UNIVERSITY OF CALIFORNIA, MERCED

Protein folding as diffusion on a free energy surface: Rates and mechanisms from advanced single-molecule fluorescence techniques

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Chemistry and Chemical Biology

by

Nivin Mothi

Committee in charge:

Professor Andy LiWang, Chair
Professor Anand Bala Subramaniam
Professor Jing Xu
Professor Victor Muñoz

The dissertation of Nivin Mothi, titled Protein folding as diffusion on a free energy surface: Rates and mechanisms from advanced single-molecule fluorescence techniques, is approved, and it is acceptable in quality and form for publication in print and electronically:

_____ Date _____

Prof. Andy LiWang, Chair

_____ Date _____

Prof. Anand Bala Subramaniam

_____ Date _____

Prof. Jing Xu

_____ Date _____

Prof. Victor Muñoz, Graduate advisor

University of California, Merced

2021

Acknowledgments

I would like to thank my advisor Prof. Victor Muñoz for his support and guidance throughout my Ph.D. His dedication and commitment to science are inspiring. His mentoring has prepared me well for my future career in research. I would also like to thank my committee members, Prof. Andy LiWang, Prof. Jing Xu, and Prof. Anand Bala Subramaniam, for all the feedbacks and suggestions. They have helped me grow as a researcher and always encouraged me to put my best work at display. I have had the opportunity to work with Prof. LiWang as his TA, and he has been a great mentor throughout my Ph.D.

I would like to thank all Muñoz lab members with special thanks to Mourad, who knew how to fix pretty much everything in the lab. I would like to thank Abhi and Rama for all the guidance and tips they provided about solving different experimental techniques, and research in general. I would like to extend my heartfelt appreciation and gratitude to my lab mates Suhani, Ameed and Thinh for their support. 6 years of working together, struggling together, and succeeding together have made us more than lab mates. I appreciate every discussion I have had with them.

I would like to thank my friends and family who stood by through this journey. I am forever grateful for the unconditional support and encouragement from my parents (Jameela and Bichu). They always believed in me and stood by my side; I could not have completed my Ph.D. without their support. I would like to offer a special thanks to my sister, Ivin, for her support and for being there for my parents while I was away. I cannot express the love and gratitude I have for my friends Som, Amanda, Kat, John, Alex, and Eric. They made my time in Merced one of the best experiences of my life. They provided the much-needed motivation to keep my spirits high throughout the inevitable obstacles during grad school.

I would like to acknowledge the faculty and staff of the School of Natural Sciences at UC Merced for all their help and services. I would also like to thank the Chemistry and Chemical Biology graduate department; they maintained an active research atmosphere on campus and kept the bureaucratic formalities simple.

Table of contents

LIST OF ABBREVIATIONS ---- XIV

CURRICULUM VITA ---- XV

ABSTRACT ---- XVIII

1 INTRODUCTION ---- 1

1.1 PROTEINS ---- 1

1.2 THE PROTEIN-FOLDING PROBLEM ---- 3

1.3 PROTEIN FOLDING ENERGY LANDSCAPE ---- 5

1.3.1 Landscape topography: Frustrated and minimally frustrated energy landscape. ---- 6

1.3.2 Projecting folding energy landscape into a one-dimensional free energy surface ---- 8

1.4 RATE THEORY FOR PROTEIN FOLDING ---- 10

1.5 TRANSITION PATH TIME ---- 12

1.6 FAST FOLDING PROTEINS ---- 14

1.7 EXPERIMENTAL INVESTIGATION OF FAST FOLDING MECHANISM ---- 16

1.8 SINGLE-MOLECULE EXPERIMENTAL TECHNIQUES ---- 20

1.9 RESEARCH OBJECTIVES AND DISSERTATION OUTLINE ---- 21

2 RESOLVING FOLDING DYNAMICS FROM SM-FRET EXPERIMENTS: METHODS AND ANALYSIS ---- 22

2.1 SINGLE-MOLECULE FRET ---- 22

2.1.1 Förster resonance energy transfer (FRET) ---- 23

2.1.2 Single-molecule FRET: Instrumentation ---- 25

2.1.3 Fluorescence labelling ---- 27

2.1.4 Histogram analysis of SM-FRET data ---- 28

2.1.5 Data processing of photon trajectories from free-diffusion SM-FRET experiments ---- 29

2.1.5.1 Identification of photon bursts ---- 29

2.1.5.2 Removal of bursts with an inactive acceptor fluorophore ---- 30

2.2 GOPICH-SZABO MAXIMUM LIKELIHOOD ANALYSIS (GS-MLA) METHOD ---- 30

2.3 ESTIMATING AVERAGE TRANSITION PATH TIME FROM LIKELIHOOD DIFFERENCE ---- 32

2.4 FREE ENERGY SURFACE-BASED MODELS ---- 34

2.4.1 1-dimensional free energy surface (1D-FES) ---- 34

2.4.2 Landau free energy surface model ---- 36

2.4.3 Folding kinetics along the free energy surface: Rate matrix formalism ---- 37

3 DECONSTRUCTING THE TERMS IN THE FOLDING RATE EQUATION AND RESOLVING (UN)FOLDING TRANSITION PATH OF ENGRAILED-HD USING HIGH-RESOLUTION SINGLE-MOLECULE FÖRSTER RESONANCE ENERGY TRANSFER (SM-FRET) TECHNIQUES ---- 39

3.1 Engrailed Homeodomain ---- 39
3.2 Sample preparation and SM-data collection ---- 40
3.3 Thermodynamic stability check after labelling ---- 41
3.4 Confocal volume characterization and power dependence ---- 43
3.5 Burst identification and binned FRET efficiency histograms ---- 44
3.6 MLA with the two-state model for folding ---- 46
3.7 MLA with the 1D-FES model ---- 47
3.7.1 Comparison of the 1D-FES model and Landau free energy surface model---- 48
3.7.2 Probing the coordinate dependence of diffusion coefficient in protein folding ---- 49
3.7.3 Resolving transition path times.---- 50
3.8 Concluding remarks ---- 54
4 PROTEIN FOLDING DYNAMICS AS DIFFUSION ON A FREE ENERGY SURFACE: RATE EQUATION TERMS, REACTIVE TRANSITION PATHS AND ANALYSIS OF SINGLE-MOLECULE PHOTON TRAJECTORIES---- 55
4.1 Stochastic simulations of folding dynamics on Landau free energy surface ---- 56
4.1.1 Landau one-dimensional free energy surface ---- 56
4.1.2 Stochastic kinetic simulations. ---- 58
4.2 Analysis of transition paths from stochastic kinetic trajectories.---- 61
4.3 Simulating time-stamped photon trajectories. ---- 62
4.4 Direct measurement of transition path times and commitment probabilities ---- 63
4.4.1 Effects of the free energy barrier on the transition paths---- 65
4.4.2 Asymmetry of the free energy surface and transition paths ---- 67
4.4.3 Comparing different methods of transition path time calculations. ---- 69
4.5 MLA with a one-dimensional FES model ---- 73
4.6 Conclusion ---- 76
5 INVESTIGATING THE ROLE OF TOPOLOGY AND SECONDARY STRUCTURAL ELEMENTS IN FOLDING DYNAMICS ---- 78
5.1 Introduction---- 78
5.2 Results and Discussion---- 80
5.3 Conclusion ---- 89
6 EFFECTS OF DENATURANTS AND IONIC STRENGTH ON THE FOLDING DYNAMICS OF ENGRAILED-HD. ---- 90
6.1 Introduction---- 90
6.2 Results and Discussion---- 92
6.3 Conclusion ---- 96
BIBLIOGRAPHY ---- 97

7 APPENDIX ---------- 113

7.1 SINGLE MOLECULE FRET MICROSCOPE: OPTICAL SET-UP ---------- 113

7.2 PROPERTIES OF THE FLUORESCENT DYES ---------- 117

7.3 TRANSFORMATION OF CELLS WITH RECOMBINANT PLASMIDS ---------- 118

7.4 DOUBLE CYSTEINE LABELLING OF EN-HD ---------- 120

7.5 SM-FRET DATA ON EN-HD AT DIFFERENT TEMPERATURES ---------- 121

7.6 BULK CHARACTERIZATION OF WW DOMAINS ---------- 123

List of figures

Figure 1-1 : Protein folding inside the cell. A newly synthesized chain is bound to chaperons during translation and released on completion. Unfolded protein (B) undergoes reversible folding to the functional native structure (C). Folded protein further forms larger functional assemblies (D). Activity of protein cell is by maintained proteosomes as they degrade unfolded protein, hence this process is dependent on the folding-unfolding equilibrium. Transient formation of misfolded protein (E) leads to aggregation (F). From [Munoz and Cerminara When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches]. Figure reprinted with permission. 3

Figure 1-2 : Schematic representation of funnelled energy landscape proposed by energy landscape theory. Landscape shows multiple high energy unstructured conformations and few low energy structured conformations. Folding occurs via multiple microscopic trajectories. From [Dill and MacCallum The Protein-Folding Problem, 50 Years On]. Figure reprinted with permission from AAAS 6

Figure 1-3 : Examples of folding energy landscapes. Left : A minimally frustrated landscape and has fewer deep valleys or high hills. Right : A highly frustrated landscape with multiple deep valleys separated by high hills. From [Wolynes *The Protein Folding Energy Landscape: A Primer*]. Figure reproduced with permission @Copyright RSC 7

Figure 1-4 : (A) A minimally frustrated landscape and is characterized by T_F being larger than T_G of compact non-native ensemble (B) Highly frustrated landscape characterized by a small number of discrete traps at T_G. Figure reproduced with permission from Wolynes, 2008 @Copyright RSC 7

Figure 1-5 : Protein folding is described as a diffusive process along a 1-dimensional FES with unfolded conformation (U) and folded conformation (F) separated by a free energy barrier. 1-dimensional FES is obtained by a projection of a multidimensional free energy landscape on to a suitable reaction coordinate (q). From [Eaton, 2015 Science)]. Figure reprinted with permission from AAAS 9

Figure 1-6 : Schematic of folding transition path for a protein with folded and unfolded state separated by a free energy barrier. (A) Transition path is the part of the folding trajectory that crosses the reaction coordinate x at x_0 and reaches x_1 without recrossing to x_0. (B) Transition path time appears as a sudden jump between the states on a FRET efficiency trajectory when used with a two-state description of folding kinetics. From [Chung, 2012 Science)]. Figure reprinted with permission from AAAS 13

Figure 1-7 : A typical chevron plot for a two-state folding protein. Each point on the chevron is obtained from the relaxation rates for folding (data to the left of C_m) and refolding (data to the right of C_m) transient mixing experiments. 19

Figure 2-1 : Principles of fluorescence and FRET explained through Jablonski diagrams (A) Fluorescence: Absorption of a photon by the fluorophore excites the ground state electron to a higher energy state. Electron relaxes back to the lowest vibrational level within this excited state through non-radiational vibrational relaxation. Relaxation back to the electronic ground state takes place via photon emission leading to

fluorescence (B) FRET : Excited state energy of the donor is taken up by a suitable acceptor in close proximity through non-radiational dipole-dipole coupling, leading to the excitation of acceptor electrons and further emission of a photon by the acceptor (C) Dependence of transfer efficiency on distance between acceptor-donor 23

Figure 2-2 :Schematic of the two-colour single-molecule FRET instrumentation set-up 26

Figure 2-3: SM-FRET data analysis and generation of FRET efficiency histograms (FEH) 28

Figure 2-4 : (A) Landau free energy surface as a function of the order parameter *x*. (B) Kinetic amplitudes of the three slowest non-zero eigenvalues obtained from the rate matrix 38

Figure 3-1 : Ribbon representation of En-HD structure 39

Figure 3-2 : Resolving folding dynamics from SM-FRET photon trajectories using ML analysis with 1D-FES model A) Fast folding protein labelled with FRET pair, and high count-rate free diffusion SM-FRET experiments performed on a confocal fluorescence microscope. B) Photon bursts are identified (0.2 ms binning) and then extended to include the entire diffusive trajectory through the confocal volume (section 2.1.5.1) C) Time-stamped photon trajectories are subjected to maximum likelihood analysis (MLA) and the 1D-FES model D) The FES (barrier, curvatures of wells and barrier top) and diffusion coefficient are determined by subsequent MLA of photon trajectories with a diffusive 1D free energy surface model. E) Photon trajectories are converted to most likely molecular state trajectories using the Viterbi algorithm. Transition paths are resolved if they are longer than the average inter-photon time (~4 μs). Figure reproduced from preprint (A. Sengupta, N. Mothi, M. Sadqi and V. Munoz 2021) 40

Figure 3-3: Chemical unfolding curves for En-HD. Experimental data for labelled (blue) is monitored using change in FRET and unlabelled (red) is monitored using change in intrinsic tryptophan fluorescence. C_m=3.3 M was obtained from the global fit (solid lines) of the experimental data (filled circles). 42

Figure 3-4 : (A) Dilution series of Atto-488 in water. Concentration of Atto-488 spans the concentration range between 40 nM to 25 pM. (B) Saturation curve of count rate as a function of laser power measured for 20 pM doubly labelled protein sample. 44

Figure 3-6 : Properties of bursts identified (A) Average FRET efficiency distribution for the bursts (B) Distribution of residence time, the effective observation time in the confocal volume in free diffusion SM-FRET experiments (C) Distribution of count rate 45

Figure 3-5 : FEH for binned SM-FRET data of En-HD in 3.3 M Urea (C_m) (A) 1 ms binning with photon threshold, N_T =160 photons (B) 150 μs binning, N_T =50 photons. Red curve shows the distribution around mean FRET if the source of broadening was solely shot noise. Broadening beyond the red curve suggests dynamic exchange between different populations under the binning time. 45

Figure 3-7 : Dwell time distribution in folded state and unfolded state obtained from the state trajectories generated using parameters from 2-state model. 46

Figure 3-8 : Free energy surface and probability distribution from MLA with 1D-FES model. (A) FES obtained as a function of reaction coordinate, nativeness and *E*. Blue and magenta vertical lines indicate *E* of states identified by two-state and three-state model (B) Probability distribution from the 1D-FES model overlayed on 150 μs binning FEH. 47

Figure 3-9 : Free energy surface obtained from the 1D-FES (blue) and the Landau FES model (red) for En-HD and gpW. 48

Figure 3-10 : Examples of state trajectories generated from the SM photon trajectories using Viterbi procedure according to different models. State trajectory corresponding to two-state, three-state and 1D-FES are represented by blue, magenta and black curve respectively. Green and red circles denote donor and acceptor photons and are overlayed on the 1D-FES state trajectory. Left y-axis represents the microstates corresponding to 1D-FES and right y-axis shows FRET efficiency. Shaded region shows the transition region defined for the molecular trajectory from 1D-FES model for measuring transition path time. Note that the duration of bursts follows an exponential distribution around a mean of 601 μs, hence a few long trajectories (e.g., G) are also observed. 51

Figure 3-11 : Distribution of transition path time for folding and unfolding transitions. Transition path time was directly measured from the state trajectories generated using Viterbi algorithm corresponding to 1D-FES. Transition region was defined as the part of the trajectory shown by the shaded region along the trajectories in Figure 3-10. 52

Figure 3-12 : Estimation of TPT from Δlog *L* curves. Δlog $L_{2si\text{-}2s}$ (magenta, left scale) and Δlog $L_{3s\text{-}2s}$ (blue, right scale) methods give TPT 10 μs and 58 μs respectively. The vertical dashed line signals the mean TPT obtained from the distribution in Figure 3-11. 53

Figure 4-1 : The Landau 1D free energy surfaces used in this study. Left) 1D free energy surfaces corresponding to the two extremes of the range we used to investigate the relationship between barrier height and transition paths. The light blue surface represents the highest barrier we explored (10 k_BT), and the navy blue the lowest (0.4 k_BT). All the free energy surfaces are generated with $\phi = 0.65$ (same shape). Right) 1D free energy surfaces corresponding to the two extremes of the range we used to investigate the effects of the position of the barrier along the reaction coordinate. The light blue represents a surface with $\phi = 0.5$ (symmetric) and the navy blue a surface with $\phi = 0.9$ (highest asymmetry). 57

Figure 4-2 : Schematic of stochastic simulations of folding dynamics on Landau FES (A) Landau FES is generated (B) Molecular trajectory for free diffusive protein folding on the FES is generated (C) Photon trajectories corresponding to the molecular trajectory is generated from FRET efficiency and an exponential distribution of interphoton times 58

Figure 4-3 : Examples of molecular trajectories generated. 20 ms segment of the full-length trajectory (20 s) is shown for (A) 1 kJ/mol (B) 3 kJ/mol (C) 6 kJ/mol 60

Figure 4-4 : Boundaries for the transition region. The transition region is defined by its boundaries (x_0 and x_1) on the reaction coordinate. Here five examples are given in which the position of x_0 and x_1 are determined

as a fraction of the distance between the minimum and the barrier top (x = 0). The halfway definition of the transition region is highlighted as a green swath with a schematic representation of a possible transition path as example. The diagram shows the reaction coordinate (x) on the bottom, and the corresponding changes in FRET efficiency for the simulations of photon trajectories on top _____ 61

Figure 4-5 : TPT (A) and commitment probabilities (B) for different definitions of transition region following the colour code in Figure 4-4 as a function of free energy barrier. Dependence of the TPT (C) and commitment probability (D) on the boundaries for the transition region for barrier of 1.2 k_BT. The dashed horizontal line indicates the inverse of the second slowest non-zero eigenvalue of the rate matrix ($1\lambda 2$). _____ 64

Figure 4-6 : A) Transition path time (black, left axis) and the curvature for free energy barrier (green, right axis) as a function of free energy barrier. The barrier curvature is shown normalized with respect to the highest barrier used in this study. B) Barrier dependence of the average transition path time multiplied by the barrier curvature_____ 66

Figure 4-7 : Transition paths as a function of the asymmetry of the free energy surface (ϕ). (A) average transition path time in the folding direction ($U \rightarrow F$), showing the climbing (navy) and descent (cyan) segments of the path. (B) As in top panel but for transitions in the unfolding direction ($F \rightarrow U$). (C) likelihood per photon (multiplied by 10^4) obtained from the MLA of the photon trajectories simulated on a ϕ = 0.65 surface using a Landau free energy surface with fixed asymmetry. _____ 68

Figure 4-8 : $\Delta\log L$ estimates of the experimental (not normalized) average TPT obtained from photon trajectory simulations on free energy surfaces with varying barrier. Results for simulations with increasing barrier are shown from top left to bottom right with the barrier height indicated in k_BT: $\Delta\log L_{2si\text{-}2s}$ (magenta, left scale) and $\Delta\log L_{3s\text{-}2s}$ (blue, right scale) methods. The vertical dashed line signals the average TPT directly measured from the stochastic kinetic simulations using 1/3 boundaries for the transition region. N_T is the total number of million photons used in the analysis _____ 71

Figure 4-9 : TPT as a function of the free energy surface barrier (normalized as before). TPT directly measured is shown in black (dashed line Figure 4-8). TPT from $\Delta\log L_{2si\text{-}2s}$ and $\Delta\log L_{3s\text{-}2s}$ are shown in magenta and blue, respectively (τ_s corresponding to the peaks in Figure 4-8) _____ 72

Figure 4-10 : Free energy surfaces corresponding to different folding scenarios showing E for folded state (E_F) and unfolded state (E_U) obtained from two-state (magenta lines) and E_F, E_U and E_S from three-state kinetic models (blue vertical lines). FES in light and dark green shows the original FES used for the simulation and the FES retrieved from MLA analysis of the simulated photon trajectories respectively _ 74

Figure 5-1 : Cartoon representation of the fluorophore labelled in the folded structure of three fast-folding domains. The mesh representation shows the 3D anisotropic orientations of the fluorophores indicating the site-specific labelling. The proteins are labelled at their flexible C-terminal and N-terminal ends with Alexa Fluor-488 and Alexa Fluor-594 as FRET pair, as shown in the figure. Figure also indicates the calculated average end to end distance for the FRET pair ($R_0 \sim 5$ nm). FBP11-WW1 and Nedd4-WW4 shares the same structure. _____ 79

