protein. Nat. Struct. Biol. **3**, 868–873 (1996)
16. B. Rost, V.A. Eyrich, EVA: large-scale analysis of secondary structure prediction. Proteins **Suppl 5**, 192–199 (2001)
17. V. Grantcharova, E.J. Alm, D. Baker, A.L. Horwich, Mechanisms of protein folding. Curr. Opin. Struct. Biol. **11**(1), 70–82 (2001)
18. K.W. Plaxco, K.T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. **277**(4), 985–994 (1998)
19. M.M. Gromiha, S. Selvaraj, Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J. Mol. Biol. **310**(1), 27–32 (2001)
20. J.T. Huang, J.P. Cheng, H. Chen, Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins **67**(1), 12–17 (2007)
21. D.N. Ivankov, A.V. Finkelstein, Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc. Natl. Acad. Sci. U S A **101**(24), 8942–8944 (2004)
22. H. Zhou, Y. Zhou, Folding rate prediction using total contact distance. Biophys. J. **82**(1 Pt 1), 458–463 (2002)
23. C.A. Waudby, C.M. Dobson, J. Christodoulou, Nature and regulation of protein folding on the ribosome. Trends Biochem. Sci. **44**(11), 914–926 (2019)
24. A. Borgia, K.R. Kemplen, M.B. Borgia, A. Soranno, S. Shammas, B. Wunderlich, D. Nettels, R.B. Best, J. Clarke, B. Schuler, Transient misfolding dominates multidomain protein folding. Nat. Commun. **6**, 8861 (2015)
25. E.P. O'Brien, M. Vendruscolo, C.M. Dobson, Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. Nat. Commun. **5**, 2988 (2014)
26. A.M.E. Cassaignau, L.D. Cabrita, J. Christodoulou, How does the ribosome fold the proteome? Annu. Rev. Biochem. **89**, 389–415 (2020)
27. A. Javed, J. Christodoulou, L.D. Cabrita, E.V. Orlova, The ribosome and its role in protein folding: looking through a magnifying glass. Acta Crystallogr. D Struct. Biol. **73**(Pt 6), 509–521 (2017)
28. O.B. Nilsson, R. Hedman, J. Marino, S. Wickles, L. Bischoff, M. Johansson, A. Müller-Lucks, F. Trovato, J.D. Puglisi, E.P. O'Brien, R. Beckmann, G. von Heijne, Cotranslational protein folding inside the ribosome exit tunnel. Cell Rep. **12**(10), 1533–1540 (2015)
29. W. Holtkamp, G. Kokic, M. Jäger, J. Mittelstaet, A.A. Komar, M.V. Rodnina, Cotranslational protein folding on the ribosome monitored in real time. Science **350**(6264), 1104–1107 (2015)
30. M. Liutkute, E. Samatova, M.V. Rodnina, Cotranslational folding of proteins on the ribosome. Biomol. Ther. **10**(1), 97 (2020)
31. I. Braakman, D.N. Hebert, Protein folding in the endoplasmic reticulum. Cold Spring Harb. Perspect. Biol. **5**(5), a013201 (2013)

32. E. Swanton, N.J. Bulleid, Protein folding and translocation across the endoplasmic reticulum membrane. Mol. Membr. Biol. **20**(2), 99–104 (2003)
33. M. Sarkar, C. Li, G.J. Pielak, Soft interactions and crowding. Biophys. Rev. **5**(2), 187–194 (2013)
34. A.P. Minton, Effect of a concentrated "inert" macromolecular cosolute on the stability of a globular protein with respect to denaturation by heat and by chaotropes: a statistical-thermodynamic model. Biophys. J. **78**(1), 101–109 (2000)
35. B. Van den Berg, R. Wain, C.M. Dobson, R.J. Ellis, Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. EMBO J. **19**(15), 3870–3875 (2000)
36. N. Tokuriki, M. Kinjo, S. Negi, M. Hoshino, Y. Goto, I. Urabe, T. Yomo, Protein folding by the effects of macromolecular crowding. Protein Sci. **13**(1), 125–133 (2004)
37. T. Niwa, R. Sugimoto, L. Watanabe, S. Nakamura, T. Ueda, H. Taguchi, Large-scale analysis of macromolecular crowding effects on protein aggregation using a reconstituted cell-free translation system. Front. Microbiol. **6**, 1113 (2015)
38. L. Stagg, S.Q. Zhang, M.S. Cheung, P. Wittung-Stafshede, Molecular crowding enhances native structure and stability of alpha/beta protein flavodoxin. Proc. Natl. Acad. Sci. U S A **104**(48), 18976–18981 (2007)
39. X. Ai, Z. Zhou, Y. Bai, W.Y. Choy, 15N NMR spin relaxation dispersion study of the molecular crowding effects on protein folding under native conditions. J. Am. Chem. Soc. **128**(12), 3916–3917 (2006)
40. B.R. Somalinga, R.P. Roy, Volume exclusion effect as a driving force for reverse proteolysis. Implications for polypeptide assemblage in a macromolecular crowded milieu. J. Biol. Chem. **277**(45), 43253–43261 (2002)
41. A. Christiansen, Q. Wang, A. Samiotakis, M.S. Cheung, P. Wittung-Stafshede, Factors defining effects of macromolecular crowding on protein stability: an in vitro/in silico case study using cytochrome c. Biochemistry **49**(31), 6519–6530 (2010)
42. R.W. Newberry, R.T. Raines, Secondary forces in protein folding. ACS Chem. Biol. **14**(8), 1677–1686 (2019)
43. J.K. Myers, C.N. Pace, Hydrogen bonding stabilizes globular proteins. Biophys. J. **71**(4), 2033–2039 (1996)
44. C. Nick Pace, J.M. Scholtz, G.R. Grimsley, Forces stabilizing proteins. FEBS Lett. **588**(14), 2177–2184 (2014)
45. C.N. Pace, Energetics of protein hydrogen bonds. Nat. Struct. Mol. Biol. **16**(7), 681–682 (2009)
46. K. Tsemekhman, L. Goldschmidt, D. Eisenberg, D. Baker, Cooperative hydrogen bonding in amyloid formation. Protein Sci. **16**(4), 761–764 (2007)
47. C. Camilloni, D. Bonetti, A. Morrone, R. Giri, C.M. Dobson, M. Brunori, S. Gianni, M. Vendruscolo, Towards a structural biology of the hydrophobic effect in protein folding. Sci. Rep. **6**, 28285 (2016)
48. L. Lins, R. Brasseur, The hydrophobic effect in protein folding. FASEB J. **9**(7), 535–540 (1995)
49. H.J. Dyson, P.E. Wright, H.A. Scheraga, The role of hydrophobic interactions in initiation and propagation of protein folding. Proc. Natl. Acad. Sci. U S A **103**(35), 13057–13061 (2006)
50. J. Li, Y. Wang, L. An, J. Chen, L. Yao, Direct observation of CH/CH van der Waals interactions in proteins by NMR. J. Am. Chem. Soc. **140**(9), 3194–3197 (2018)
51. H.X. Zhou, X. Pang, Electrostatic interactions in protein structure, folding, binding, and condensation. Chem. Rev. **118**(4), 1691–1741 (2018)
52. A. Azia, Y. Levy, Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. J. Mol. Biol. **393**(2), 527–542 (2009)
53. A.L. Darling, V.N. Uversky, Intrinsic disorder and posttranslational modifications: the darker side of the biological dark matter. Front. Genet. **9**, 158 (2018)