Figure 5-2 : Chemical denaturation at SM conditions for Nedd4-WW4. Histograms show the distribution of *E* for the bursts identified using algorithm described earlier for different concentration of GdnCl (mean of the duration of burst is ~0.6 ms). Red and green vertical lines (0.69 and 0.90) indicate the shift of mean *E* for the lowest and highest concentration used. FEH for 1.4 M GdnCl (~C_m) is further analysed with different binning times. 82

Figure 5-3 : Chemical denaturation at SM conditions for FBP11-WW1. Histograms shows the distribution of *E* for the bursts identified using algorithm described earlier for different concentration of GdnCl (mean of the duration of burst is ~0.6 ms). Red and green vertical lines (0.72 and 0.90) indicate the shift of mean *E* for the lowest and highest concentration used. FEH for 2.4 M GdnCl (~C_m) is further analysed with different binning times. 83

Figure 5-4: Chevron for Nedd4-WW4 obtained from 2-state fit to single-molecule data. Blue and red circles show unfolding and folding rates directly obtained from the fit. Green curve shows sum of the fits to the rates 84

Figure 5-5 : FEH for all four proteins at different binning time (T_b) at ~C_m. Photon threshold of 150 photons and 40 photons were used for 1 ms (left column) and 0.15 ms (right column) respectively. Red curve shows the shot noise limited distribution around mean *E*. Broadening suggests dynamics happening within the timescale of binning. 85

Figure 5-6: Free energy surface obtained from MLA analysis with the 1D-FES model at C_m for the four proteins under study 87

Figure 5-7 : Rate analysis for Nedd4-WW4 (A) Kinetic amplitudes of the three slowest non-zero eigenvalues obtained from the rate matrix (B) Relaxation kinetics simulated with the rates from the matrix **K** 89

Figure 6-1 : FEH for bursts collected for En-HD at midpoint denaturation condition (C_m) in GdnCl and urea. Mean residence time for the bursts shown here is ~0.6 ms 92

Figure 6-2: FES obtained from MLA with 1D-FES model for En-HD at midpoint denaturation concentration (A) GdnCl (B) Urea 93

Figure 6-3: SM-FRET analysis at 7 M urea (~C_m) in presence of 2.3 M NaCl (A) FEH for the bursts collected (B) FES obtained from 1D-FES model in 7 M urea + 2.3 NaCl (red). Blue curve shows the FES at urea C_m discussed earlier (3.3 M urea) for comparison 95

Figure 7-1: Schematic of 2-colour SM-FRET optical set-up 113

Figure 7-2 : Spectral properties and chemical structure of the organic fluorophores used as FRET acceptor-donor pair. Excitation spectra (blue) and emission spectra (red) of dyes are shown (As reported by Thermo Fischer). 117

Figure 7-3 : FEH for bursts collected at urea midpoint at different temperatures for En-HD. The rates from MLA analysis with a two-state model are also shown. The dynamics at higher temperature was found to be too high to be resolved by the 1D-FES with the rate matrix formalism we employ. _____ 121

Figure 7-4 : 7-5 FEH for GdnCl midpoint condition at different temperature. Ionic strength was maintained at 2.3 M by adding NaCl for the decrease in the C_m for higher temperature. _____ 122

Figure 7-6 : Chemical denaturation curves for FBP11-WW1 and Ned4-WW4 monitored using steady state tryptophan fluorescence and tryptophan lifetime. Denaturation midpoint of 1.6 and 2.4 M found from global fit to two-state _____ 123

Figure 7-7: Chemical denaturation curves of FBP11-WW1and Ned44-WW4 at different conditions___ 124

List of abbreviations

1. FES : Free energy surface
2. 1D-FES : One-dimensional free energy surface
3. SM-FRET : Single-molecule Förster resonance energy transfer
4. MLA : Maximum likelihood analysis
5. U and N : Unfolded conformation and folded conformation
6. TPT : Transition path time
7. C_m : Denaturation midpoint concentration
8. FEH : FRET efficiency histograms
9. GdnCl : Guanidine hydrochloride
10. R : Universal gas constant
11. k_B : Boltzmann constant

Curriculum Vita

NIVIN MOTHI

RESEARCH INTEREST

Single-molecule fluorescence (FRET) spectroscopy, Optical spectroscopy, Protein folding, Biophysics, Protein engineering, Molecular Biology

EDUCATION

Ph. D Candidate (2021), Chemistry and Chemical Biology at University of California Merced, California

Integrated M. Sc. (2015), Chemistry at Centre for Excellence in Basic Sciences – Mumbai, India

RESEARCH EXPERIENCE

- **Graduate researcher** (2015 - current) | University of California, Merced

  Investigated protein folding dynamics and resolved folding transition path times using single-molecule fluorescence techniques

- **Master's research** (2014-2015) | UM-DAE Centre for Excellence in Basic Sciences, Mumbai, India

  Investigated the amorphous aggregation of human serum albumin and devised strategies to prevent it.

  Characterized proteins and aggregates using fluorescence spectroscopy, light scattering techniques, circular dichroism, and various chemical assays.

- **Research intern** (2014) | National Centre for Biological Sciences, Bangalore, India

  Studied the role of the C-terminal region of prion proteins in aggregation by monitoring aggregation of various disease-related mutants.

- **Research intern** (2013) | Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India

  Synthesized a photochromic spiropyran derivative which undergoes reversible isomerization on UV-exposure with application in artificial light-harvesting assemblies.

- Research intern (2012) | Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India

  Worked on synthesizing antimicrobial peptides and developed a cheap synthetic route to the monomers. Gained experience in organic chemistry synthesis techniques

HIGHLIGHTS & FELLOWSHIPS

- UC Merced graduate dean's dissertation fellowship 2020
- Selected as NSF-CREST Center for Cellular and Biomolecular Machines Scholar (2017-20)
- Center for Cellular and Biomolecular Machines travel fellowship for 2019 & 2020
- Center for Cellular and Biomolecular Machines -CREST training award for 2019
- Chemistry and Chemical Biology summer fellowship for 2016-2017
- Diploma in Project Oriented Chemistry Education (2010-13) at Jawaharlal Nehru Centre for Advanced Scientific Research (worked on different research projects and gave scientific presentations over a span of 3 years)
- Innovation in Science Pursuit for Inspired Research (2010-15) Fellow by Department of Science and technology (selected for a fully funded bachelor's and master's program)

PUBLICATIONS

1. Mothi, N., & Muñoz, V. (2021). Protein folding dynamics as diffusion on a free energy surface: rate equation terms, transition paths and analysis of single-molecule photon trajectories. *The journal of physical chemistry. B* (submitted)

2. Castellanos, M., Mothi, N., & Muñoz, V. (2020). Eukaryotic transcription factors can track and control their target genes using DNA antennas. *Nature communications*, *11*(1), 540

3. Jain, A., Achari, A., Mothi, N., Eswaramoorthy, M., & George, S. J. (2015). Shining light on clay-chromophore hybrids: layered templates for accelerated ring closure photo-oxidation. *Chemical science*, *6*(11), 6334–6340

4. Mothi, N., Muthu, S. A., Kale, A., & Ahmad, B. (2015). Curcumin promotes fibril formation in F isomer of human serum albumin via amorphous aggregation. *Biophysical chemistry*, *207*, 30–39.

5. Muthu, S. A., Mothi, N., Shiriskar, S. M., Pissurlenkar, R. R., Kumar, A., & Ahmad, B. (2016). Physical basis for the ofloxacin-induced acceleration of lysozyme aggregation and polymorphism in amyloid fibrils. *Archives of biochemistry and biophysics*, *592*, 10–19.

ABSTRACT

Protein folding as diffusion on a free energy surface: Rates and mechanisms from advanced single-molecule fluorescence techniques

by

Nivin Mothi

Doctor of Philosophy in Chemistry and Chemical Biology

University of California, Merced, 2021

Professor Victor Munoz, Graduate Advisor

Proteins act as cellular nanomachines by carrying out a wide variety of functions, enabling life as we know it today. They exist in a dynamic equilibrium between various structural conformations ranging from their functional native structure to the nascent unfolded chain. The diversity of the protein structures qualifies proteins for carrying out diverse functions. Protein conformational dynamics play a key role in functional control, as it is the dynamic equilibrium between different conformations under cellular conditions that enables the regulation of functionalities.

Single-molecule FRET experiments combined with maximum likelihood analysis (MLA) offers a unique opportunity for experimentalists to determine various parameters involved in the folding dynamics which were previously accessible only through simulations. All details about conformational dynamics happening in the protein within the interphoton time is embedded in the single-molecule photon trajectory and the MLA method has the potential to extract these details with the limitation being the model being used to describe folding. Traditional models used to describe protein folding are simple kinetic models with inherent assumptions about the nature of the folding. While the simplicity of these models makes them an attractive approach, they come with serious limitations in resolving the dynamics. Here, we demonstrate a free energy surface-based approach to extract and dissect various dynamic processes happening along protein folding reactions.

Everything that living things do can be understood in terms of the jiggling and wiggling of atoms." – Richard Feynman

Through this dissertation, we are trying to gain a better understanding of the jiggling and wiggling of protein chain as it (un)folds along a one-dimensional free energy surface.

Chapter 1

1 Introduction

1.1 Proteins

Biological processes that sustain life as we know of today are possible because of the calculated and precise interactions between a group of biomolecules. These biomolecules act together in a concerted manner and exist in a dynamic equilibrium under cellular conditions, responding to stimuli to maintain life processes. Proteins are one of the most diverse and ubiquitous biomolecules. They are the primary building blocks of life, and they take part in performing nearly every task required for perpetuating cellular life. While nucleic acids are the key to storing and passing down genetic information, it is proteins that enable this process and act as the workforce that carries out cellular processes. Millions of years of evolution have equipped proteins to carry out various cellular functions like cell signalling, enzymatic activity, structural integrity, cellular storage, cellular transport, etc. The diverse array of highly specific three-dimensional structures that proteins can fold into on-demand enables them to carry out these multitudes of functions.

Proteins are synthesized as linear unstructured heteropolymers made of individual amino acid units at the site of protein synthesis, the ribosome machinery. The unfolded protein chain then self-assembles into a three-dimensional structure capable of performing cellular functions with high selectivity and specificity. A folded protein has an intricate network of surface interactions and can selectively participate in specific interactions leading to binding with analytes or can assemble into larger macromolecular assemblies leading to the functionality of proteins. This self-assembly of the linear heteropolymer into a specific compact three-dimensional conformation with distinctive secondary and tertiary structural elements and a specific overall topology is called protein folding. The three-dimensional structure a protein folds into is determined by its amino acid sequence, which is determined by the genes encoding them. Thus, the blueprint for the folding of protein to the functional conformation with precision and fidelity is embedded in the amino acid sequence through the genetic code itself.

The functionality of a protein is strictly correlated to the structure of the protein. Hence the failure to correctly fold into the functional conformation can result in a myriad of

problems. Proteins under cellular conditions exist in a dynamic equilibrium between the unfolded conformation and the folded conformation. The conformational equilibrium between different states determines the protein's ability to carry out specific functions. They undergo conformational changes in response to stimuli associated with the functions on demand. This balance of conformational flexibility and structural integrity enables proteins to cycle between different conformations carrying out various functions under biologically relevant timescales. A thorough understanding of the mechanism of the folding-unfolding process is thus essential to harvest the functional flexibility of proteins. This can ultimately help to design and engineer synthetic protein molecules that can function as nanomachines at the cellular level.

The search to understand the mechanism of protein folding, 'the protein folding problem', emerged along with the development of protein structural elucidation using X-ray analysis. Structural elucidation of Myoglobin was a turning point in this direction[1]. The packing of helices to form a complex globin structure and the lack of symmetry in this compact three-dimensional structure revealed through the X-ray analysis baffled scientists. This, in turn, initiated the novel field of research on protein folding. The classic Levinthal's paradox, a heteropolymer like protein, can populate an astronomical number of conformations, and searching for the native conformation randomly cannot be done in biologically relevant timescales, is now a well-understood question. Decades of research in the field and development of computational techniques and the advancements in experiential techniques have provided enormous information about the mechanism of protein folding[2–4].

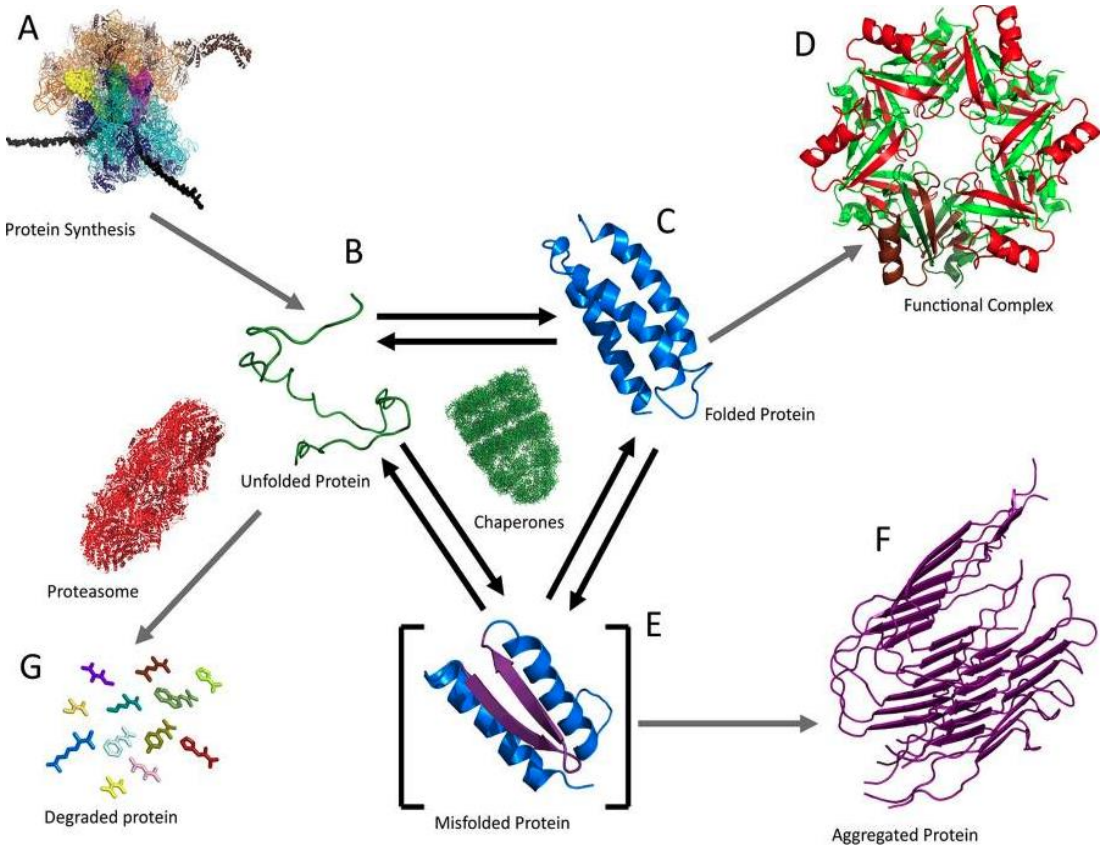


Figure 1-1 : Protein folding inside the cell. A newly synthesized chain is bound to chaperons during translation and released on completion. Unfolded protein (B) undergoes reversible folding to the functional native structure (C). Folded protein further forms larger functional assemblies (D). Activity of protein cell is by maintained proteosomes as they degrade unfolded protein, hence this process is dependent on the folding-unfolding equilibrium. Transient formation of misfolded protein (E) leads to aggregation (F). From [Munoz and Cerminara When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches]. Figure reprinted with permission.

1.2 The Protein-Folding Problem

The protein folding problem, which was once regarded as a grand challenge, is composed of three closely related problems (1) The thermodynamic folding code: What is the physical code that drives a linear sequence of amino acids to fold into a specific dimensional native structure (2) The folding mechanism: How does a polypeptide chain that has an astronomical number of possible conformations search to find the native protein structure quickly? (3) Structure prediction: Can we computationally predict the structure of proteins from their amino acid sequence with high accuracy?[3–6]

Half a century after being posed as an insurmountable challenge, the protein folding field has now progressed, and our understanding of the process has advanced significantly. Designing of foldable proteins and nonbiological polymers is routinely done in protein chemistry labs, and such designed polymers are successfully being employed in various biological applications. Applying statistical thermodynamic principles from polymer physics and glass transitions have provided detailed information on the funnelled energy landscape for protein folding, essentially reducing the search problem into gradual downhill steps along a funnel-shaped energy landscape[7,8]. All atom-molecular dynamics simulations carried out on a special-purpose machine demonstrated the folding-unfolding transition of small proteins with atomic resolution [9–11]. Fully automated structure prediction tools like ROSETTA and HHPred have made significant contributions in structure prediction[12,13]. DeepMind's artificial intelligence program AlphaFold2 has made a gigantic leap in protein structure prediction during CASP-2020 (a community-wide experiment where researchers from all over the world try to predict the unknown structure of proteins), improving significantly over the performance of its own predecessor Alphafold in CASP-2018[14,15].

While the developments in machine learning algorithms and structure prediction tools have brought in a paradigm shift in the research focused on structure prediction, studies aimed at resolving protein folding dynamics are still extremely relevant. Proteins are highly dynamic molecules and undergo structural fluctuations between different conformations at physiological conditions. The conformational flexibility of proteins is functionally and biologically relevant. Understanding the evolution of protein sequences to achieve the level of plasticity required to carry out specific functions is of great interest. The timescale of protein dynamics is interlinked with the control and regulation of protein functions inside cells.

The discovery of functionally relevant unstructured proteins in nature has raised another intriguing question in protein science. These functional proteins without unique structures are called intrinsically disordered proteins (IDPs) and are found in abundance in nature. Recognition of IDPs as a widely prevalent occurrence in proteome, rather than an exception raised a new ‘protein non-folding problem’[16–18]. IDPs demonstrate remarkable conformational flexibility and structural plasticity and stand out from the traditional description of protein folding and functional principles. Many IDPs, despite their lack of fixed structure, are promiscuous binders that form complexes, and invariably engage with partners via many binding situations, producing static, semi-static, fuzzy, or dynamic complexes[19]. Obtaining a comprehensive understanding of IDPs' mechanism of action and functions is one of the most interesting problems in protein science today. Interpretation of the mechanism of folding and other conformational dynamics on the energy landscape holds the key to designing flexible functional biomolecules.

1.3 Protein folding energy landscape

Proteins are complex polymers with large degrees of motion and can populate an astronomical number of conformations of varying stability. The big question in the protein folding field was how proteins search through this vast conformational space to reach the thermodynamically stable native structure. Simple chemical reactions are explained by a mechanism where a reactant molecule converts to a product molecule through a series of intermediates following a fixed pathway. Proteins, however, occupy a vast ensemble of unstructured conformations, and hence a protein folding pathway starting from a single microscopic structure won't describe the process in its entirety. The protein folding process needs to be described as a conformational transition from a structurally disordered ensemble to a structured ensemble.

Energy landscape theory takes a statistical mechanical approach to describe protein folding by mapping the free energy averaged over solvent coordinates of all conformations populated by protein and thus provides the shape and nature of the energy landscape[7,20,21]. The key challenge was obtaining the density of states (DOS) which is the number of different possible conformations at each energy level that can directly yield the conformational entropy. DOS and the conformational entropy are obtained by employing simplified treatments like spin-glass transition theories and polymer models, and minimalistic lattice models[22,23]. The outcome from this statistical approach to the thermodynamics of protein folding is a funnel-shaped energy landscape with very few low energy conformations, and much more high energy unfolded conformations.