54. S. Ramazi, J. Zahiri, Posttranslational modifications in proteins: resources, tools and prediction methods. Database (Oxford) **2021**, baab012 (2021)
55. N. Georgopoulou, M. McLaughlin, I. McFarlane, K.C. Breen, The role of post-translational modification in beta-amyloid precursor protein processing. Biochem. Soc. Symp. **67**, 23–36 (2001)
56. L. Ellgaard, N. McCaul, A. Chatsisvili, I. Braakman, Co- and post-translational protein folding in the ER. Traffic **17**(6), 615–638 (2016)
57. J. Breitling, M. Aebi, N-linked protein glycosylation in the endoplasmic reticulum. Cold Spring Harb. Perspect. Biol. **5**(8), a013359 (2013)
58. H.S. Lee, Y. Qi, W. Im, Effects of N-glycosylation on protein conformation and dynamics: protein data bank analysis and molecular dynamics simulation study. Sci. Rep. **5**, 8926 (2015)
59. P.J. Robinson, N.J. Bulleid, Mechanisms of disulfide bond formation in nascent polypeptides entering the secretory pathway. Cell **9**(9), 1994 (2020)
60. R.B. Freedman, P. Klappa, L.W. Ruddock, Protein disulfide isomerases exploit synergy between catalytic and specific binding domains. EMBO Rep. **3**(2), 136–140 (2002)
61. S. Parakh, J.D. Atkin, Novel roles for protein disulphide isomerase in disease states: a double edged sword? Front. Cell Dev. Biol. **3**, 30 (2015)
62. E.A. Craig, Hsp70 at the membrane: driving protein translocation. BMC Biol **16**(1), 11 (2018)
63. R. Rosenzweig, N.B. Nillegoda, M.P. Mayer, B. Bukau, The Hsp70 chaperone network. Nat Rev Mol Cell Biol **20**(11), 665–680 (2019)
64. A. Hoter, M.E. El-Sabban, H.Y. Naim, The HSP90 family: structure, regulation, function, and implications in health and disease. Int J Mol Sci **19**(9), 2560 (2018)
65. C.M.S. Kumar, S.C. Mande, G. Mahajan, Multiple chaperonins in bacteria—novel functions and noncanonical behaviors. Cell Stress Chaperones **20**(4), 555–574 (2015)
66. J.P. Hendrick, F.U. Hartl, The role of molecular chaperones in protein folding. FASEB J. **9** (15), 1559–1569 (1995)
67. H. Saibil, Chaperone machines for protein folding, unfolding and disaggregation. Nat. Rev. Mol. Cell Biol. **14**(10), 630–642 (2013)
68. S. Polier, Z. Dragovic, F.U. Hartl, A. Bracher, Structural basis for the cooperation of Hsp70 and Hsp110 chaperones in protein folding. Cell **133**(6), 1068–1079 (2008)
69. M.P. Mayer, B. Bukau, Hsp70 chaperones: cellular functions and molecular mechanism. Cell. Mol. Life Sci. **62**(6), 670–684 (2005)
70. F.U. Hartl, A. Bracher, M. Hayer-Hartl, Molecular chaperones in protein folding and proteostasis. Nature **475**(7356), 324–332 (2011)
71. J. Lu, X. Zhang, Y. Wu, Y. Sheng, W. Li, W. Wang, Energy landscape remodeling mechanism of Hsp70-chaperone-accelerated protein folding. Biophys. J. **120**(10), 1971–1983 (2021)
72. F. Motojima, How do chaperonins fold protein? Biophysics (Nagoya-shi) **11**, 93–102 (2015)
73. S. Haldar, A.J. Gupta, X. Yan, G. Miličić, F.U. Hartl, M. Hayer-Hartl, Chaperonin-assisted protein folding: relative population of asymmetric and symmetric GroEL:GroES complexes. J. Mol. Biol. **427**(12), 2244–2255 (2015)
74. A.J. Gupta, S. Haldar, G. Miličić, F.U. Hartl, M. Hayer-Hartl, Active cage mechanism of chaperonin-assisted protein folding demonstrated at single-molecule level. J. Mol. Biol. **426** (15), 2739–2754 (2014)
75. T. Lopez, K. Dalton, J. Frydman, The mechanism and function of group II chaperonins. J. Mol. Biol. **427**(18), 2919–2930 (2015)
76. J.M. Flynn, P. Mishra, D.N. Bolon, Mechanistic asymmetry in Hsp90 dimers. J. Mol. Biol. **427**(18), 2904–2911 (2015)
77. P. Mishra, D.N. Bolon, Designed Hsp90 heterodimers reveal an asymmetric ATPase-driven mechanism in vivo. Mol. Cell **53**(2), 344–350 (2014)
78. L. Jiang, P. Mishra, R.T. Hietpas, K.B. Zeldovich, D.N. Bolon, Latent effects of Hsp90 mutants revealed at reduced expression levels. PLoS Genet. **9**(6), e1003600 (2013)
79. P. Mishra, J.M. Flynn, T.N. Starr, D.N.A. Bolon, Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. Cell Rep. **15**(3), 588–598 (2016)