Protein folding energy landscape provides a distribution of conformations along the free energy axis on a multidimensional conformational space. Unfolded state of the proteins is devoid of any significant defined structural features and has very few stabilizing native-like interactions. The conformational change from the unfolded state to the compact tertiary structure occurs through formation of more stabilizing interactions between the amino acid residues. Native-like interactions are, on average, more stable than non-native interactions. Hence, proteins can find the energetically stable structure by a stochastic search of conformations accessible to the polypeptide chain. Non-native interactions formed along the conformational transition are counterproductive to the folding process and adds to frustration on the energy landscape. As protein forms more native-like interactions, the number of conformational states of the same energy that can be sampled decreases. This is represented by the funnel shape of the energy landscape[7,20]. An unstructured ensemble occupies the top of the funnel with the radial axis representing the number of states of the same energy (conformational entropy). The energy landscape gradually narrows to take a funnel shape as more structured conformations are formed en route to get to the final native conformation. This essentially reduces the protein search problem to be comprised of random incrementally downhill steps towards the thermodynamically stable folded state[24].

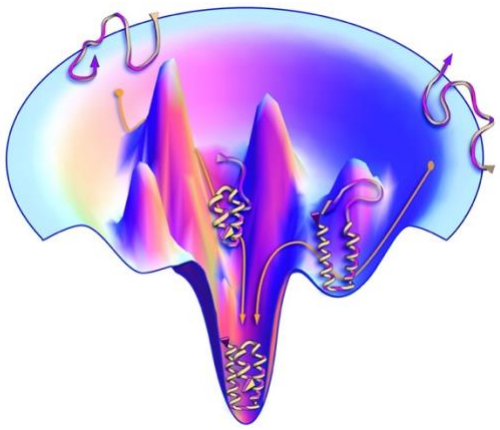


Figure 1-2 : Schematic representation of funnelled energy landscape proposed by energy landscape theory. Landscape shows multiple high energy unstructured conformations and few low energy structured conformations. Folding occurs via multiple microscopic trajectories. From [Dill and MacCallum The Protein-Folding Problem, 50 Years On]. Figure reprinted with permission from AAAS

The cartoon representations of funnelled energy landscape have been a popular way of explaining the protein folding energetics. However, it can be deceiving and misinterpreted if not explained carefully. Folding along the funnelled landscape appears thermodynamically downhill at all instances and the origin of the free energy barrier is not evident from the representation. It is important to emphasize that the vertical axis is the contribution of stabilization energy to the free energy and the radial axis represents the conformational entropy. There is a loss in conformational entropy as we move down the funnel which is manifested as an unfavourable term in the free energy. Along with this, there is a gain in the stabilization energy due to formation of native contacts. The free energy for folding is made up of contributions from both the conformational entropy and stabilization energy. It is this delicate balance between the conformational entropy and the stabilization energy that determines the course of protein folding. Asynchronous compensation of loss of conformational entropy by the stabilization energy gives rise to a free energy barrier for folding, while a perfect compensation for the conformational entropy by stabilization energy results in barrier-less or downhill folding scenario[25,26].

1.3.1 Landscape topography: Frustrated and minimally frustrated energy landscape.

Energy landscape theory makes extensive comparison of the landscape for a random heteropolymer and naturally evolved proteins[27]. A protein with highly optimized native interactions has a smooth downhill slope for the interaction energy along the funnel,

while proteins optimized for non-native interactions to facilitate function have a frustrated energy landscape[27]. Proteins with smooth landscapes have fewer deep valleys or high hills. Their phase space is broadly divided into two main parts, high energy high entropy unstructured state and low energy low entropic structured state resulting in single exponential folding kinetics behaviour on longer time scales and the coincidence of transitions measured through various experimental probes. Processes like helix-coil transition and simple crystallization of materials show this kind of landscape.

Proteins with rough energy landscapes show distinct kinetic and thermodynamic properties from smoother landscapes and are characterized by the presence of many deep valleys separated by high energy barriers. They show a large number of low energy structures when coarse-grained. The roughness of the landscape is a result of many competing interactions involved in the interaction energy and is termed as energetic frustration. As multiple conflicting interactions add up to the frustration, folding into one global minimum will involve finding the optimal orientation for the polypeptide chain to reduce these conflicting interactions. A completely random polypeptide chain, not selected to fold into a specific 3-dimensional structure either by an algorithm or nature, cannot access its global free energy minima at a reasonable time and will fail to fold into the native fold. Experiments on proteins designed with random amino acid sequence have shown that folding into multiple folded structures are possible and reliable folding into global free energy minima is not guaranteed[28,29].

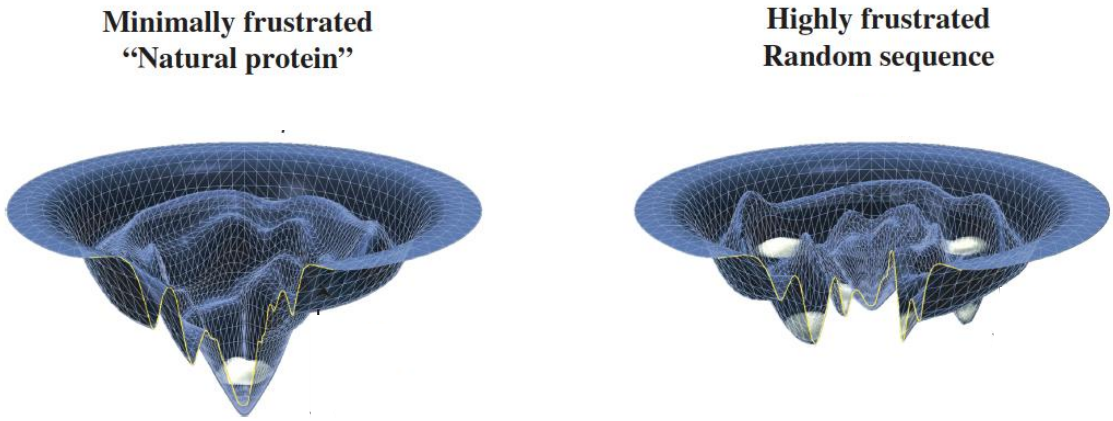


Figure 1-3 : Examples of folding energy landscapes. Left : A minimally frustrated landscape and has fewer deep valleys or high hills. Right : A highly frustrated landscape with multiple deep valleys separated by high hills. From [Wolynes *The Protein Folding Energy Landscape: A Primer*]. Figure reproduced with permission @Copyright RSC

Energy landscape theory explains the reliable folding of natural proteins to their native fold by the principle of minimal frustration. The smoothness of folding landscapes in naturally occurring proteins is attributed to the evolution of selecting protein sequences

to fold in a way to facilitate its function. Proteins with minimal frustration are selected to fold quickly, as seen in β-barrel proteins and leucine zippers, where the sidechains make complementing interactions to form symmetric super secondary structures. This is postulated by Go as a self-consistency principle that ensures compatibility between local secondary structural elements to form higher-order super secondary structures[30]. The hypothesis of minimal frustration has also been utilized to guide protein structure prediction and design new folds through a reverse engineering strategy[31]. However, it is important to note that there are naturally occurring proteins with regions of local frustration in the landscape and multiple distinct stable structures. These are attributed to the functional necessity of some proteins to allow dynamic flexibility and interaction with multiple binding partners[32,33].

1.3.2 Projecting folding energy landscape into a one-dimensional free energy surface

The protein folding process is a complex assembly process involving many degrees of motion occurring on a multidimensional free energy landscape. However, one of the real surprises here is that the protein folding can adequately be described as a diffusive process along a 1-dimensional free energy surface, given that we can select a reaction coordinate representative of the ordering of unfolded structural ensemble to the folded ensemble. The progression of folding along a funnelled energy landscape can be described by choosing one or more collective reaction coordinates that can monitor the evolution of different ensembles of partially ordered structures[34–37]. Folding along the funnelled energy landscape is a progressive formation of partially structured conformations that needs to navigate through the local minima present along the pathway as it proceeds down the funnel towards native conformation.

The free energy of the funnel is characterized by competing contributions from conformational entropy and stabilization enthalpy and the concerted change in both the terms as one proceeds along the reaction coordinate leading to the folding. Both energy and entropy decrease as protein proceeds down the funnel. The gradient of free energy thus determines the average drift along the funnel[38]. This concerted change in enthalpy and entropy as one moves down the funnel determines the average drift up or down the funnel. Folding takes place on this drift as a stochastic process which to first approximation can be defined as a diffusive process, making the reduction of dimensionality of FES possible[34–36].

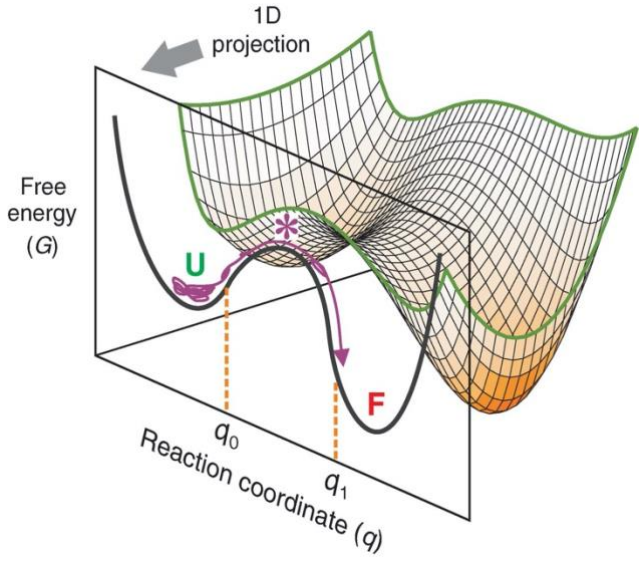


Figure 1-5 : Protein folding is described as a diffusive process along a 1-dimensional FES with unfolded conformation (U) and folded conformation (F) separated by a free energy barrier. 1-dimensional FES is obtained by a projection of a multidimensional free energy landscape on to a suitable reaction coordinate (q). From [Eaton, 2015 Science)]. Figure reprinted with permission from AAAS

The stochastic motion of protein along this gradient is impeded by the extent of roughness and is dependent on how protein jumps between the local minima. This process of navigating along the local minima as a diffusive process determines the timescale of various dynamics along the surface. As a result, folding rates are determined by the free energy profile of the funnel and the stochastic diffusive motion along with the reaction coordinates. An exciting extension of this reduction is the ability to predict the folding speed limit, which will be diffusion-limited, on a landscape with a perfect downhill gradient[39,40]. The adequate representation of folding dynamics as a diffusive process on a lower-dimensional surface is thus an implication of energy landscape theory. This reduction of the multidimensional landscape onto a suitable reaction coordinate helps in identifying various configurations present along the reaction pathway as protein fold into the native structure. This also makes it easier to interpret experimental data from techniques like single-molecule FRET or optical tweezers or atomic force microscopy as these techniques directly measure dynamics along with a distance coordinate[41–43].

1.4 Rate theory for protein folding

The simplest approach to model folding dynamics along a 1-dimensional fee energy surface is by considering it as a chemical reaction in which the reactant molecule has to cross an activation energy barrier to convert to the product. Nearly all treatments of folding kinetics treat the rate to be dependent on the height of the energy barrier and a probability factor that accounts for the number of states that are accessible for the molecule in the ground state and the activated state[44]. Thermal energy is required to cross the energy barrier, and the rate of such a reaction is given by the Arrhenius equation.

$$k = A \exp\left({}^{-E_A}/_{RT}\right)$$

where E_A is the activation energy, and A is the preexponential term. Transition state theory (TST) which was developed for reactions of small molecules in the gas state, is one of the most influential rate theories. It assumes that the top of the free energy barrier represents a quasi-thermodynamic state (x = $x^{\ddagger}$ along the reaction coordinate) which is in equilibrium with the reactant, and any reactant molecule that crosses the energy at x = $x^{\ddagger}$ is considered to form the product. The rate of folding according to transition state theory is given by.

$$k = \kappa\, {}^{k_B T}/_{h} \exp\left({}^{-\Delta G^{\ddagger}}/_{RT}\right) = \kappa\, {}^{k_B T}/_{h} \exp\left({}^{\Delta S^{\ddagger}}/_{R}\right) \exp\left({}^{-\Delta H^{\ddagger}}/_{RT}\right)$$

where $\Delta G^{\ddagger}$ is the Gibbs free energy of activation and k_B and h is Boltzmann's constant and Plank's constant. The parameter κ is an *ad hoc* fudge factor (transmission coefficient) introduced to correct for the unreactive attempts at barrier crossing (recrossing on the trajectories)[45]. Thermodynamic activation parameters $\Delta S^{\ddagger}$ and $\Delta H^{\ddagger}$ can be calculated from the preexponential term and its temperature dependence in barrierless conditions. Transition state theory always overestimates the absolute rates owing to its simplistic treatment of a complex process, that proceeds through simultaneous breaking and making of bonds in solution. The value of ${}^{k_B T}/_{h} = 6 \times 10^{12}$ s^{-1} for T = 300 K is several orders of magnitude faster for molecular motions of a polypeptide chain which is constantly interacting with the solvent molecules. The assumption of TST that there is a "point of no return" along the reaction trajectory and the crossing of this is a "moment of decision" can be stated with precision only within classical mechanics[45].

Kramer's rate theory originally developed for condensed phase reactions is extended to better suit protein folding as these occur in solutions with simultaneous making and breaking of multiple weak interactions[46]. Rate according to Kramer's theory is given by

$$k = \left(\frac{D\beta\omega\omega^{\ddagger}}{2\pi}\right)\exp(-\beta\Delta G^{\ddagger})$$

where D is the intramolecular diffusion coefficient at the barrier top, ω^2 is the curvature of the well (e.g. folded and unfolded), $-(\omega^{\ddagger})^2$ is the inverted curvature at the barrier top, and $\Delta G^{\ddagger}$ is the free energy barrier on the projected free energy surface (FES)[47]. The terms of this rate equation contain very important mechanistic information regarding the self-assembly process. While the exponential term depends on the free energy barrier and temperature, the preexponential term depends on the frequency of the system in the ground state and on the rate of escape from the activated state (transition state) thus depending on the local mobility of the system while passing through the transition state. In aqueous solutions, this local mobility depends on the macroscopic viscosity η of the surrounding medium. Viscosity dependence of protein folding rates are often used as an indicator for dynamics following Kramer's rate theory. Folding rates are expected to be inversely proportional to the solvent viscosity when the movement of the protein chain determines the diffusion coefficient for protein dynamics through the solvent. Deviations from this trend suggest a contribution from the protein energy landscape called 'internal friction', which arises due to interactions of the polypeptide chain with itself.

Kramer's expression offers the possibility of evaluating different factors affecting protein folding rate independently. However, separating different terms of this rate expression using conventional experimental methods is not straightforward. It is done by determining protein folding rate constants at different conditions with similar stability and assigning the difference in the rates found to the preexponential factor. Polyols like glycerol are often added into the solution to increase the viscosity, but these also increase the stability of proteins. Hence such experiments are carried out by adding chemical denaturants like urea or guanidinium hydrochloride to compensate for this stability difference. Polyols are known to modulate the polarity and surface tension and hence can affect protein folding kinetics. There is also the issue of extrapolating a macroscopic property like viscosity to describe a phenomenon at the molecular level at the protein-solvent interface. It is not known with certainty whether the measured macroscopic viscosity and microscopic viscosity are linearly related over the entire range of viscogen concentration. The addition of chemical denaturants has nonlinear viscosity dependence. Viscosity correction is small when the concentration of denaturant is less than 2 M, however, correction becomes significant at higher denaturant concentrations which are required to access the unfolded state. Even though control experiments are generally done to account for the changes in the folding induced

by varying multiple solvent properties, the cumulative nonlinear effect of multiple properties makes the dependence complex.

The experimental resolution of the preexponential factor on protein folding rates is challenging as the rate pre-factor and diffusion coefficient are unknown and cannot be determined independently. Hence, the diffusion coefficient for protein folding has stayed as the most elusive term in the rate expression with very less information about how they change for different types of proteins and how they correlate with solvent effects. T-jump relaxation experiments done on a fast-folding beta-sheet protein, FiP35, with a small activation energy barrier have shown the friction controlling preexponential factor in Kramer's equation to not scale with solvent viscosity[48]. They observed an increase in the preexponential factor with temperature, and it was attributed to the roughness along the reaction coordinate, which a coordinate-dependent diffusion coefficient can model. Molecular dynamics simulations on different peptides have shown that local frustration of the protein energy landscape can give rise to apparent internal friction as the barrier crossing events on the rough surface are insensitive to the solvent viscosity[49]. It has been shown that proteins with lower folding barriers are more likely to exhibit internal friction as the folding dynamics should be sensitive to the diffusion coefficient over a wider region of the energy landscape[50]. Independent experimental determination of diffusion coefficient from experimental data is hence of utmost importance.

1.5 Transition path time

Theory suggests that protein molecules exploit a multitude of microscopic pathways along their search to get to the functional native conformation, resulting in a wide distribution of folding mechanisms. Folding time and transition path time corresponds to two key relaxation processes taking place during this conformational search and detailed inspection of these parameters together can give valuable mechanistic information about the heterogeneity of the folding pathways. The transition path is the segment of the protein folding trajectory where the successful reactant to product conversion takes place crossing the free energy barrier separating the states[47,51]. For a protein populating an unfolded state and a folded state as shown in Figure 1-6, folding transition path corresponds to the section of the molecular trajectory from unfolded to folded state for which a position closer to the unfolded well, x_0 is crossed and reaches the position x_1, close to the folded minima without re-crossing to x_0.

Transition path times appear as an instantaneous jump between the states in the molecular trajectory as it is the actual time spent on the barrier during successful folding-unfolding transitions. For a free energy surface with a parabolic barrier with $\Delta G_f^* \gg 1\ RT$, transition path time is given by,

$$t_{TP} = \frac{1}{\beta D^*(\omega^*)^2} \ln(2e^{\gamma}\beta\Delta G^*)$$

where $D^*$is the diffusion coefficient for the top of the barrier, $\beta = {}^{1}/_{k_B T}$, k_B is the Boltzmann constant, T is the temperature, $(\omega^*)^2$ is the curvatures at the barrier top and ΔG^* is the free energy barrier[51,52].

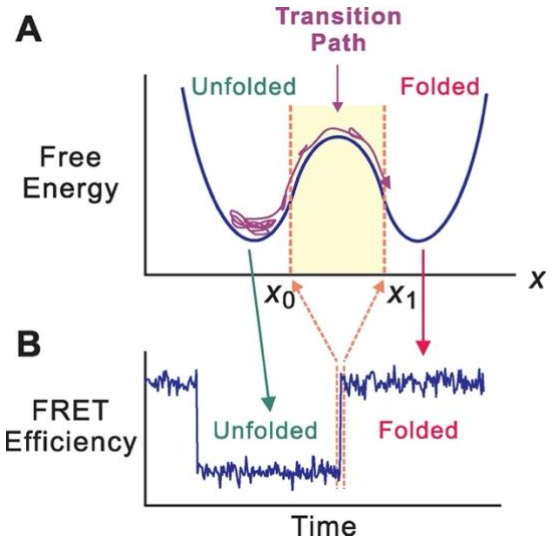


Figure 1-6 : Schematic of folding transition path for a protein with folded and unfolded state separated by a free energy barrier. (A) Transition path is the part of the folding trajectory that crosses the reaction coordinate *x* at x_0 and reaches x_1 without recrossing to x_0. (B) Transition path time appears as a sudden jump between the states on a FRET efficiency trajectory when used with a two-state description of folding kinetics. From [Chung, 2012 Science)]. Figure reprinted with permission from AAAS

Transition path is an important property for resolving folding mechanisms as it is insensitive to the free energy barrier and can provide direct mechanistic information about Kramer's diffusion coefficient. Folding time on the other hand, depends exponentially on the free energy barrier and cannot provide information on the intramolecular diffusion coefficient. A protein spends most of the time on the molecular trajectory in either folded state or unfolded state. Experiments probing either of these states can thus provide information about the conformational fluctuations happening in unfolded state or unfolded state. Transition path on the other hand is where the actual transition between the states takes place and contains all the necessary details about the self-assembly process. The dynamics while crossing the barrier, more specifically, the

diffusion coefficient along the transition state contains information about the energetic frustration along the protein folding pathway on the folding energy landscape.