80. N.W. Pursell, P. Mishra, D.N. Bolon, Solubility-promoting function of Hsp90 contributes to client maturation and robust cell growth. Eukaryot. Cell **11**(8), 1033–1041 (2012)
81. T. Morán Luengo, M.P. Mayer, S.G.D. Rüdiger, The Hsp70-Hsp90 chaperone cascade in protein folding. Trends Cell Biol. **29**(2), 164–177 (2019)
82. E.G. Mimnaugh, P.J. Worland, L. Whitesell, L.M. Neckers, Possible role for serine/threonine phosphorylation in the regulation of the heteroprotein complex between the hsp90 stress protein and the pp60v-src tyrosine kinase. J. Biol. Chem. **270**(48), 28654–28659 (1995)
83. S.J. Backe, R.A. Sager, M.R. Woodford, A.M. Makedon, M. Mollapour, Post-translational modifications of Hsp90 and translating the chaperone code. J. Biol. Chem. **295**(32), 11099–11117 (2020)
84. P. Chan, J. Warwicker, Evidence for the adaptation of protein pH-dependence to subcellular pH. BMC Biol. **7**, 69 (2009)
85. K. Talley, E. Alexov, On the pH-optimum of activity and stability of proteins. Proteins **78**(12), 2699–2706 (2010)
86. T.C. Bjorndahl, G.P. Zhou, X. Liu, R. Perez-Pineiro, V. Semenchenko, F. Saleem, S. Acharya, A. Bujold, C.A. Sobsey, D.S. Wishart, Detailed biophysical characterization of the acid-induced PrP(c) to PrP(β) conversion process. Biochemistry **50**(7), 1162–1173 (2011)
87. C.R. Brown, L.Q. Hong-Brown, W.J. Welch, Correcting temperature-sensitive protein folding defects. J. Clin. Invest. **99**(6), 1432–1444 (1997)
88. F. Macchi, M. Eisenkolb, H. Kiefer, D.E. Otzen, The effect of osmolytes on protein fibrillation. Int. J. Mol. Sci. **13**(3), 3801–3819 (2012)
89. M.M. Santoro, Y. Liu, S.M. Khan, L.X. Hou, D.W. Bolen, Increased thermal stability of proteins in the presence of naturally occurring osmolytes. Biochemistry **31**(23), 5278–5283 (1992)
90. T. Arakawa, D. Ejima, K. Tsumoto, N. Obeyama, Y. Tanaka, Y. Kita, S.N. Timasheff, Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects. Biophys. Chem. **127**(1–2), 1–8 (2007)
91. R. Liu, H. Barkhordarian, S. Emadi, B.P. Chan, M.R. Sierks, Trehalose differentially inhibits aggregation and neurotoxicity of beta-amyloid 40 and 42. Neurobiol. Dis. **20**(1), 74–81 (2005)
92. A.L. Clos, R. Kayed, C.A. Lasagna-Reeves, Association of skin with the pathogenesis and treatment of neurodegenerative amyloidosis. Front. Neurol. **3**, 5 (2012)
93. A. Nevone, G. Merlini, M. Nuvolone, Treating protein misfolding diseases: therapeutic successes against systemic amyloidoses. Front. Pharmacol. **11**, 1024 (2020)
94. M.D. Geschwind, Prion diseases. Continuum (Minneap. Minn.) **21**(6 Neuroinfectious Disease), 1612–1638 (2015)
95. E.N. Minakawa, Y. Nagai, Protein aggregation inhibitors as disease-modifying therapies for polyglutamine diseases. Front. Neurosci. **15**, 621996 (2021)
96. Shao J, Diamond MI. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. Hum. Mol. Genet. **16**(Spec No. 2), R115–R123 (2007)
97. A. Sun, Lysosomal storage disease overview. Ann. Transl. Med. **6**(24), 476 (2018)
98. S. Naehrig, C.M. Chao, L. Naehrlich, Cystic fibrosis. Dtsch. Arztebl. Int. **114**(33–34), 564–574 (2017)
99. R.A. Williams, C.D. Mamotte, J.R. Burnett, Phenylketonuria: an inborn error of phenylalanine metabolism. Clin. Biochem. Rev. **29**(1), 31–41 (2008)
100. R.V. Gardner, Sickle cell disease: advances in treatment. Ochsner J. **18**(4), 377–389 (2018 Winter)
101. D. Bockenhauer, D.G. Bichet, Pathophysiology, diagnosis and management of nephrogenic diabetes insipidus. Nat. Rev. Nephrol. **11**(10), 576–588 (2015)
102. L.G. Goldfarb, M. Olivé, P. Vicart, H.H. Goebel, Intermediate filament diseases: desminopathy. Adv. Exp. Med. Biol. **642**, 131–164 (2008)
103. N. Rivlin, R. Brosh, M. Oren, V. Rotter, Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis. Genes Cancer **2**(4), 466–474 (2011)

104. R. Butti, S. Das, V.P. Gunasekaran, A.S. Yadav, D. Kumar, G.C. Kundu, Receptor tyrosine kinases (RTKs) in breast cancer: signaling, therapeutic implications and challenges. Mol. Cancer **17**(1), 34 (2018)

105. G. Pepe, B. Giusti, E. Sticchi, R. Abbate, G.F. Gensini, S. Nistri, Marfan syndrome: current perspectives. Appl. Clin. Genet. **9**, 55–65 (2016)

106. K. Wang, H. Jiang, W. Li, M. Qiang, T. Dong, H. Li, Role of vitamin C in skin diseases. Front. Physiol. **9**, 819 (2018)

107. S.C. Bergheanu, M.C. Bodde, J.W. Jukema, Pathophysiology and treatment of atherosclerosis: current view and future perspective on lipoprotein modification treatment. Neth. Heart J. **25**(4), 231–242 (2017)

108. W.A. Baumgartner, Etiology, pathogenesis, and experimental treatment of retinitis pigmentosa. Med. Hypotheses **54**(5), 814–824 (2000)

109. M. Torres-Durán, J.L. Lopez-Campos, M. Barrecheguren, M. Miravitlles, B. Martinez-Delgado, S. Castillo, A. Escribano, A. Baloira, M.M. Navarro-Garcia, D. Pellicer, L. Bañuls, M. Magallón, F. Casas, F. Dasí, Alpha-1 antitrypsin deficiency: outstanding questions and future directions. Orphanet J. Rare Dis. **13**(1), 114 (2018)

110. M. Goldklang, R. Stockley, Pathophysiology of emphysema and implications. Chronic Obstr. Pulm. Dis. **3**(1), 454–458 (2016)

111. S. Ghaemmaghami, T.G. Oas, Quantitative protein stability measurement in vivo. Nat. Struct. Biol. **8**(10), 879–882 (2001)

112. Z. Ignatova, L.M. Gierasch, Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling. Proc. Natl. Acad. Sci. U S A **101**(2), 523–528 (2004)

113. E. Luchinat, L. Banci, In-cell NMR: a topical review. IUCrJ **4**(Pt 2), 108–118 (2017)

114. L. Barbieri, E. Luchinat, L. Banci, Characterization of proteins by in-cell NMR spectroscopy in cultured mammalian cells. Nat. Protoc. **11**(6), 1101–1111 (2016)

115. K. Inomata, A. Ohno, H. Tochio, S. Isogai, T. Tenno, I. Nakase, T. Takeuchi, S. Futaki, Y. Ito, H. Hiroaki, M. Shirakawa, High-resolution multi-dimensional NMR spectroscopy of proteins in human cells. Nature **458**(7234), 106–109 (2009)

116. C. Li, G.F. Wang, Y. Wang, R. Creager-Allen, E.A. Lutz, H. Scronce, K.M. Slade, R.A. Ruf, R.A. Mehl, G.J. Pielak, Protein (19)F NMR in Escherichia coli. J. Am. Chem. Soc. **132**(1), 321–327 (2010)