Direct measurement of transition path times is possible from molecular trajectories obtained from atomistic simulations with an explicit solvent model that provides protein conformations in picosecond time intervals. The challenge is to reach the long timescales required to watch folding and unfolding, but recent developments have led to simulations that result in multiple folding and unfolding events for some of the fastest folding proteins[53,54]. However, these simulations are based on imperfect physical models (force fields) of the interatomic interactions that underlie the dynamics of biomolecular systems[55]. Current protein force fields are, for the most part, derived by fitting parameters to data from quantum level calculations or experiments on small molecules as model compounds. Obtaining high-resolution experimental data on proteins is thus essential to demonstrate theoretical predictions and benchmark and refine simulation methods. Experimental approaches employed to measure transition path times are discussed in later sections of this chapter.

1.6 Fast folding proteins

Traditional protein folding experiments have, for the most part, probed single domain proteins that folded slowly in a highly cooperative (all or none transitions) fashion (tens of milliseconds to minutes) and have led to the generalization of two-state folding kinetics across the protein world. Two state folding kinetics implies that the intermediate structures along the protein folding trajectory during a transition between folded and unfolded ensemble, transition state ensemble, are highly unstable and hence inaccessible to experimental probes. This is a manifestation of large free energy barriers, which are sparsely populated at experimental conditions for such proteins. The origin of this free energy barrier, according to energy landscape theory, is the asynchronous compensation of the conformational entropy and stabilization energy as the folding proceeds. Hence, a perfect compensation of the energy terms should theoretically result in a scenario where protein undergoes folding in a downhill fashion without crossing any free energy barrier. Two-state behaviour introduces serious limitations to resolving protein folding trajectory as it doesn't offer any scope for investigating the part of the trajectory that corresponds to the actual conformational transition.

Technological developments on the experimental front have permitted researchers to access kinetic processes occurring in the microsecond and nanosecond timescales. Ultrafast kinetic measurements using optical triggering with nanosecond laser pulses have made it possible to study the fastest-folding proteins along with the fundamental processes in folding. Relaxation dispersion NMR spectroscopy techniques have been used to probe conformational excursions of protein molecules. This has led to the experimental observation of fast-folding proteins that fold over very small or no free energy barrier and opened up the unique opportunity to resolve the partially

(un)folded intermediates along folding pathway[26,56–58]. These proteins have broad population distributions which is a direct outcome of not having a significant free energy barrier and show complex kinetics suggestive of gradual protein folding as opposed to highly cooperative two-state kinetics. The absence of a free energy barrier results in the timescale of folding dynamics and conformational dynamics being comparable and makes it possible to measure conformational dynamics from macroscopic kinetics[59]. Protein folding on a free energy landscape is limited by the free energy barrier and the diffusion coefficient. Downhill fast-folding proteins are ideal model proteins to experimentally probe the diffusion coefficient, a factor indicative of the friction within the protein molecule (internal friction) and the friction due to the solvent molecules (solvent viscosity).

Protein stability, folding rate and cooperativity are closely connected with faster folding proteins, often showing minimal stability and minimal cooperativity. A large number of proteins that exhibit disordered structure at native conditions have been identified recently[17]. These proteins, identified as intrinsically disordered proteins (IDPs), exist partially disordered at the physiological condition and folds to form stable native structures under thermodynamically favourable conditions[60]. The functional link between the noncooperative folding and ability to perform various binding modes depending on the binding partner and thermodynamic conditions makes these attractive targets for functional protein engineering. The gradual folding mechanism has enabled these proteins with a built-in mechanism to act as conformational rheostat-based nanomachines which have potential applications in biological protein-based sensors[61].

We have chosen fast-folding proteins to carry out mechanistic investigation on protein folding using the SM-FRET technique. We exploit the merits of employing fast-folding proteins to benchmark results from simulations and to obtain previously inaccessible features of folding[62]. In brief, using fast-folding proteins as experimental models to dissect the rates equation has the following advantages.

- Transition paths contain important mechanistic information about the folding process, experimentally resolving them has been challenging because these are rare and nearly instantaneous jumps on the folding trajectory. Fast-folding proteins interconvert between the native and unfolded state in sub-milliseconds, and hence the number of transitions that can be observed under experimental duration is maximized.
- Obtaining the population distribution of various species present along the folding reaction coordinate under different conditions is important in validating different folding behaviour (whether there are multiple pathways or a single dominant pathway). Energy and population are related by the exponential Boltzmann relationship, $p \propto \exp(-E/k_B T)$. Fast folding proteins have a minimal (<3 RT) free energy barrier if any, and hence the probability of observing population over the free energy barrier is maximized.

- Sub-millisecond rates facilitate performing SM-FRET experiments under high count-rate conditions and free diffusion (immobilized experiments introduce unwanted artifacts due to protein-surface interaction that affects the folding dynamics). The experimental observation time window for free diffusion experiments is between 0.5 ms - 2.0 ms and sub-millisecond folders have relatively high transition probabilities within this time frame.

- We have discovered through simulations that transition path times increase with lower free energy (Chapter 3), and hence the transition paths of fast folders should be easier to resolve. This is opposite to what is generally perceived of transition path times. A detailed analysis of this relationship is done in chapter 3.

1.7 Experimental investigation of fast folding mechanism

Protein folding dynamics and their dependence on various factors have been studied extensively for decades by introducing mutations and changing the physicochemical environment of the protein[63,64]. Folding time on a 1-dimensional projection can be defined as the average time a molecule spends in the unfolded well before it successfully crosses the barrier to get to the folded well, and the average unfolding time is the average time spent in folded well before crossing over to the unfolded well. These wait times on the molecular trajectories are distributed exponentially and the average (un)folding time is given by the corresponding rate coefficients. Average folding time for a protein with two states separated by a free energy barrier ΔG_f^* on a 1-dimensional free energy surface derived from Kramers expression is given by[65]

$$t_f = \frac{2\pi}{\beta D^* \omega^* \omega_u} \exp\left(\beta \Delta G_f^*\right)$$

where $\beta = {}^{1}/_{k_B T}$, k_B is the Boltzmann constant, T is the temperature, D^* is the diffusion coefficient for the top of the barrier, $(\omega^*)^2$ and $(\omega_u)^2$ are the curvatures at the top of the barrier and the bottom of the unfolded minima respectively. Mutational analysis of the folding rate constants attributes the changes in observables to the exponential term of the rate equation and is based on the fundamental assumption of a significant free energy barrier. This assumption of folded and unfolded states being separated by a high energy barrier, effectively reduces folding to a two-state model and makes it easier to interpret the experimental data. However, it limits the possibility of monitoring/resolving the conformational evolution of protein during the self-assembly process. Direct measurement of folding dynamics from macroscopic kinetics is done for fast-folding proteins with minimal free energy barriers[59]. Such proteins are ideal candidates for resolving the mechanism of folding.

Experimental studies of folding dynamics involve deducing information about the microscopic dynamics and energetics involved in folding from measurable macroscopic kinetics[66]. Two main experimental strategies have been commonly used to study protein folding kinetics. The first category is a transient perturbation method where the equilibrium between folded and unfolded population is perturbed quickly and the relaxation to the new equilibrium is measured using experimentally observable properties[67]. The second includes equilibrium measurements where fluctuations between the folded and the unfolded states under experimental conditions are probed. Accurate measurement of folding kinetic from equilibrium requires monitoring signals that can be observed for folded and unfolded population without averaging. Both methods require experimental probes with sufficient time resolution to reliably resolve various conformational states. Spectroscopic tools like fluorescence[68,69], infrared (IR) spectroscopy[70], circular dichroism and nuclear magnetic resonance (NMR)[71,72] are common tools utilized to monitor structural change and protein conformational change.

Extracting kinetic information from equilibrium methods unlike transient perturbation methods do not require a fast initiation. But they require a significant population of folded and unfolded states to be present at experimental conditions. The interconversion rates between the states are measured from partial averaging of the observed signal when the intrinsic timescale of the experimental technique is comparable to the protein dynamics. NMR methods have been extensively used due to the sensitivity of line broadening to exchange rates between different states[73,7469]. For relatively stable proteins, equilibrium experiments are usually carried out under denaturing conditions, and the results are extrapolated to physiological conditions. One of the disadvantages of NMR technologies is that they only provide indirect information regarding dynamics. NMR Relaxation or line broadening experiments can be used to derive folding or unfolding rates, which are frequently in excellent agreement with those determined directly, such as by laser T-jump experiments. However, any departure from simple kinetics, such as fast kinetic phases or non-exponential progressions, are not detectable[76] unless used with comprehensive atom-by-atom analysis approaches[25].

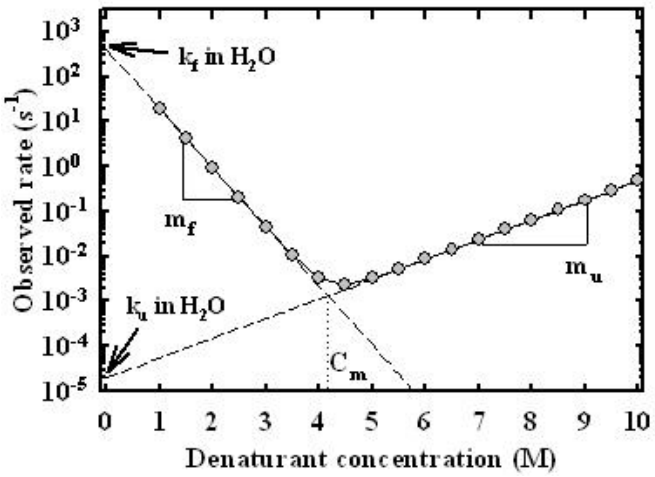
Chemical denaturation is a common approach for experimentally assessing the conformational stability of proteins. Experiment involves monitoring the signal from an experimental probe (absorbance, fluorescence, calorimetry etc) at increasing denaturant concentration to obtain a chemical denaturation curve (e.g. Figure 7-6). The data is commonly analyzed with two-state models, by defining an equilibrium between folded and unfolded state. It reports on the thermodynamic stability of the native conformation relative to the unfolded conformation and the equilibrium constant ($K_{eq}=k_u/k_f$)[77]. The m-value is a key parameter obtained from this analysis, it is a measure of cooperativity and is related to the steepness of the transition region and the solvent surface area. The concentration of denaturant at which there is an equal fraction of folded and unfolded population is called midpoint denaturation concentration (C_m). Conformational stability in the absence of denaturant, $\Delta G_u(H_2O)$, is determined as the product of C_m and m-value (refer to section 3.3 for the two-state description of unfolding curves). While the two-state approach for analysing data works for slow folding two-state proteins,

comprehensive analysis strategies are available for fast-folding proteins that fold gradually with minimal cooperativity[78]. When analysing multiple data sets (such as unfolding monitored by different probes), the fitting is done globally. This helps in keeping global parameters like ΔG and m-values constant while accounting for the differences in individual folding and unfolding baselines.

Transient techniques include mixing methods like the stopped-flow[79] and the continuous flow[80] where buffer conditions are rapidly changed by mixing two solutions, initiating protein folding or unfolding of the protein sample. Chemical denaturants like urea and guanidinium chloride are utilized here due to their ability to unfold almost all proteins. Concentration of denaturant is changed to trigger folding or unfolding by either diluting out or increasing the concentration respectively. Stopped flow methods are limited by ~1 ms dead time, while continuous flow methods significantly improve the time resolution to ~50 μs. Plot of the log of apparent rate constant versus denaturant concentration obtained from refolding and folding mixing techniques is very useful in interpreting the folding kinetic data. This plot is known as chevron plot due to the V-shape of the data. Chevron plot analysis provides folding rate (k_f) and unfolding rate (k_u) in the absence of denaturant (obtained from extrapolating the folding and unfolding arms of chevron), the m-values for folding (m_f) and unfolding (m_u) and information on the position of transition state. The m-value determined kinetically (m_{kin}) from the slope of the two limbs of chevron is compared with the m-value from equilibrium experiments (m_{eq}). The ratio m_{kin}/m_{eq} is expected to be smaller than 1 as the effect of barrier is more pronounced on the kinetics and can serve as an indicator of the barrier height at midpoint conditions[78].

Temperature and pressure jump methods are also used extensively to introduce perturbations in samples and have substantially higher time resolutions than mixing techniques[81–83]. The improvement in the time-resolution of these ultrafast kinetic techniques is utilized to study the kinetics of fast-folding proteins (μs folders). Laser T-jump experiments makes use of infrared laser pulse to heat the surrounding solvent molecules around proteins. They can achieve ~10-15 °C temperature change in few nanoseconds, thus introducing a sudden perturbation[84,85]. The relaxation back to equilibrium is monitored using spectroscopic tools leading to the determination of relaxation rates from the rate of decay of the signal. The amplitude of change in the signal can report on the changes in the population and the microscopic rates of interconversion between species when analysed with a suitable kinetic model[57]. It is however, important to mention here that comparison of folding rates and speed limits obtained from T-jump experiments and denaturation kinetic experiments need to be done with caution as T-jump experiments are based on data at high temperature (~340 K)[78]. While temperature and chemical denaturants produce similar effects on the free energy barrier, temperature dependence of the diffusion coefficient makes the effect of temperature on folding kinetics complex. FES based analysis (discussed in detail in

Chapter 2) of folding kinetics is particularly attractive in this regard as they can report on the diffusion coefficient and the free energy barrier directly.



Transition path time corresponds to the timescale of actual structural reorganization happening during folding. They are the segments of any stochastic trajectory in which the native structure is structured. They inform directly on the underlying molecular mechanisms. As the transition state population doesn't accumulate and cannot be synchronized, bulk experiments can only provide indirect information about the

Figure 1-7 : A typical chevron plot for a two-state folding protein. Each point on the chevron is obtained from the relaxation rates for folding (data to the left of C_m) and refolding (data to the right of C_m) transient mixing experiments.

transition state. The ϕ-value analysis is the most commonly used experimental technique to probe the transition state, and it uses perturbation in energetics to make structure prediction[86–88]. The method introduces mutations in proteins and studies the relative change in folding stability (equilibrium constant) and folding kinetics (rate coefficient). Mutations that slow down the folding reaction destabilize the transition states and imply that the mutated residue is involved in native contacts. A series of systematically selected mutations can map out the network of interactions involved in the transition states. Transition paths are rare and fast events. Resolving them requires time resolution in microseconds to catch the fast process as well as long observation times to catch the sparse events. Experimentally resolving transition path times is possible using single-molecule measurements. Transition path times correspond to the time spent crossing the barrier and hence are often inaccessible during experimental inspection for proteins with a high free energy barrier. However, downhill folding proteins and proteins with minimal free energy barrier offers a unique possibility of

probing the transition path time as they have the barrier significantly populated at experimental conditions.

1.8 Single-molecule experimental techniques

Single-molecule methods are extremely attractive techniques that can, in principle, resolve the stochastic transitions of individual protein molecules while they undergo folding and unfolding transitions in real-time. While bulk experiments can provide ensemble-averaged information, they do not have the ability to observe the myriads of different microscopic pathways and mechanisms that can lead to the folded state. Hence mapping the topography of the folding energy landscape or measuring the relevant conformational motions involved in the process cannot be achieved using bulk experiments. With sufficient time resolution, single-molecule methods permit to measure the dwell times of all the relevant states (folded and unfolded), which are connected to the rates measured in bulk experiments, but also the transition paths that protein molecules follow when they jump between states. They can recapitulate the isotropic (un)folding conditions of bulk experiments while they can also resolve folding events one molecule at a time and uncover rare species or transitions that are hidden within the averaging that takes place in ensemble experiments.

The development of single-molecule fluorescence and force spectroscopy (atomic force microscopy, optical tweezers, and magnetic tweezers) techniques has made it possible to experimentally observe biomolecules at equilibrium in their ground states and as they undergo folding-unfolding transitions[89–94]. Fundamental questions of protein folding like whether the folding proceeds through multiple pathways or if there are obligatory intermediates along the folding pathway and how are the various conformational states distributed can now be answered through single-molecule experimental methods. Equilibrium populations along the folding pathway can be observed without ensemble averaging through SM methods, and this has been utilized to study the folding nature of proteins with direct access to subpopulations[95–97]. SM experiments allow direct measurement of folding and unfolding rate constants from the distribution of wait times in the unfolded and folded state, respectively. AFM techniques study the extension of proteins held between a flexible cantilever and a surface[98,99]. Optical tweezers make use of a similar principle where the molecule is held between two beads, and the extension is studied while applying forces to the molecule[100–102]. The end-to-end distance of the protein molecule acts as a good reaction coordinate and allows the monitoring of dynamics directly.

However, interactions between the surface to which protein is attached and the protein itself is a disadvantage of these methods. SM methods based on Förster resonance energy transfer (FRET) make use of the distant dependent non-radiative energy transfer between suitable fluorophore pairs to track the protein (un)folding[96,103]. This method has the added advantage of measuring folding under free diffusion conditions or on

immobilized molecules. Free diffusion SM-FRET experiments are particularly attractive methods as they eliminate the need to immobilize protein to a surface and hence are free of spurious events caused by protein-surface interaction. Additionally, free diffusion experiments make it possible to measure tens of thousands of single-molecule trajectories easily. This large dataset ensures an accurate statistical representation of folding dynamics and enables the capture of rare events. SM-FRET methods have been widely utilized to estimate protein folding transition path times[47,51,104–106]. A statistical analysis method based on likelihood calculation is employed with SM-FRET to obtain the average transition path time from the photon trajectories. This likelihood calculation-based approach extends the practical time resolution of the technique to interphoton time, making it possible to resolve transition path times (few μs). We will be expanding this method to be used along with a 1-dimensional free energy surface-based model to analyses SM-FRET data on fast-folding proteins. Methodology for MLA based investigation of protein folding dynamics with SM-FRET data will be described later in the dissertation (chapter 2).

1.9 Research objectives and dissertation outline

Experimental investigation of folding dynamics of fast-folding proteins with very small barriers (smaller than 3 *RT*) or no barrier with single-molecule detection techniques hold the promise of providing the type of experimental information required to thoroughly benchmark and refine atomistic simulations. All important details about conformational dynamics happening in the protein within the interphoton time are embedded in the SM photon trajectory. The MLA method can extract these details, with the limitation being the assumptions included in the model used to describe folding.

This dissertation will demonstrate a free energy surface-based approach that can be implemented with MLA for use with data from free-diffusion SM-FRET experiments. We then use the method to (a) deconstruct the terms of rate expression and (b) resolve the folding transition path times for fast-folding proteins. SM-FRET technique and the methodology to implement MLA with 1D-FES is described in chapter 2. Chapter 3 utilizes this method of analysis to resolve the dynamics of fast folding protein, En-HD. We further test the performance of FES based approaches in comparison to other exiting methods to estimate the transition path times for folding (chapter 4). This is done by utilizing a diverse set of simulated SM-FRET trajectories. Chapter 4 further dissects the contribution of different parameters (namely height of the free energy barrier, curvature of the barrier and asymmetry of the overall FES) to the overall transition path time.

We then utilize the same methodology to investigate the effect of intrinsic protein properties like topology and secondary structural content and experimental conditions like ionic strength, and nature of denaturant have on various folding parameters (chapter 5 and chapter 6).

Chapter 2

2 Resolving folding dynamics from SM-FRET experiments: Methods and Analysis

2.1 Single-molecule FRET

Single-molecule methods have evolved into one of the mainstream techniques used to probe conformational dynamics and interactions of biomolecules. It is now used to address various problems ranging from understanding the mechanism of enzymatic action, protein folding dynamics, protein-protein interaction, protein-nucleic acid interaction, macromolecular assembly of proteins, and many more[43,47,51,107–109]. The ability of single-molecule experiments to access information without ensemble averaging makes it an attractive technique to obtain distributions of molecular properties.