117. Z. Serber, W. Straub, L. Corsini, A.M. Nomura, N. Shimba, C.S. Craik, P. Ortiz de Montellano, V. Dötsch, Methyl groups as probes for proteins and complexes in in-cell NMR experiments. J. Am. Chem. Soc. **126**(22), 7119–7125 (2004)

118. F.X. Theillet, A. Binolfi, B. Bekei, A. Martorana, H.M. Rose, M. Stuiver, S. Verzini, D. Lorenz, M. van Rossum, D. Goldfarb, P. Selenko, Structural disorder of monomeric α-synuclein persists in mammalian cells. Nature **530**(7588), 45–50 (2016)

119. A. Binolfi, A. Limatola, S. Verzini, J. Kosten, F.X. Theillet, H.M. Rose, B. Bekei, M. Stuiver, M. van Rossum, P. Selenko, Intracellular repair of oxidation-damaged α-synuclein fails to target C-terminal modification sites. Nat. Commun. **7**, 10251 (2016)

120. S. Zhang, C. Wang, J. Lu, X. Ma, Z. Liu, D. Li, Z. Liu, C. Liu, In-cell NMR study of tau and MARK2 phosphorylated tau. Int. J. Mol. Sci. **20**(1), 90 (2018)

121. D.R. Rosen, T. Siddique, D. Patterson, D.A. Figlewicz, P. Sapp, A. Hentati, D. Donaldson, J. Goto, J.P. O'Regan, H.X. Deng, et al., Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature **362**(6415), 59–62 (1993) Erratum in: Nature. 1993;364(6435):362

122. E. Luchinat, L. Banci, In-cell NMR in human cells: direct protein expression allows structural studies of protein folding and maturation. Acc. Chem. Res. **51**(6), 1550–1557 (2018)

123. S. Cabantous, Y. Rogers, T.C. Terwilliger, G.S. Waldo, New molecular reporters for rapid protein folding assays. PLoS One **3**(6), e2387 (2008). Erratum in: PLoS One. 2008;3(6).

124. L. Foit, G.J. Morgan, M.J. Kern, L.R. Steimer, A.A. von Hacht, J. Titchmarsh, S.L. Warriner, S.E. Radford, J.C. Bardwell, Optimizing protein stability in vivo. Mol. Cell **36**(5), 861–871 (2009)

125. K. Okamoto, K. Hibino, Y. Sako, In-cell single-molecule FRET measurements reveal three conformational state changes in RAF protein. Biochim. Biophys. Acta Gen. Subj. **1864**(2), 129358 (2020)

126. M. Sustarsic, A.N. Kapanidis, Taking the ruler to the jungle: single-molecule FRET for understanding biomolecular structure and dynamics in live cells. Curr. Opin. Struct. Biol. **34**, 52–59 (2015)

127. I. Guzman, M. Gruebele, Protein folding dynamics in the cell. J. Phys. Chem. B **118**(29), 8459–8470 (2014)

128. A. Dhar, K. Girdhar, D. Singh, H. Gelman, S. Ebbinghaus, M. Gruebele, Protein stability and folding kinetics in the nucleus and endoplasmic reticulum of eukaryotic cells. Biophys. J. **101** (2), 421–430 (2011)

129. L. Kisley, K.A. Miller, D. Guin, X. Kong, M. Gruebele, D.E. Leckband, Direct imaging of protein stability and folding kinetics in hydrogels. ACS Appl. Mater. Interfaces **9**(26), 21606–21617 (2017)

130. Z. Ignatova, L.M. Gierasch, Extended polyglutamine tracts cause aggregation and structural perturbation of an adjacent beta barrel protein. J. Biol. Chem. **281**(18), 12959–12967 (2006)

131. J.M. Yon, Protein folding: a perspective for biology, medicine and biotechnology. Braz. J. Med. Biol. Res. **34**(4), 419–435 (2001)

132. C. Li, R. Zhang, J. Wang, L.M. Wilson, Y. Yan, Protein engineering for improving and diversifying natural product biosynthesis. Trends Biotechnol. **38**(7), 729–744 (2020)

133. P.S. Huang, S.E. Boyken, D. Baker, The coming of age of de novo protein design. Nature **537** (7620), 320–327 (2016)