Single-molecule techniques are associated with specific challenges that have to do with detecting signals coming from an individual biomolecule of interest. Single-molecule detection in condensed phase like liquids and solids is challenging than in the gas phase because of the contribution of background signal arising from Raman and Rayleigh scattering. Contaminants present in the medium can also have a significant effect on the signal, and hence the purity of the medium in which samples are present is of utmost importance. It is also vital to ensure that the solvents don't produce any signal as even small contributions from a large number of solvent molecules present in comparison to one molecule of interest can cause excessive background. These limitations make fluorescence spectroscopy well suited to use as a signal for single-molecule detection because of its high sensitivity[41,110,111]. Stokes shift, the shift in the wavelength of emitted light compared to the excitation light, can be used to filter out contributions from excitation scattering, effectively reducing background. Another strategy used to decrease the effect of unwanted signal coming from solvent is decreasing the observation volume and hence making the ratio of the solvent molecule to the molecule of interest smaller. This is achieved using confocal detection, where the excitation light is focused to a diffraction-limited spot of femtoliter volume.

An important limitation of single-molecule fluorescence is the low photostability of fluorophores in the condensed phase. Fluorophores can only undergo a limited number of excitation-emission cycles before becoming photo-inactive through photobleaching

and photo-blinking processes. Even though there are techniques available to improve the photostability of fluorophores, this remains a major limitation in expanding the time resolution of single-molecule fluorescence techniques[112–114].

2.1.1 Förster resonance energy transfer (FRET)

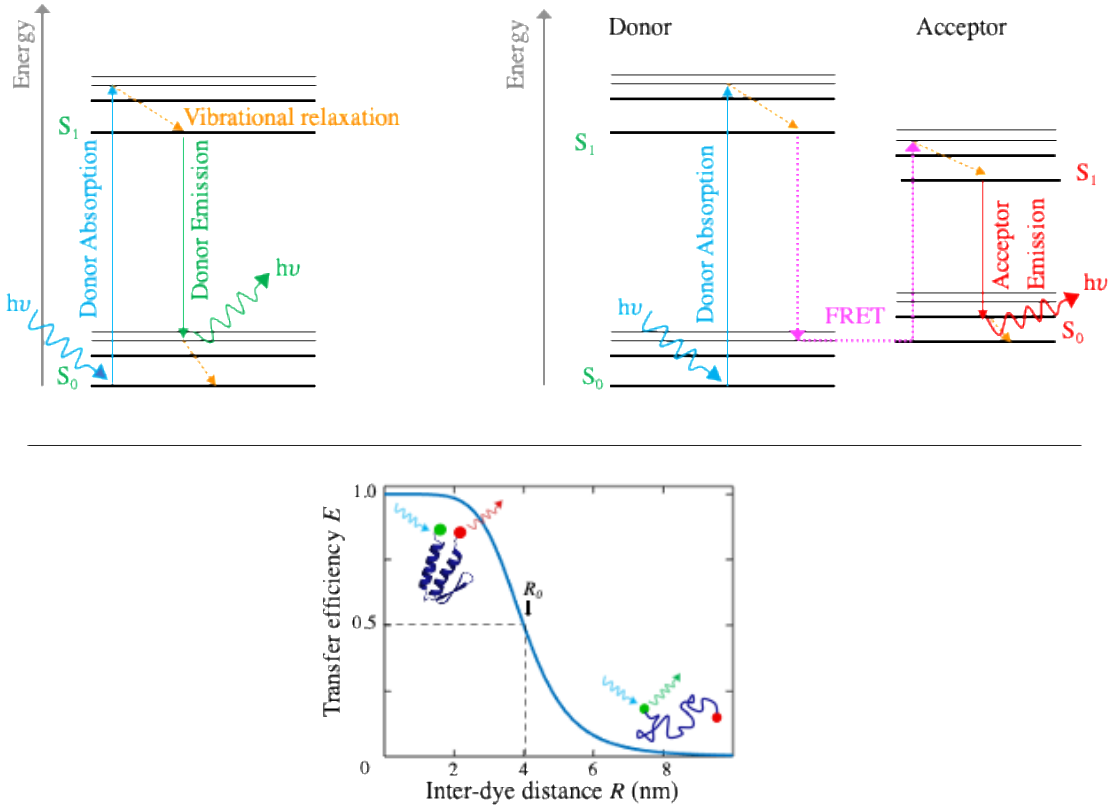


Figure 2-1 : Principles of fluorescence and FRET explained through Jablonski diagrams (A) Fluorescence: Absorption of a photon by the fluorophore excites the ground state electron to a higher energy state. Electron relaxes back to the lowest vibrational level within this excited state through non-radiational vibrational relaxation. Relaxation back to the electronic ground state takes place via photon emission leading to fluorescence (B) FRET : Excited state energy of the donor is taken up by a suitable acceptor in close proximity through non-radiational dipole-dipole coupling, leading to the excitation of acceptor electrons and further emission of a photon by the acceptor (C) Dependence of transfer efficiency on distance between acceptor-donor

Förster resonance energy transfer is a non-radiative excitation energy transfer between a donor (D) fluorophore in the excited state and an acceptor (A) fluorophore in the ground state. Given that there is suitable spectral overlap between the fluorophores,

FRET takes place through long-range dipole-dipole coupling. The transfer efficiency (E) depends on the inverse sixth power of the distance between the fluorophores (R), as expected for such energy transfer processes[115].

Rate of transfer efficiency is given by

$$k_T(R) = \frac{1}{\tau_D}\left(\frac{R_0}{R}\right)^6$$

where τ_D is the decay time of donor in the absence of acceptor and, R_0 is called the Förster radius. The efficiency of energy transfer is defined as the fraction of photons absorbed by the donor that is transferred to the acceptor through FRET and hence is given by the ratio of transfer rate to the total decay rate of the donor in the presence of the acceptor

$$E(R) = \frac{k_T(R)}{\tau_D^{-1} + k_T(R)}$$

$$E(R) = \frac{R_0^6}{R_0^6 + R^6}$$

Förster distance R_0 is the inter-dye distance at which transfer efficiency is 50%. It can be calculated from the spectral properties of the fluorophores using the following expression

$$R_0^6 = \frac{9000\ \ln 10\ \kappa^2 J Q_D}{128\pi^4 n^4 N_A}$$

κ^2 is a factor describing the relative orientation of the acceptor and donor transition dipoles and can range from 0 to 4

$$\kappa^2 = (\cos\theta_T - 3\cos\theta_D \cos\theta_A)^2$$

where θ_T, θ_D and θ_A are angles between the emission transition dipole of the donor and the transition absorption dipole of the acceptor, the angles between these dipoles and the vector joining the donor and the acceptor dipole, respectively. It takes a value of 0 for perpendicular dipoles, 1 for parallel dipoles, and 4 for head-to-tail parallel dipoles.

It is usually taken to be 2/3, which is the orientation factor for the dipoles randomized through rotational reorientation prior to energy transfer and holds true if the fluorophores are not constrained. As κ^2 has a sixth power dependence on distance, a variation from 1 to 4 only changes the calculated distance by 1/4th, though dipoles being oriented perpendicular to each other can lead to serious errors in distance calculations and hence needs to be dealt with caution. It is crucial to confirm the orientation factor by carrying out anisotropy measurements, especially in cases where fluorophores are introduced into the cores of proteins or near binding sites. Q_D is the quantum yield of the donor in the absence of acceptor, n is the refractive index of the medium and is usually taken to be 1.33 for biomolecules in aqueous solutions, N_A is the Avogadro's number (6.023×10^{23}), and J is the overlap integral which depends on the extent of spectral overlap of donor emission and acceptor absorption. These derivations assume that the lifetime of the donor is not altered by the presence of an acceptor other than by Förster resonance energy transfer. Detailed analysis of apparent transfer efficiency with considerations from donor quenching and enhanced acceptor emission will have to be done in cases where this is not true.

As evident from the equation for FRET efficiency, the distance dependence of transfer efficiency is the strongest when the inter-dye distance is near R_0. Transfer efficiency changes from 0.015 to 0.985 for a distance change from $2R_0$ to $0.5R_0$. Organic fluorophore pairs with R_0 in the range of 2 to 9 nm are commercially available and are being used as convenient "spectroscopic ruler" to monitor conformational changes of biomolecules [103,116]. Since the first report of the use of FRET in combination with single-molecule detection by Ha et al., the SM-FRET technique has improved immensely and is widely used in the study of biomolecular structure and dynamics[117].

2.1.2 Single-molecule FRET: Instrumentation

Experimental set-up for single-molecule FRET involves an excitation light source, confocal microscope, optics for collecting and separating donor and acceptor emission, single-photon detectors, and photon counting device. The excitation laser beam is focused to a diffraction-limited spot by a high numerical aperture (*N.A.* = 1.49) objective to get a small confocal volume (femtoliter). In free diffusion-based experiments, sample molecules are freely diffuse in and out of this small volume. Immobilized experiments have a biomolecule immobilized on to a surface within the confocal volume. The donor fluorophore in the confocal volume, when excited, can relax back either by emitting a photon or by transferring excitation energy to the acceptor, thus resulting in acceptor emission. The number of donor and acceptor photons observed depends on the energy transfer efficiency, which in turn depends on the distance between the fluorophores. This emission is collected by the same objective used to focus excitation light. The numerical aperture (*N.A.*) of the microscope determines lateral resolution (*d*) of the microscope.

$$d = \frac{1.22 \times \lambda}{N.A.}$$

where λ is the wavelength of light. A high numerical aperture objective is essential for single-molecule measurements to ensure femtoliter confocal volume. High numerical aperture also provides high collection efficiency as the emitted light is collected by the same objective in an epifluorescence arrangement. The emitted light is passed through suitable optical filters to remove excitation leakage and scattered light. It then passes through a pinhole that acts as a spatial filter to remove the out-of-focus light. This completes the confocal detection and, when used in conjunction with a low sample concentration (picomolar) ensures the detection of signal coming from a single biomolecule except for a few background photons. Emission is then separated into acceptor and donor emission by using suitable optics. Emission from both channels is then collected by two avalanche photodiodes (APD) which have the resolution to detect single photons and are registered by a multichannel photon counter. Detailed information on the components of the optical set-up and procedure to align and optimize the instrumentation is described in the appendix.

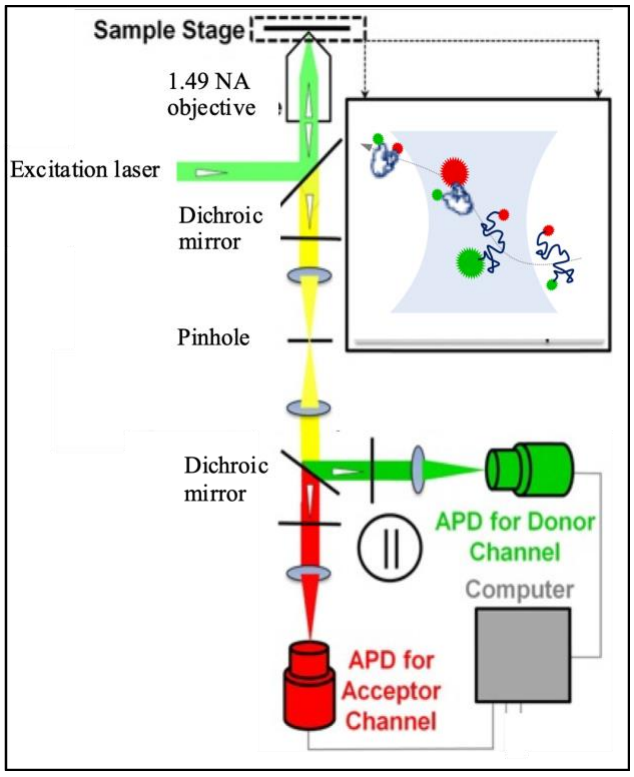


Figure 2-2 :Schematic of the two-colour single-molecule FRET instrumentation set-up

2.1.3 Fluorescence labelling

Biomolecules need to be conjugated with fluorophores prior to making single-molecule FRET measurements. Natural amino acids like tryptophan are not suitable for single-molecule detection due to their low extinction coefficient and photostability. Fluorescent proteins are gaining popularity in this regard as they can be expressed in cells with the proteins of interest, thus removing the need for external labelling of fluorophores. However, they are of limited use in studying the dynamics of smaller proteins owing to their large size and disadvantageous photophysical properties. Labelling with bright and photo-stable extrinsic fluorophores is thus inevitable for single-molecule fluorescence techniques. Two fluorophores with suitable spectral properties need to be carefully inserted into specific positions in proteins so that conformational change in the molecule can be probed as FRET change. Strategies including making use of chemistry between side groups of natural amino acids to introducing non-natural amino acids are available for this purpose[118–120]. One of the most common approaches is cysteine derivatization using maleimide chemistry. Specificity is achieved by removing naturally occurring unwanted cysteines in the sequence and introducing cysteines in the desired positions. Dyes derivatized with maleimide functional group are then reacted with the thiol groups of cysteine residues to achieve the conjugation. It is important to note that this strategy cannot be utilized for proteins with cysteines required for the structure.

Labelled protein is further purified using multiple chromatography techniques to achieve separation from unlabelled and partially labelled protein. Established protocols for commercially available organic fluorophores functionalized with maleimide are available (refer to appendix). The labelling sites are chosen carefully to avoid the interaction of organic dye molecules with the protein and thus affecting the dynamics of proteins. Quantum yield of dyes can decrease on interaction with other residues, and rotational reorientation of dyes can be restricted as well. These issues are more pronounced when labels are introduced in the middle of the protein than when introduced in the ends. As we are interested in monitoring the end-to-end distance change of proteins, the labelling sites were chosen at C-terminal and N-terminal, reducing the interaction of dye molecules with the rest of the protein. Alexa Fluor™ 488 C5 Maleimide and Alexa Fluor™ 594 C5 Maleimide were used as FRET pairs for our study owing to their high extinction coefficient, high quantum yield, better photostability, high solubility in water and Förster distance of 5.4 nm, making it ideal for studying dynamics happening in 2.5 nm to 8 nm[96] (See appendix for labelling protocol).

2.1.4 Histogram analysis of SM-FRET data

Raw data from the single-molecule FRET experiment is a time-stamped stream of donor and acceptor photons. The colour of the photon depends on the transfer efficiency and hence contains information about the distance between the acceptor and the donor molecule. The interphoton time depends on the dynamics happening within the biomolecule. The simplest way of analysing single-molecule data is by dividing the data into small time bins and grouping the donor and acceptor photons in the bins. Transfer efficiency is then be calculated ratiometrically for the bins with photon count greater than a set threshold and is given by

$$E = \frac{n_A}{n_A + n_D}$$

where n_A and n_D are the total number of acceptor and donor photons in the bin. Experiments are done for long enough time to obtain a statistically relevant number of bins to get a distribution of transfer efficiency. The distribution of transfer efficiencies is called FRET efficiency histogram (FEH) and gives information about various conformations of protein present under the experimental conditions and about the rates of dynamics involved to an extent[41,94].

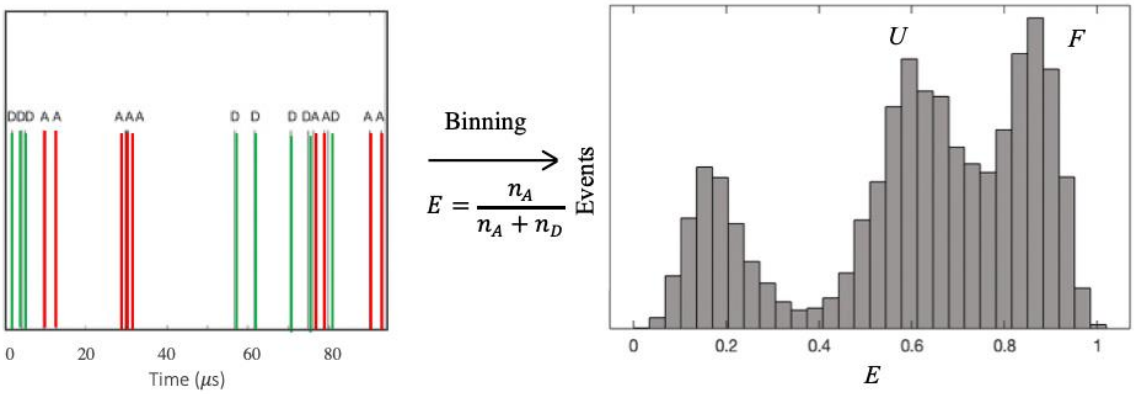


Figure 2-3: SM-FRET data analysis and generation of FRET efficiency histograms (FEH)

As transfer efficiency is inversely proportional to the sixth power of inter-dye distance, higher efficiency corresponds to the conformations in which dyes are in proximity and lower efficiency corresponds to dyes being separated from each other. This is illustrated in the FEH in Figure 2-3, unfolded state of the protein shows lower efficiency, and the higher efficiency peak corresponds to the population of the folded protein. This demonstrates an important advantage of single-molecule techniques, the ability to

resolve various sub populations even when they coexist at an experimental condition while the bulk measurement would have only provided an averaged-out value of transfer efficiency. An additional low FRET peak is commonly observed in FEH corresponding to molecules without an active acceptor which could be a result of incomplete labelling or acceptor getting photobleached.

Photon emission is a stochastic process and hence single-molecule photon detection is accompanied with a statistical fluctuation in the number of photons detected. This results in the broadening of the transfer efficiency distribution even when the inter-dye distance is not changing. The variance caused by this stochastic fluctuation of photons is called shot noise and is given by

$$\sigma_{SN}^2 = \frac{E(1-E)}{N_T}$$

where E is the transfer efficiency and N_T is the number of photons in the bin/burst. As a result of shot noise broadening, the time resolution of SM-FRET with binning analysis is set by the minimum time required to collect ~25 photons (25 photons correspond to $\sigma \pm 0.10$ for $E = 0.5$). The low collection efficiency of the confocal microscope set-up (2-4 % collection efficiency) and the fluorescence quantum yield of dye-pairs limits the smallest time bin with 25 photons and thus time resolution of SM-FRET technique.

Detailed analysis of FEH using a multi-Gaussian approximation to resolve various dynamically averaged populations is possible[121]. Although significant progress has been made in making use of FRET efficiency histograms, extracting quantitative information about the conformational dynamics is still challenging. FEH is a reduced representation of the whole-time stamped photon trajectory and thus cannot be sufficiently sensitive to the parameters of the conformational dynamics. Additionally, when the conformational fluctuation occurs on a timescale comparable or faster than the bin times, FRET efficiency histograms get dynamically averaged, and conformational dynamics cannot be resolved. Finally, the application of FEH analysis methods to single-molecule data from free diffusing molecules is demanding owing to the fluctuation in photon count rates as the molecule passes through the laser spot.

2.1.5 Data processing of photon trajectories from free-diffusion SM-FRET experiments

2.1.5.1 Identification of photon bursts

Single-molecule data was recorded as time-stamped photon sequences with picosecond resolution. We devised a simple method to extract the protein crossing events from the free diffusion single-molecule FRET data (most of the data acquired will be discarded

as only ~2% of total data collected account for protein crossing events). Regions of high photon density were picked by an initial binning of the data into 200 μs bins and the bins with more than 55 photons were identified. We further scanned along the edges of these 200 μs photon trajectories to identify the beginning and end of the photon bursts. This was done by scanning till we got to the regions with a count rate lower than 2.5 times the background count rate.

2.1.5.2 Removal of bursts with an inactive acceptor fluorophore

Photon bursts from free diffusion SM experiments were checked for acceptor photobleaching or photo blinking, and only the bursts free of such dye artefacts were selected for final analysis. This was achieved by following a simple 3 step algorithm: i) Parameters for a 3-state model with a low FRET state corresponding to the inactive acceptor state and having $E = 0.2$ were determined using MLA analysis. The crosstalk between two channels results in a 20% emission leakage to the acceptor channel in our SM-FRET set-up and corresponds to $E < 0.2$ ii) The Viterbi algorithm was used to generate the state trajectories corresponding to this 3-state model[122] iii) Trajectories in which the 0.2 *E* state was visited at least once were identified and corresponding bursts were discarded.

2.2 Gopich-Szabo Maximum likelihood analysis (GS-MLA) method

Resolving biomolecular dynamics using SM-FRET experiments through transfer efficiency histogram analysis has practical limitations. The time-stamped photon trajectory from an SM-FRET experiment contains invaluable information about the kinetics and dynamics happening in the biomolecule. Methods of extracting this information without grouping photons together can extend the resolution of this powerful experimental technique. Gopich and Szabo developed a likelihood analysis method to extract kinetic and dynamic parameters from the sequence of the colour of photons and interphoton times. In an SM-FRET experiment, the probability that a photon is emitted by either the donor or acceptor depends on the distance separation between the fluorophores. Thus, the pattern of the colour of photons contains information about the inter-dye distance, which is in turn modulated by the conformational dynamics happening in the biomolecule. GS-MLA model calculates the likelihood that a given set of parameters of a predefined model describes the observed time-stamped photon trajectory. This likelihood function is then maximized to obtain the parameters that best describe the data[123,124].

The MLA method doesn't involve binning photons together or explicit count rates, hence the time resolution is not limited by the count rate and extracts most information from the photon sequence. The method is particularly useful when photon stream is

affected by high background signal or has contributions from photophysical properties of the various conformations present. The method also has the added benefit of being easily applicable to single-molecule data from free diffusion experiments as well as the ones from biomolecules immobilized onto a surface. Versions of MLA with considerations for conformation dependent count rate of donor and acceptor fluorophore have also been developed for cases when one or both dyes undergo significant quenching[124].

The output from a free diffusion single-molecule FRET experiment is a time stamped photon trajectory and is characterized by photon colours and interphoton time. Interphoton time observed in SM-FRET experiments with organic fluorophore pairs acting as FRET acceptor-donor are generally in the range few microseconds. Photon statistics are influenced by processes like translational diffusion or photophysical processes of interest to us (e.g., absorption and emission of photons). Translational diffusion through the confocal volume is a slow process compared to interphoton times and occurs in millisecond timescale. Photophysical processes like excitation, decay and energy transfer are fast and take place in sub-microsecond timescales. Fast conformational dynamics happening in proteins are also in this fast regime.

The count rates depend on the conformational state and the position of the molecule in the laser spot and fluctuate when either one of these changes. The distribution of interphoton times, regardless of the colour of the photon, is defined by the sum of acceptor and donor count rate. If the total count rate is independent of the conformation, then the interphoton depends only on the location in the laser spot and hence contains information about translational diffusion and not conformational dynamics. The probability that a photon of a particular colour is emitted depends on the ratio of the acceptor and donor count rates of the conformation emitting the photon. The Colour of a photon is exclusively dependent on the conformation and is independent of the position in the laser spot. Hence the key parameters of a model describing the conformational dynamics can be extracted by decoding the pattern of colours from SM-FRET experiments.

The model assumes that every photon of a photon trajectory is originating from one of the many conformational states present. Each state is characterized by an apparent FRET efficiency which is defined as the ratio of acceptor count rate to the total count rate. These efficiencies are related to the inter-dye distance. The exact relationship between distance and apparent FRET efficiency depends on the dye photophysics, fast orientational dynamics and the detection efficiency for acceptor and donor photons. The transition between the states is defined by a rate matrix **K**. Likelihood function for a photon trajectory with *N* photons is then given by

$$L_t = 1^T \prod_{j=2}^{N} \left[\mathbf{F}(c_j) \exp(\mathbf{K}\tau_j)\right] \mathbf{F}(c_1) p_{eq} \qquad (1)$$

where **K** is the rate matrix, τ_j is the time interval between j^{th} and $(j-1)^{\text{th}}$ photon, p_{eq} is the vector with equilibrium probabilities of different states, c_1 is the colour of the first photon in the trajectory, c_j is the colour of the j^{th} photon. $\mathbf{F}(acceptor) = E$ and $\mathbf{F}(donor) = I - E$ where E is a diagonal matrix with FRET efficiencies of states included in the model. The total likelihood for multiple bursts was obtained by summing individual likelihoods, $\log L = \sum_N \log L_t$. Conventionally, GS-MLA is used in combination with simple kinetic models like two-state or three-state model, but we will be implementing the ML analysis with free energy surface-based models (section 2.4 and section 2.4.2).

2.3 Estimating average transition path time from likelihood difference

The procedure commonly utilized by various groups determines the average transition path time from the difference in the maximum likelihood of a two-state model (i.e. instantaneous transition: path not resolved) and that of a model with a virtual intermediate state (its lifetime provides the transition path time)[47,51,104].

Two-state model of protein folding

For a two-state system with instantaneous transitions, protein molecules populate either the folded state (native) or the unfolded state. If p_F is the fraction of protein in the native state and p_U is the fraction of protein in the unfolded state with FRET efficiencies E_F and E_U respectively and k_F and k_U are folding and unfolding rate coefficients, the equilibrium and kinetic parameters are given by

$$p_F + p_U = 1$$

$$k = k_f + k_u \qquad (2a)$$

$$k_f p_U = k_u p_F$$

The rate matrix, **K** is then diagonalized to get the eigen spectra from which relaxation rates are obtained.

$$\mathbf{K} = \begin{pmatrix} -k_f & k_u \\ k_f & -k_u \end{pmatrix}$$

Three-state model of protein folding

The three-state model represents a finite transition path time, defined by introducing an intermediate state S with lifetime $\tau_S = 1/2k_s$ that represents the average transition path time $\langle t_{TP} \rangle$. The kinetic scheme and parameters for the three-state model are

$$U \leftrightarrows S \rightleftarrows F$$

$$p_{F'} + p_S + p_{U'} = 1$$

$$k_f p_{U'} = k_u p_{F'} = p_S k_s$$

$$\mathbf{K} = \begin{pmatrix} -k_f & k_s & 0 \\ k_f & -2k_s & k_u \\ 0 & k_s & -k_u \end{pmatrix}$$

where $p_{U'}$, $p_{F'}$ and p_S are the fraction of protein molecules in the unfolded, folded and intermediate states respectively.

$\Delta \log L(\tau_S) = \log L(S) - \log L(0)$ is then calculated for a grid of τ_S and ε_S values, where $\log L(0)$ is the likelihood for the two-state model with instantaneous transition and $\log L(S)$ is the likelihood for the three-state model with a finite transition. Each kinetic model was fit to the photon trajectory data using the MATLAB multidimensional unconstrained nonlinear minimization (Nelder-Mead) algorithm *fminsearch* by minimizing the negative of log L. The two-state model was fit to obtain ε_F, ε_U, k_f and k_u. For the three-state model (finite transition path time), k_f and k_u were fixed to the values obtained with the two-state model. The FRET efficiency for the three states were obtained in two ways:

a) ε_F and ε_U from the two-state model and ε_S fixed to be halfway. This is a simpler, more common method of obtaining average transition path time[51], and will be referred to as $\Delta \log L_{2si\text{-}2s}$ in this dissertation.

b) ε_F and ε_U obtained from the GS-MLA of the three-state model by fixing k_f and k_u to the values obtained from the two-state fit. $\Delta \log L(S)$ is then calculated for a 2D grid of values of ε_S (in the range $\varepsilon_U \leq \varepsilon_S \leq \varepsilon_F$) and τ_S. The combination of ε_S and τ_S values that gives the highest $\Delta \log L(\tau_S)$ is selected. We will be referring to this method as $\Delta \log L_{3s\text{-}2s}$ in this dissertation.

These two methods are essentially equivalent with one difference: method *a* is simpler and assumes that the estimates of ε_F and ε_U are not affected by the introduction of the

virtual intermediate state (ε_F and ε_U values from two-state fit are correct); method *b* allows for relaxing ε_F and ε_U to accommodate for the presence of the virtual intermediate state. We introduced the latter because we realized that for surfaces with small free energy barriers, the two-state fit tends to merge the two minima (moves ε_F and ε_U away from their true values and towards each other).

2.4 Free energy surface-based models

One of the implications of energy landscape theory is the ability to model folding as a diffusive process along a one-dimensional free energy surface, given that a reaction coordinate indicative of the conformational transition is chosen[34–37]. Folding along the funnelled energy landscape is described as a progressive formation of partially structured conformations that needs to navigate through the local minima present along the pathway as it proceeds down the funnel towards native conformation. This funnel can thermodynamically be defined by a free energy term comprising competing contributions from conformational entropy and stabilizing enthalpy. This concerted change in enthalpy and entropy as one moves down the funnel determines the average drift up or down the funnel. Folding takes place on this drift as a stochastic process which to first approximation can be defined as a diffusive process, making the reduction of dimensionality of FES possible.

2.4.1 1-dimensional free energy surface (1D-FES)

A simple 1-dimensional free energy surface where the multidimensional folding energy landscape is projected on to a single order parameter that represents the degree of native structure in terms of local contact has been proven sufficient to capture the folding dynamics [78]. According to the model, stabilization energy and conformational entropy are described by continuous functions and the non-synchronous decay of conformational entropy and gain of stabilization free energy results in a free energy barrier. The contribution of stabilization energy (β_{stab}) and conformational entropy (β_{conf}) to the free energy barrier is dependent on the denaturational stress and hence can change drastically with the protein stability. An important outcome of the free energy surface approach, that emphasizes its model-independent universal description of protein folding, is that it follows Hammond's postulates. The position of the barrier top along the reaction coordinate shifts with the change in the stabilization energy, this change corresponds to changes in β_{stab} and β_{conf} and is dependent on the protein stability. It is interesting to note that despite the changes in β_{stab} and β_{conf}, the free energy surface approach has been able to reproduce V-shaped chevron plots, which are conventionally associated with a rigid transition state[56].

1D-FES model treats protein folding in the full essence of the energy landscape model. It is simple enough in terms of parameters to be combined with the maximum likelihood analysis method, which is used widely for analyzing photon trajectories from SM-FRET

experiments. The performance of the model in combination with the likelihood analysis method has already been tested on simulated single-molecule photon trajectories. This approach was able to distinguish between various folding scenarios with limited count rates and was found to be resilient to a moderate level of background photon counts[125].

Free energy, as a function of reaction coordinate *n*, ($\Delta G(n)$) is made up of contributions from conformational entropy ($\Delta S^{conf}(n)$) and stabilization energy ($\Delta H(n)$), both of which scales linearly with the total number of residues (*N*). Nativeness (*n*) is defined as the average probability of finding any residue in its native like conformation and hence is a continuous function from 1 to 0 with 1 corresponding to the native state and 0 being a fully unfolded state. Stabilization energy is calculated as an exponential function of *n* in agreement with mean-field theory and conformational entropy is calculated from Gibb's entropy formula.

$$\Delta H(n) = N\Delta H_{res}[1 + (\exp\ (\kappa_{\Delta H} n) - 1)/(1 - \exp(\kappa_{\Delta H}))]$$

$$\Delta S^{conf}(n) = N(-R[n\ln(n) + (1-n)\ln(1-n)] + (1-n)\Delta S_{res}) \qquad (1)$$

$$\Delta G(n) = \Delta H^{total}(n) - T\Delta S^{conf}(n)$$

where ΔS_{res} is the entropy penalty of fixing a residue in native conformation, we fixed a previously derived value of 16.5 J mol^{-1} K^{-1} for all the calculations in this study[126]. ΔH_{res} is the average stabilization energy per residue and $\kappa_{\Delta H}$ is a parameter that defines the curvature of the free energy surface. The free energy barrier arises because of non-synchronous compensation due to loss of conformational entropy and stabilization due to interaction energy.

We used a discretized version of the model where the potential of the mean force is calculated for *N* microstates (*N*=101), and the interconversion rate between the species is determined by intramolecular diffusion coefficient *D* as described in section 2.4.3. The probability of i^{th} microstate (p_i) is given by the expression

$$p_i = \exp(-V(i)/RT)/\left[\sum_{j=1}^{N} \exp(-V(j)/RT)\right] \qquad (5)$$

where *T* is the temperature of the system, *R* is the universal gas constant, *N* is the total number of states accessible to the system and *V(i)* is the energy function for the 1D-FES.

2.4.2 Landau free energy surface model

1D-FES described in previous section is an actual model for protein folding based on the contributions of different energy terms (entropic and enthalpic) to the overall free energy. It makes assumptions and predictions about the folding process. Here we are using a pure phenomenological model to obtain the one-dimensional FES. To represent the resulting one-dimensional free energy analytically, we use a simple description of the surface based on a Landau quartic polynomial. The model is similar to the variable barrier model developed in our group for the analysis of differential scanning calorimetry experiments[127], but in this case it is defined according to a generic order parameter (x) rather than the enthalpy. The model is parametrically as simple as a three-state model, but it provides a 1D free energy surface with adjustable barrier height and shape. The model is more flexible than the 1D-FES as the curvatures of the wells and the barrier can be adjusted. In this model, the Gibbs free energy as a function of the order parameter is defined at isostability conditions ($G_0 \Rightarrow G(\alpha_U) = G(\alpha_F)$) as:

$$G_0(x) = -2\beta\left(\frac{x}{\alpha_U}\right)^2 + |\beta|\left(\frac{x}{\alpha_U}\right)^4 \quad \text{for } x \leq 0$$

$$G_0(x) = -2\beta\left(\frac{x}{\alpha_F}\right)^2 + |\beta|\left(\frac{x}{\alpha_F}\right)^4 \quad \text{for } x > 0$$

where α_U and α_F are the positions of the unfolded and folded minima along the order parameter x, respectively, and x is defined with the condition that $\max(x) - \min(x) = 1$. The parameter β defines the free energy barrier (placed at $x = 0$) that separates U and F. The sign of β determines whether the region between the two minima is convex (positive barrier) or concave (downhill: unimodal); and, when positive, its magnitude determines the height of the free energy barrier ($\Delta G^{\ddagger}$). The ratio $\phi = |\alpha_U|/(|\alpha_U| + \alpha_F)$ defines the asymmetry of the free energy surface, with $\varphi = 0.5$ indicating a perfectly symmetric surface (i.e. barrier halfway between U and F). Therefore, the model can produce bimodal-unimodal one-dimensional free energy surfaces with shape (height of the barrier and curvatures of minima and barrier) that is controlled using these four parameters. To account for the changes in the free energy surface as a function of the (un)folding thermodynamic bias, we introduce the linear function $G_{bias}(x) = \lambda x$, where $\lambda \equiv G_{bias}(x_{max}) - G_{bias}(x_{min})$. The probability density as a function of the order parameter is hence:

$$p(x) = \frac{\exp(-(G_0(x) - \lambda x)/RT)}{\int \exp(-(G_0(x) - \lambda x)/RT)\, dx}$$

The kinetics of the process is described as diffusion on the free energy surface $G_0(x) - \lambda x$ as determined by the intramolecular diffusion coefficient *D*, which is assumed to be constant (i.e. independent of the order parameter). For convenience, we calculated the dynamics using a discretized version (discretized into *N*=121 microstates) of the free energy surface and the following diffusive rate matrix[125](section 2.4.3)

2.4.3 Folding kinetics along the free energy surface: Rate matrix formalism

We used a discretized version of the model where the potential of the mean force is calculated for *N* microstates and the interconversion rate between the species are determined by an intra molecular diffusion coefficient, *D* following the rate matrix

$$\mathbf{K} = \begin{pmatrix} -k_{1,2} & k_{2,1} & 0 & 0 & 0 \\ k_{1,2} & -(k_{2,1}+k_{2,3}) & k_{3,2} & \dots & 0 \\ 0 & k_{2,3} & \dots & k_{n-1,n-2} & 0 \\ 0 & \dots & k_{n-2,n-1} & -(k_{n-1,n-2}+k_{n-1,n}) & k_{n,n-1} \\ 0 & 0 & 0 & k_{n-1,n} & -k_{n,n-1} \end{pmatrix} \quad (4)$$

where the microscopic rates to go from the i^{th} microstate to the previous (i-1) and following (i+1) microstates are given by

$$k_{i,i-1} = \frac{D}{2}\left(\frac{p_i}{p_{i-1}} + 1\right)$$

$$k_{i,i+1} = \frac{D}{2}\left(\frac{p_i}{p_{i+1}} + 1\right).$$

p_i, p_{i-1}, p_{i+1} are the probability of i^{th}, $(i-1)^{th}$ and $(i+1)^{th}$ microstate respectively.

The rate matrix **K** is diagonalized to yield the eigen value spectrum, which gives the relaxation rates involved in the (un)folding dynamics. The slowest non-zero eigen value (λ_1) from the rate matrix corresponds to the overall relaxation time between folded and unfolded states ($k_f + k_u$). The second slowest non-zero eigenvalue (λ_2) represents the equilibration between the barrier top and the two minima, and hence it is closely related to the transition paths. It reflects the flux from either basin to the top rather than the crossing from one basin to the other. The third slowest relaxation rate (λ_3) gives information about the dynamics happening in each minimum (folded and unfolded). While the relaxation rates are well separated for proteins with a defined barrier, the difference between them gets increasingly blurred for downhill proteins which is an

indication of folding-unfolding rates being limited by the timescale of internal dynamics along the FES.

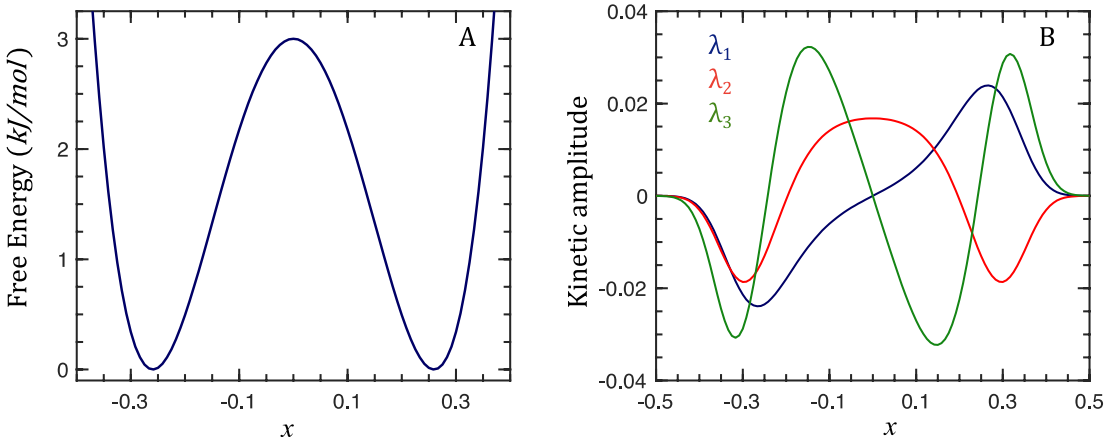


Figure 2-4 : (A) Landau free energy surface as a function of the order parameter *x*. (B) Kinetic amplitudes of the three slowest non-zero eigenvalues obtained from the rate matrix

Chapter 3

3 Deconstructing the terms in the folding rate equation and resolving (un)folding transition path of Engrailed-HD using high-resolution single-molecule Förster resonance energy transfer (SM-FRET) techniques

3.1 Engrailed Homeodomain

Engrailed is a transcription factor (TF) from *Drosophila melanogaster*. The DNA binding domain of this TF, engrailed homeodomain (En-HD) is a three-helix bundle with 62 amino acid residues. En-HD is a well-studied protein using both experiments and molecular dynamics simulations[83,91,128–130]. En-HD is known to show complex thermodynamic unfolding behaviour and kinetics. NMR studies on a destabilizing mutant of En-HD have characterized the presence of an intermediate and the complex multiexponential kinetic features have been attributed to the formation of an intermediate. Bulk experiments have found En-HD to be a fast folder that folds over a marginal free energy barrier or folding in a completely downhill scenario[130,131]. Long timescale MD simulations have demonstrated one-state downhill folding on En-HD[9]. This makes En-HD an ideal model protein to implement and test the performance of MLA with 1D-FES to analyse single-molecule photon trajectories and to study the complex folding of small single-domain proteins.