134. M.T. Koday, J. Nelson, A. Chevalier, M. Koday, H. Kalinoski, L. Stewart, L. Carter, T. Nieusma, P.S. Lee, A.B. Ward, I.A. Wilson, A. Dagley, D.F. Smee, D. Baker, D.H. Fuller, A computationally designed Hemagglutinin stem-binding protein provides in vivo protection from influenza independent of a host immune response. PLoS Pathog. **12**(2), e1005409 (2016)

135. G. Bhardwaj, V.K. Mulligan, C.D. Bahl, J.M. Gilmore, P.J. Harvey, O. Cheneval, G.W. Buchko, S.V. Pulavarti, Q. Kaas, A. Eletsky, P.S. Huang, W.A. Johnsen, P.J. Greisen, G.J. Rocklin, Y. Song, T.W. Linsky, A. Watkins, S.A. Rettie, X. Xu, L.P. Carter, R. Bonneau, J.M. Olson, E. Coutsias, C.E. Correnti, T. Szyperski, D.J. Craik, D. Baker, Accurate de novo design of hyperstable constrained peptides. Nature **538**(7625), 329–335 (2016)

136. L. Cao, I. Goreshnik, B. Coventry, J.B. Case, L. Miller, L. Kozodoy, R.E. Chen, L. Carter, A. C. Walls, Y.J. Park, E.M. Strauch, L. Stewart, M.S. Diamond, D. Veesler, D. Baker, De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science **370**(6515), 426–431 (2020)

137. A.A. Vorobieva, P. White, B. Liang, J.E. Horne, A.K. Bera, C.M. Chow, S. Gerben, S. Marx, A. Kang, A.Q. Stiving, S.R. Harvey, D.C. Marx, G.N. Khan, K.G. Fleming, V.H. Wysocki, D. J. Brockwell, L.K. Tamm, S.E. Radford, D. Baker, De novo design of transmembrane β barrels. Science **371**(6531), eabc8182 (2021)

138. I. Anishchenko, S.J. Pellock, T.M. Chidyausiku, T.A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A.K. Bera, F. DiMaio, L. Carter, C.M. Chow, G.T. Montelione, D. Baker, De novo protein design by deep network hallucination. Nature **600**(7889), 547–552 (2021)

139. G.A. Khoury, J. Smadbeck, C.A. Kieslich, C.A. Floudas, Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol. **32**(2), 99–109 (2014)

140. M. Aftabizadeh, Y.J. Li, Q. Zhao, C. Zhang, N. Ambaye, J. Song, T. Nagao, C. Lahtz, M. Fakih, D.K. Ann, H. Yu, A. Herrmann, Potent antitumor effects of cell-penetrating peptides targeting STAT3 axis. JCI Insight **6**(2), e136176 (2021)

141. A. Shteinfer-Kuzmine, Z. Amsalem, T. Arif, A. Zooravlov, V. Shoshan-Barmatz, Selective induction of cancer cell death by VDAC1-based peptides and their potential use in cancer therapy. Mol. Oncol. **12**(7), 1077–1103 (2018)

142. J. Hao, A.W. Serohijos, G. Newton, G. Tassone, Z. Wang, D.C. Sgroi, N.V. Dokholyan, J.P. Basilion, Identification and rational redesign of peptide ligands to CRIP1, a novel biomarker for cancers. PLoS Comput. Biol. **4**(8), e1000138 (2008)