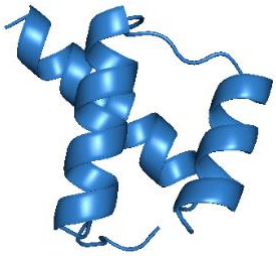


Figure 3-1 : Ribbon representation of En-HD structure

We are implementing GS-MLA (section 2.2) in combination with different FESs (section 2.4 and section 2.4.2) introduced in Chapter 2 to directly extract the FES for folding and the dynamics parameters from SM-FRET photon trajectories. We further utilize this information to resolve the transition paths and obtain a distribution of transition path times. A schematic representation of obtaining different terms involved in the rate expression and resolving transition path times according to our methodology is given in Figure 3-2

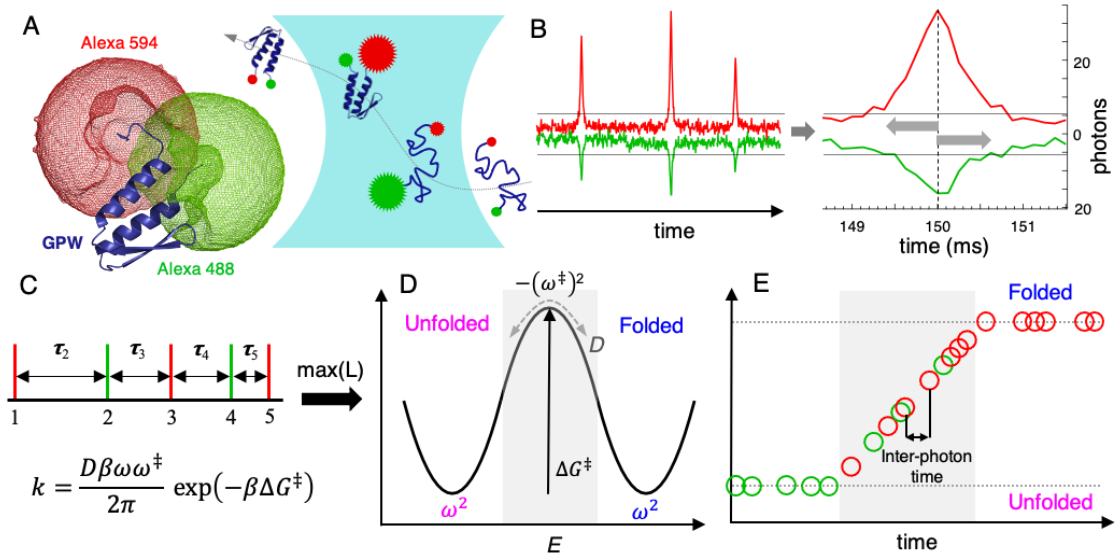


Figure 3-2 : Resolving folding dynamics from SM-FRET photon trajectories using ML analysis with 1D-FES model A) Fast folding protein labelled with FRET pair, and high count-rate free diffusion SM-FRET experiments performed on a confocal fluorescence microscope. B) Photon bursts are identified (0.2 ms binning) and then extended to include the entire diffusive trajectory through the confocal volume (section 2.1.5.1) C) Time-stamped photon trajectories are subjected to maximum likelihood analysis (MLA) and the 1D-FES model D) The FES (barrier, curvatures of wells and barrier top) and diffusion coefficient are determined by subsequent MLA of photon trajectories with a diffusive 1D free energy surface model. E) Photon trajectories are converted to most likely molecular state trajectories using the Viterbi algorithm. Transition paths are resolved if they are longer than the average inter-photon time (~4 μs). Figure reproduced from preprint (A. Sengupta, N. Mothi, M. Sadqi and V. Munoz 2021)

3.2 Sample preparation and SM-data collection

All reagents used in this study were obtained from Sigma-Aldrich or Thermo Fischer Scientific and were of high purity grade. The data discussed in this chapter were collected

in 20 mM acetate buffer with 100 mM NaCl at pH 5.5 and urea concentration of 3.3 M. This urea concentration was chosen (midpoint condition, C_m) to increase the probability of seeing folding-unfolding transitions. All buffers were treated with activated charcoal and filtered through 0.2 μM membrane filters. A photo protection cocktail was used to reduce photobleaching and other photophysical artifacts. Freshly prepared stocks were added to the buffer to have a final concentration of 1 mM (S)-Trolox methyl ether and 10 mM cysteamine. The pH of the buffer was maintained at pH 5.5 after the addition of a photoprotection cocktail, and the buffer was filtered thrice through 0.06 μM membrane filters to clean the sample and reduce scattering. A protein concentration of 20 pM was used for achieving a single-molecule condition.

All single-molecule experiments were performed at an excitation power of 150 μW (power was measured at the port of entry to the microscope) on the home-built SM-FRET set-up. Specifications of the optical set-up are explained in the appendix. 150 μW was chosen as the optimal power (maximum count rate was observed with minimal photophysical artifacts). To obtain a statistically relevant number of bursts, about 2 hours of SM data was collected. The concentration of protein is kept low (20 pM) for SM-FRET experiments under free diffusion conditions, so that the probability of a molecule diffusing through the confocal volume is ~2% (this ensures the probability of two molecules passing through the volume to an extremely low value of 0.04%). The protein sample was changed and replaced by a fresh sample every 3 minutes of data collection to minimize the effect of photobleaching and photooxidation due to prolonged exposure to excitation laser.

3.3 Thermodynamic stability check after labelling

Fluorophores used in the study are stable and well-characterized organic molecules that are widely used to fluorescently tag proteins. However, they have a single negative charge on them (structure given in the appendix) and hence it is important to characterize the thermodynamic stability of proteins before and after labelling to ensure the labelling hasn't affected the stability of the protein. Standard chemical denaturation experiments were carried out in bulk before and after labelling with extrinsic fluorophores. Intrinsic tryptophan fluorescence change was measured for unlabelled protein and change in FRET efficiency was monitored for labelled protein. Data was analysed using a two-state model for folding as described below.

$$\Delta G^{UN}([D]) = \Delta G^{UN}_{H_2O} - m[D] = m(C_m - [D])$$

$$K_{eq} = \exp(-\Delta G^{UN}([D]/RT))$$

$$p_N = 1/(1 + K_{eq})\,;\; p_U = 1 - p_N$$

$$< Signal([D]) > = S_U([D])(1 - p_N) + S_N([D])p_N$$

where ΔG^{UN} is the free energy difference between unfolded and folded state and is dependent on [D] (concentration of denaturant), $\Delta G^{UN}_{H_2O}$ is the conformational stability in the absence of denaturant, K_{eq} is the equilibrium constant, p_N and p_U are probability of folded and unfolded state respectively.

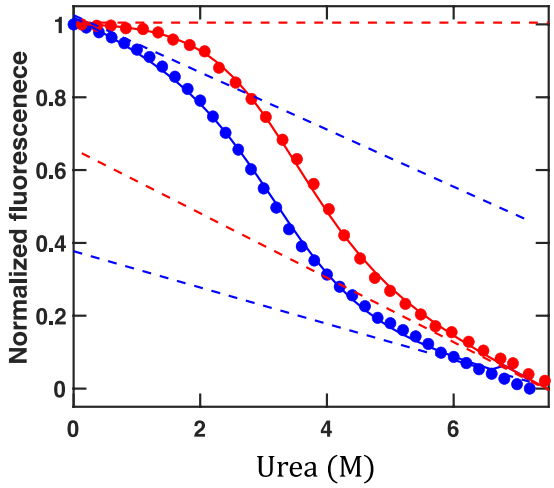


Figure 3-3: Chemical unfolding curves for En-HD. Experimental data for labelled (blue) is monitored using change in FRET and unlabelled (red) is monitored using change in intrinsic tryptophan fluorescence. C_m=3.3 M was obtained from the global fit (solid lines) of the experimental data (filled circles).

Even though analysis of individual unfolding curves separately makes it look like labelling has decreased the C_m and affected the thermodynamic stability of the protein, it is important to note that downhill folders are known to show complex probe dependent thermodynamics with probe dependent dynamics[132]. This is a result of the weak thermodynamic coupling between different regions of the protein and results in unfolding curves that differ depending on the nature of the structural probe being used. While tryptophan fluorescence is a measure of the change of environment around the tryptophan residue, FRET monitors the end-to-end distance of the protein and hence reports on a different structural aspect of the protein. Additionally, the absence of a well-defined baseline for the folded state complicates simple two-state analysis for proteins with marginal stability like En-HD. For this reason, a global analysis of these probe-dependent unfolding curves was done as it has been shown to represent the unfolding behaviour of downhill/marginal folders better. The solid lines in Figure 3-3 shows the fit obtained for the global analysis of data on labelled and unlabelled protein (filled circles). As it can be seen, there is great agreement between

the fit and the data. The global fit gave a urea concentration of 3.3 M as C_m and this concentration was used to collect SM-FRET data at midpoint denaturation condition.

SM-FRET experiments done in 3.3 M urea produced FEH with roughly equal populations of folded and unfolded ensemble (Figure 3-6). The observation that the midpoint denaturation concentration obtained from the global analysis of bulk fluorescence data matches with the C_m observed at single-molecule level demonstrates the strength of global analysis of experimental probes.

3.4 Confocal volume characterization and power dependence

Given that the fluctuation in the fluorescence intensity of a sample is only caused by diffusion through the effective volume in a confocal system, the amplitude of the cross/auto correlation function (G_0) equals the inverse of the number of fluorophores present in the effective volume (<N>). The confocal volume of the home-built SM-FRET set-up was measured by collecting FCS (fluorescence correlation spectroscopy) curves at various dilutions of a sample with known concentration. The average number of particles and concentration of the sample follows a linear dependence, and the slope of the graph is interpreted as the effective volume. The volume thus measured, V_{eff}, is not identical to the confocal volume, $\underline{V_{conf}}$. If approximated by a three-dimensional Gaussian function, the effective volume is larger than the confocal volume by a factor of $2^{3/2}$.

Atto-488 was chosen for the measurement as the dye has negligible triplet build up and hence correlation function will not be influenced by the triplet lifetime. The concentration range was chosen between the highest sample concentration that doesn't lead to saturation of detectors and the lowest concentration for which a correlation curve could be obtained without being overshadowed by noise. As expected, a linear dependence was found between the concentration and <N>. A confocal volume of 4.2 fL was measured for the SM-fluorescence optical set-up. This volume is larger than the volume observed for conventional commercial confocal microscopes (~1 fL). The reason for the larger confocal volume is the large pinhole (200 μm) used in our set-up to maximize the count rate. Single-molecule condition is still maintained throughout the measurements as the protein sample concentrations are kept to a low value of 20 pM. The optimal laser power was decided from an experiment, where a series of excitation powers (10-500 μW) was used to excite the doubly labelled protein sample and the emitted photons at individual channels were collected. The highest power at which a linear dependency was maintained between power and count rate was chosen for further experiments. In our SM instrumentation set-up, this corresponds to 150 μW.

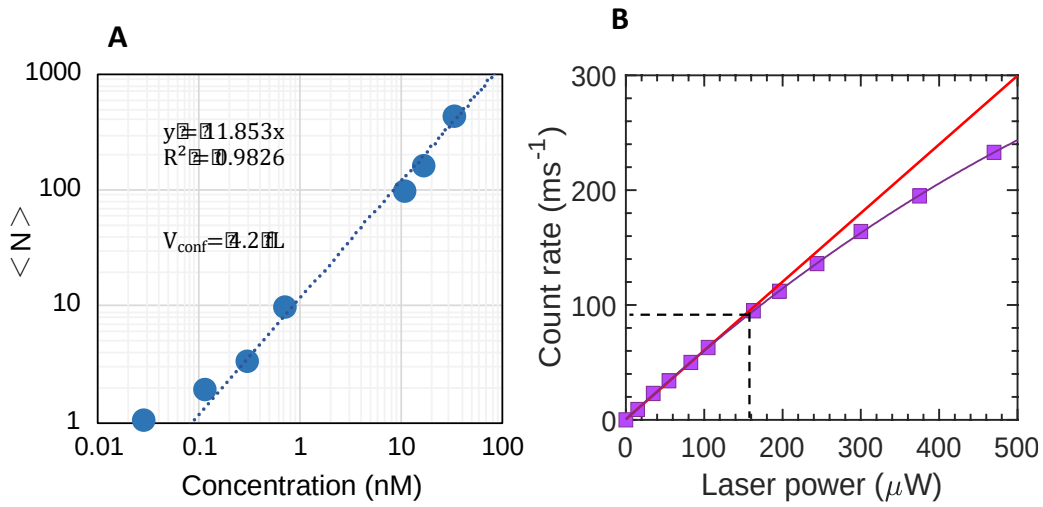


Figure 3-4 : (A) Dilution series of Atto-488 in water. Concentration of Atto-488 spans the concentration range between 40 nM to 25 pM. (B) Saturation curve of count rate as a function of laser power measured for 20 pM doubly labelled protein sample.

3.5 Burst identification and binned FRET efficiency histograms

SM-FRET for En-HD was collected at the urea denaturation midpoint of 3.3 M. FEH from time binning was generated (section 2.1.4) as a first step of the analysis to check for the quality of data and stability of protein samples. 1 ms binning time produced a unimodal and broad distribution suggesting dynamic exchange happening under that timescale (Figure 3-6A). A lower binning time of 150 μs was able to separate the unimodal distribution into two distinct populations (Figure 3-6B). This primary analysis puts the approximate timescale of (un)folding transition to be around 150 μs.

Time resolution of binning-based approaches are limited by the binning time. Burst analysis was carried out to remove the dynamic averaging arising due to binning. Photon bursts were identified from the long raw photon trajectories and the bursts were checked for photo-chemical artifacts according to the procedure explained in methods (section 2.1.5). ~6% of the total collected bursts were found to have such artifacts and were discarded. We collected about 11000 such bursts for En-HD to have a statistically significant representation of single-molecule events. FEH for these bursts is shown in Figure 3-5A. This FEH representation has dynamic averaging as the FRET was calculated for each burst by grouping photons together, but the data analysis for the rest of this study

was done on the photon trajectory directly and is devoid of such averaging. The mean residence time was found to be 601 µs allowing the possibility of observing multiple (un)folding transitions in some of the bursts (Figure 3-5B). The ML analysis was done on the time-stamped photon trajectory and hence is not affected by any kind of averaging.

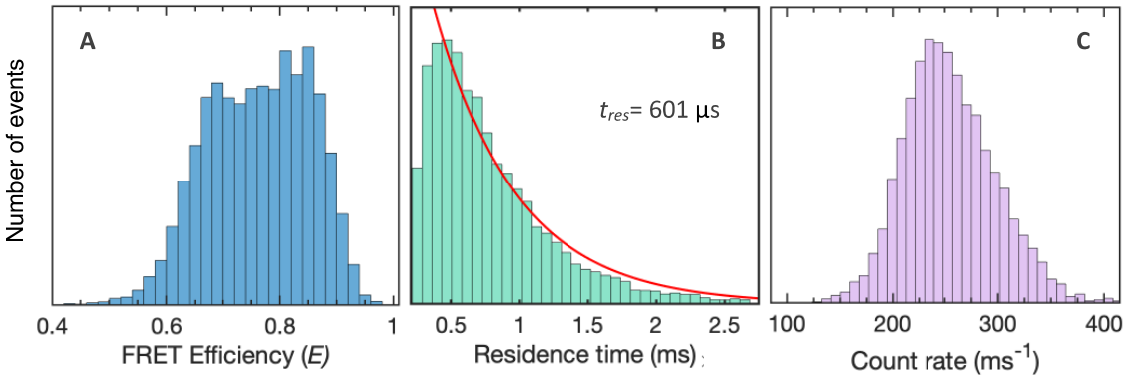


Figure 3-5 : Properties of bursts identified (A) Average FRET efficiency distribution for the bursts (B) Distribution of residence time, the effective observation time in the confocal volume in free diffusion SM-FRET experiments (C) Distribution of count rate

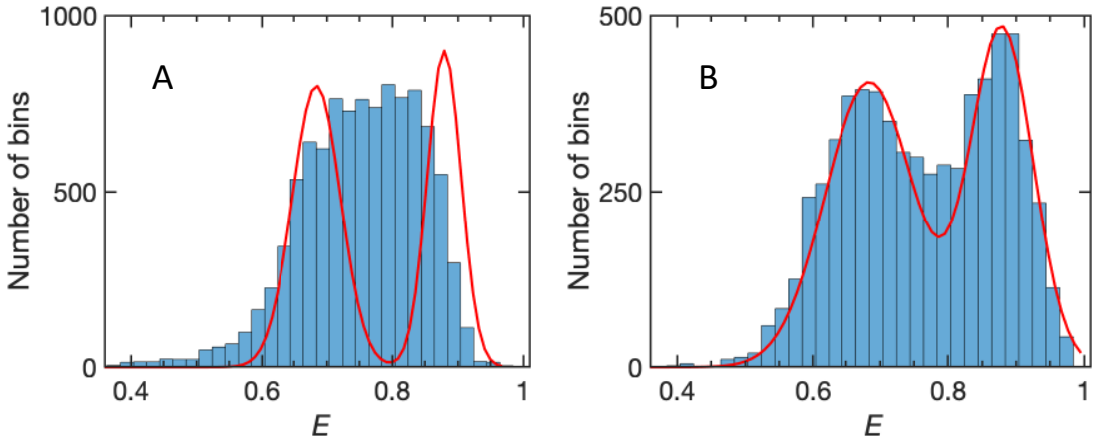


Figure 3-6 : FEH for binned SM-FRET data of En-HD in 3.3 M Urea (C_m) (A) 1 ms binning with photon threshold, N_T =160 photons (B) 150 µs binning, N_T =50 photons. Red curve shows the distribution around mean FRET if the source of broadening was solely shot noise. Broadening beyond the red curve suggests dynamic exchange between different populations under the binning time.

3.6 MLA with the two-state model for folding

Parameters for a two-state model of folding were obtained using the MLA method. A relaxation rate of 4801 s^{-1} was obtained (folding rate of 2335 s^{-1} and unfolding rate of 2466 s^{-1}) from the two-state description. This implies an average of 0.7 transitions per bin for the 150 µs binned histogram and agrees with the FEH which shows distribution broader than shot noise limited gaussian. The Viterbi algorithm was used according to the procedure explained by Chung and co-workers to generate the state trajectories from the photon trajectories and dwell time distributions of protein in folded and unfolded conformation was obtained (Figure 3-10)[122].

Folding and unfolding rate constants obtained from the dwell time distributions were found to be comparable with the rates observed from MLA analysis thus showing the ability of the Viterbi algorithm to identify transitions correctly. The underreporting of dwell time < 200 µs in the dwell time distributions is a result of the method adopted to identify bursts. Our algorithm identifies bursts of 200 µs or long (refer to section 2.1.5.1 for details). The relaxation rates obtained from the distribution are indicative of fast-folding dynamics previously observed for En-HD in bulk experiments and MD simulations.

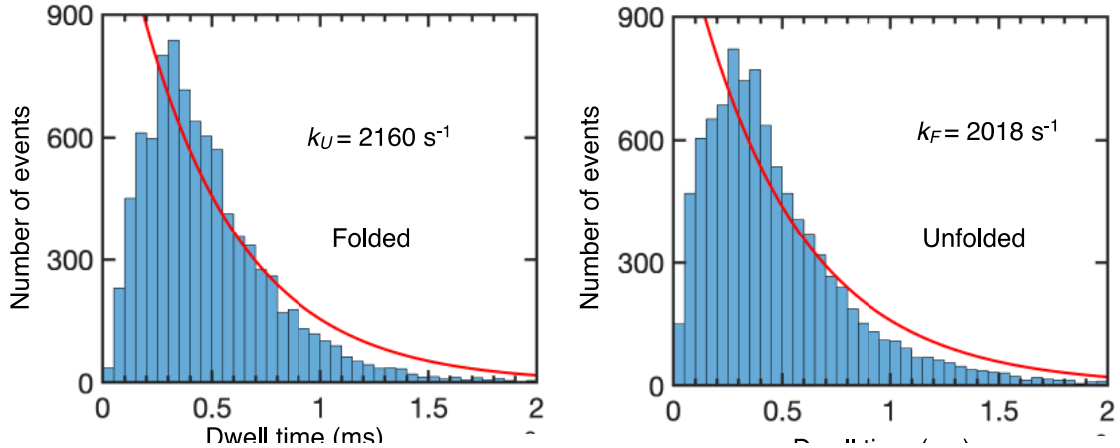


Figure 3-7 : Dwell time distribution in folded state and unfolded state obtained from the state trajectories generated using parameters from 2-state model.