143. T.S. Istivan, E. Pirogova, E. Gan, N.M. Almansour, P.J. Coloe, I. Cosic, Biological effects of a de novo designed myxoma virus peptide analogue: evaluation of cytotoxicity on tumor cells. PLoS One **6**(9), e24809 (2011)
144. K. Fosgerau, T. Hoffmann, Peptide therapeutics: current status and future directions. Drug Discov. Today **20**(1), 122–128 (2015)
145. B.E. Correia, Y.E. Ban, M.A. Holmes, H. Xu, K. Ellingson, Z. Kraft, C. Carrico, E. Boni, D.N. Sather, C. Zenobia, K.Y. Burke, T. Bradley-Hewitt, J.F. Bruhn-Johannsen, O. Kalyuzhniy, D. Baker, R.K. Strong, L. Stamatatos, W.R. Schief, Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. Structure **18**(9), 1116–1126 (2010)
146. J. Smadbeck, M.B. Peterson, G.A. Khoury, M.S. Taylor, C.A. Floudas, Protein WISDOM: a workbench for in silico de novo design of biomolecules. J. Vis. Exp. **77**, 50476 (2013)
147. M.L. Bellows, M.S. Taylor, P.A. Cole, L. Shen, R.F. Siliciano, H.K. Fung, C.A. Floudas, Discovery of entry inhibitors for HIV-1 via a new de novo protein design framework. Biophys. J. **99**(10), 3445–3453 (2010)
148. M.J. Gomara, Y. Perez, P. Gomez-Gutierrez, C. Herrera, P. Ziprin, J.P. Martinez, A. Meyerhans, J.J. Perez, I. Haro, Importance of structure-based studies for the design of a novel HIV-1 inhibitor peptide. Sci. Rep. **10**(1), 14430 (2020). https://doi.org/10.1038/s41598-020-71404-0
149. Y. Zhu, H. Chong, D. Yu, Y. Guo, Y. Zhou, Y. He, Design and characterization of cholesterylated peptide HIV-1/2 fusion inhibitors with extremely potent and long-lasting antiviral activity. J. Virol. **93**(11), e02312–e02318 (2019)
150. M.K. Siddiqi, P. Alam, T. Iqbal, N. Majid, S. Malik, S. Nusrat, A. Alam, M.R. Ajmal, V.N. Uversky, R.H. Khan, Elucidating the inhibitory potential of designed peptides against amyloid fibrillation and amyloid associated cytotoxicity. Front. Chem. **6**, 311 (2018)
151. Y.S. Cheng, Z.T. Chen, T.Y. Liao, C. Lin, H.C. Shen, Y.H. Wang, C.W. Chang, R.S. Liu, R.P. Chen, P.H. Tu, An intranasally delivered peptide drug ameliorates cognitive decline in Alzheimer transgenic mice. EMBO Mol. Med. **9**(5), 703–715 (2017)
152. R. Nelson, M.R. Sawaya, M. Balbirnie, A.Ø. Madsen, C. Riekel, R. Grothe, D. Eisenberg, Structure of the cross-beta spine of amyloid-like fibrils. Nature **435**(7043), 773–778 (2005)
153. S.A. Sievers, J. Karanicolas, H.W. Chang, A. Zhao, L. Jiang, O. Zirafi, J.T. Stevens, J. Münch, D. Baker, D. Eisenberg, Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. Nature **475**(7354), 96–100 (2011)
154. J. Rajadas, C.W. Liu, P. Novick, N.W. Kelley, M. Inayathullah, M.C. Lemieux, V.S. Pande, Rationally designed turn promoting mutation in the amyloid-β peptide sequence stabilizes oligomers in solution. PLoS One **6**(7), e21776 (2011)
155. J. Lu, Q. Cao, C. Wang, J. Zheng, F. Luo, J. Xie, Y. Li, X. Ma, L. He, D. Eisenberg, J. Nowick, L. Jiang, D. Li, Structure-based peptide inhibitor design of amyloid-β aggregation. Front. Mol. Neurosci. **12**, 54 (2019)
156. J.R. Horsley, B. Jovcevski, K.L. Wegener, J. Yu, T.L. Pukala, A.D. Abell, Rationally designed peptide-based inhibitor of Aβ42 fibril formation and toxicity: a potential therapeutic strategy for Alzheimer's disease. Biochem. J. **477**(11), 2039–2054 (2020)
157. T. Ikenoue, F.A. Aprile, P. Sormanni, F.S. Ruggeri, M. Perni, G.T. Heller, C.P. Haas, C. Middel, R. Limbocker, B. Mannini, T.C.T. Michaels, T.P.J. Knowles, C.M. Dobson, M. Vendruscolo, A rationally designed bicyclic peptide remodels Aβ42 aggregation in vitro and reduces its toxicity in a worm model of Alzheimer's disease. Sci. Rep. **10**(1), 15280 (2020)