It is important to note that the Viterbi algorithm generates the most likely molecular trajectory (according to the parameters for a model identified by MLA method) for each photon burst. As a result of this, trajectories are dominated by the most likely trajectories

and rare trajectories are overlooked (these are particularly interesting for folding mechanisms as these can point to the existence of heterogeneity in folding pathways).

3.7 MLA with the 1D-FES model

Performance of different models for describing a given data set can be done by performing MLA analysis with the models and comparing the likelihood values. We found the likelihood to be 192 log units better for the 1D-FES model than a simple 2-state model (statistical significance of $1\text{-}10^{-192}$) and 43 log units better than a 3-state kinetic model (statistical significance of $1\text{-}10^{-43}$). This improved performance of the 1D-FES model compared to the rigid two-state and three-state models indicates that the folding dynamics of En-HD is better captured by a model that has the flexibility to accommodate distributions of conformations along the folding reaction coordinate.

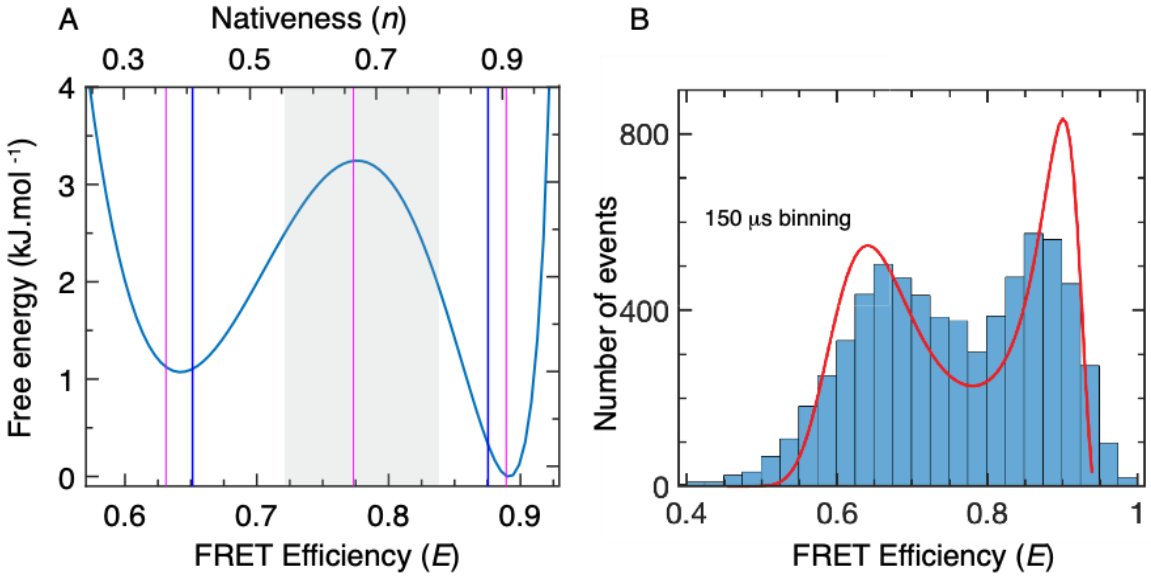


Figure 3-8 : Free energy surface and probability distribution from MLA with 1D-FES model. (A) FES obtained as a function of reaction coordinate, nativeness and *E*. Blue and magenta vertical lines indicate *E* of states identified by two-state and three-state model (B) Probability distribution from the 1D-FES model overlayed on 150 µs binning FEH.

The probability distribution from the 1D-FES model, which is the Boltzmann inversion of the free energy surface, captures the FEH produced from 150 µs binning (Figure 3-8B). This demonstrates that the model can identify and place the conformational ensembles

along the reaction coordinate with accuracy. FEH represents the dynamic population while the probability distribution from the model represents the static population distribution (without any dynamic averaging). Because of this difference, a complete overlap of the distributions is not expected as the FEH has broadening from dynamic exchange.

Free energy barriers of 0.84 k_BT and 1.2 k_BT were found for folding and unfolding, respectively, under the given conditions. The minimal free energy barrier offers the advantage of being able to see a significant population over the free energy barrier, in our case, 19% of the total population is along the transition region. This in turn ensures the observation of a statistically relevant number of folding-unfolding transitions, effectively making the resolution on transition path times possible, given that we have the required time resolution. MLA method has the advantage of time resolution being determined by interphoton times, which gives us a resolution of 4 µs events.

3.7.1 Comparison of the 1D-FES model and Landau free energy surface model

1D-FES model was compared with an empirical model where FES is approximated by a quartic Landau polynomial (section 2.4.2). The Landau FES model is a purely phenomenological and does not depend on the theoretical basis for the thermodynamic variables. This model is parametrically as simple as a three-state model, but it describes protein (un)folding kinetics as diffusion on a 1D free energy surface with adjustable barrier height and shape. Dynamics on both FES is described using the same rate matrix formalism explained earlier (section 2.4.3). The simple 1D-FES introduced in this chapter describe the free energy according to the principles derived from the energy landscape theory for

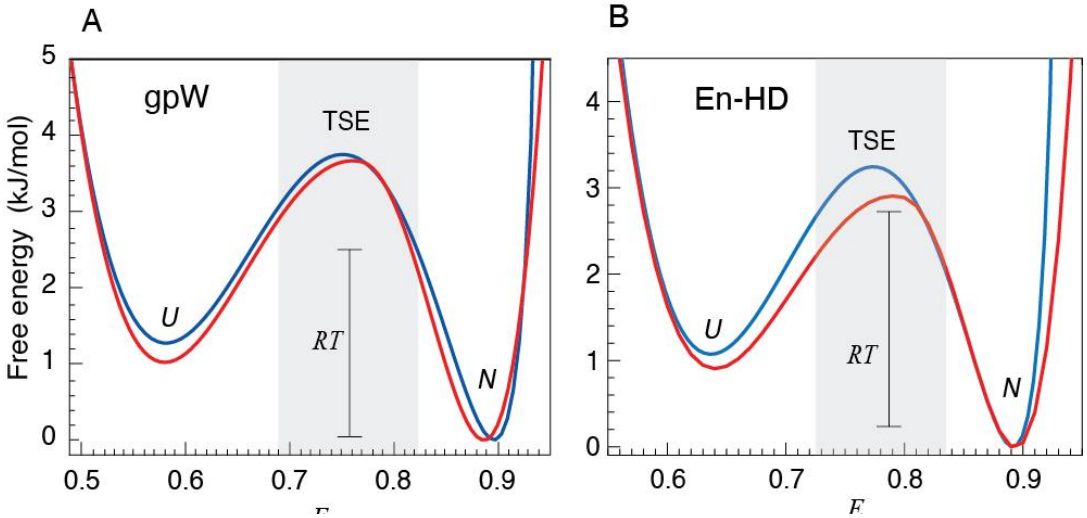


Figure 3-9 : Free energy surface obtained from the 1D-FES (blue) and the Landau FES model (red) for En-HD and gpW.

protein folding. Conformational entropy calculated from Gibb's entropy formula and stabilization energy calculated according to a mean-field description together makes up the total free energy, which scales with the total number of residues present in the protein (5 parameter). On the other hand, the Landau polynomial provides a purely geometric description of an FES (6 parameters) but with better flexibility.

FES obtained for En-HD from the simple 1D-FES model is comparable to the Landau FES, The Landau model however places the transition state closer to the folded state (Landau model produced a likelihood value 10 log units better than 1D-FES model, accounting for a statistical significance of $1\text{-}10^{-10}$). The general agreement between both the FESs shows the ability of 1D-FES to capture the folding behaviour by staying true to the principles of protein folding. Landau FES model is a more flexible model than the 1D-FES model. It can vary the barrier height and the position of transition state independently. FES obtained from both the models for another fast-folding protein, gpW (data for gpW is discussed in detail in chapter 0) is also shown for comparison. As seen from Figure 3-9A, FESs from the models are similar for gpW (likelihood value obtained for Landau model was only different by 1 log unit, accounting for a statistical significance of $1\text{-}10^{-1}$). The agreement between the FESs obtained from very different models validate the method of using MLA with FES based models for resolving protein folding dynamics.

3.7.2 Probing the coordinate dependence of diffusion coefficient in protein folding

1D-FES model defines folding along the discretized FES according to the timescale determined by an intra molecular diffusion coefficient, *D*. This diffusion coefficient accounts for the friction of the chain dynamics originating within the polypeptide chain. While Kramer's theory postulates that only the diffusion coefficient at the barrier top determines the folding rate constants, there is the question of whether the diffusion coefficient for a folding trajectory is dependent on the reaction coordinate. This becomes more of an important consideration for proteins folding over minimal or no free energy barrier as the diffusion coefficient along the entire reaction coordinate starts contributing to the overall rate[37].

We implemented a modified 1D-FES model with an extra parameter to account for varying diffusion coefficient along the FES. The diffusion coefficient was varied linearly between folded and unfolded ensembles. We found the diffusion coefficient to be larger for unfolded ensemble compared to the folded ensemble (log *D* decreased from 3.7 for unfolded ensemble to 1.7 for folded ensemble, thus showing a 100-fold difference in *D*). This significant decrease in *D* from unfolded to folded, was also observed in coarse-grained molecular simulations done to study the coordinate dependence of diffusion coefficient[37]. Performance of the model with varying diffusion coefficient was done by comparing the likelihood value, which was found to be 120 log units better than the model with a constant diffusion coefficient. This treatment demonstrates the potential of the 1D-FES model to

account for the coordinate dependence of intra molecular diffusion coefficient without adding significant complexity in terms of total parameters involved in defining the model.

3.7.3 Resolving transition path times.

Photon trajectories from SM-FRET experiments were converted into state trajectories using the Viterbi algorithm following the procedure by Chung and coworkers[122]. The Viterbi algorithm generates the most probable state trajectory resulting from the specific sequence of photons, inter-photon time and relaxation rates defined by the model. The result of the Viterbi procedure is a trajectory of different states allowed by the model corresponding to the trajectory of the photons. Essentially the algorithm determines the most likely state corresponding to each photon (Figure 3-10). Result of Viterbi procedure is the most likely molecular state trajectory (time trajectory of states allowed by the model) corresponding to each experimental burst (time-tagged photon trajectory).

State trajectories corresponding to the kinetics described by the two-state model, the three-state model and the 1D-FES model were generated. Trajectories from the two-state model produces a trajectory with molecules in either folded state (E_F = 0.87) or unfolded state (E_U = 0.65) and the ones following the three-state model produces a trajectory with the molecules in folded, unfolded, or intermediate state (E_F = 0.89, E_S = 0.77, and E_U = 0.63). For the 1D-FES model, the algorithm generates a trajectory with 101 states corresponding to each microstate with the indices of the states mapping to the nativeness along the free energy surface with 1st and 101th states corresponding to a nativeness of 0 and 1, respectively. The transition path was selected as part of the state trajectory where the protein molecule changes states from 56th through 80th state or vice versa. This maps to a change in nativeness of 0.56 through 0.80 (E_U = 0.687 and E_F = 0.824) which corresponds to the time spent crossing the free energy barrier on the 1D-FES obtained from the model (shaded region of FES in Figure 3-8). A distribution of transition path time for folding and unfolding transitions are obtained from these trajectories.

1D-FES inherently has the potential to resolve any part of the protein folding trajectory along the reaction coordinate due to the definition of protein as a diffusive process along the FES as opposed to the two-state and three-state model. This is evident from the examples of state trajectories shown in Figure 3-10. While the two-state model is limited to providing dwell time distributions in the folded and the unfolded state (k_f and k_u), and the three-state model can give the lifetime of the intermediate, 1D-FES can give information about conformational fluctuation happening in *F* and *U* ensemble along with a distribution of transition path times.

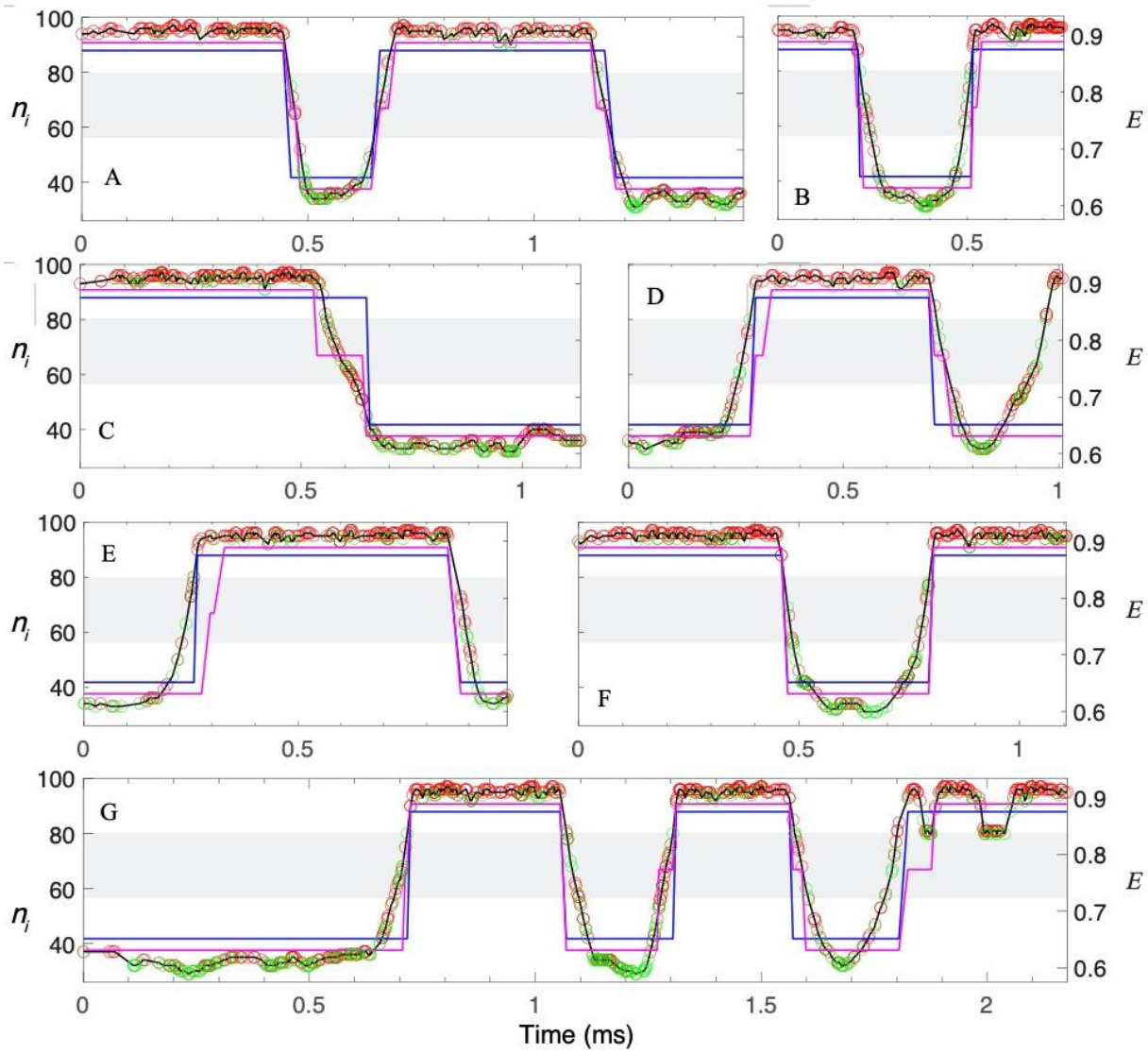


Figure 3-10 : Examples of state trajectories generated from the SM photon trajectories using Viterbi procedure according to different models. State trajectory corresponding to two-state, three-state and 1D-FES are represented by blue, magenta and black curve respectively. Green and red circles denote donor and acceptor photons and are overlayed on the 1D-FES state trajectory. Left y-axis represents the microstates corresponding to 1D-FES and right y-axis shows FRET efficiency. Shaded region shows the transition region defined for the molecular trajectory from 1D-FES model for measuring transition path time. Note that the duration of bursts follows an exponential distribution around a mean of 601 μs, hence a few long trajectories (e.g., G) are also observed.

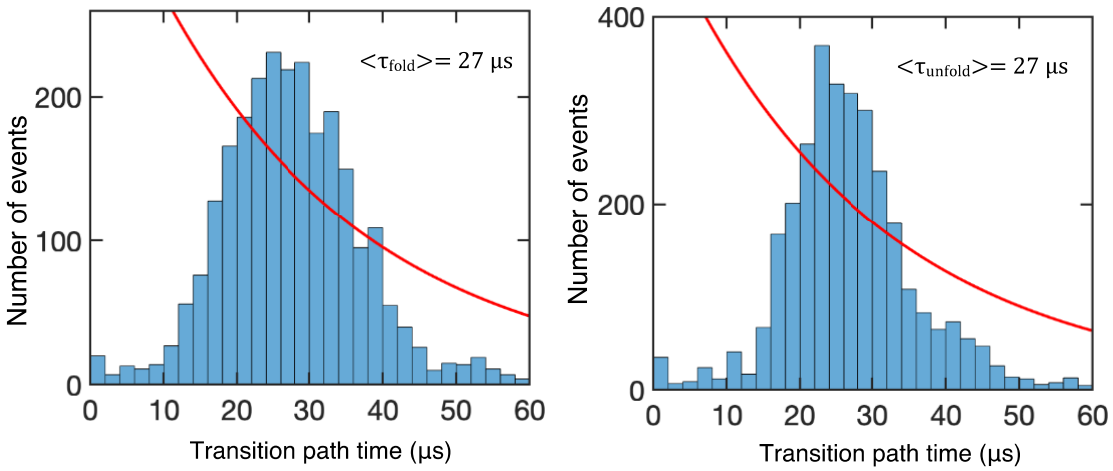


Figure 3-11 : Distribution of transition path time for folding and unfolding transitions. Transition path time was directly measured from the state trajectories generated using Viterbi algorithm corresponding to 1D-FES. Transition region was defined as the part of the trajectory shown by the shaded region along the trajectories in Figure 3-10.

We identified a total of 5623 (3016 folding transitions and 2517 unfolding transitions) transitions from the molecular state trajectories generated by Viterbi. The total number of photon bursts we utilized for the data set discussed in this chapter is ~11000 with a mean residence time of 601 µs. The high number of trajectories with transitions demonstrates the benefit of using fast-folding proteins for resolving TPT as observing statistically significant number of transitions under experimental condition is warranted. Resolving TPT of slow-folders using SM-FRET requires long data collection. They are analysed after a prior identification of regions with transitions (characterized by a sudden jump in FRET) to reduce the computational cost of analysing long photon trajectories. The comparable number of transitions observed for folding and unfolding transitions is a result of doing the experiments at denaturation midpoint condition.

We observed identical TPT for folding and unfolding transitions. This is expected as the TPT corresponds to the time spend on the transition region during successful barrier crossing events and is independent of the direction of the transition. Interestingly the mean of the distribution of the TPT obtained from our analysis, 27 µs, is relatively slower than the TPTs obtained for small single-domain proteins studied by SM-FRET till date (<10 µs)[47,65]. This timescale is of particular interest in the field as the TPT for small single-domain proteins with minimal or no barrier are utilized to predict the folding speed limit. Our observation of 27 µs TPT for a fast-folding protein with minimal barrier (~1 k_BT) indicates the roughness of the energy landscape and a higher internal friction for folding. Very few groups in the field have experimentally explored TPT estimation for fast-folding proteins. A common approach, based on comparing the likelihood of a two-state and three-

state description, is utilized for such measurements (section 2.3). We carried out similar analysis (Δlog $L_{2si\text{-}2s}$, magenta curve in Figure 3-12) and a variation of the method (Δlog $L_{3s\text{-}2s}$, blue curve in Figure 3-12). We found that the Δlog $L_{2si\text{-}2s}$ method underestimates the TPT (10 μs) while a variation of the method (it parametrizes the FRET efficiencies of two-state and three-state model used for the analysis differently) overestimated the TPT (58 μs). This difference points at the sensitivity of such methods on the mode of identification of FRET efficiencies. Additionally, the absolute Δlog L value at the peak provides a measure to compare the performance of different models. Δlog $L_{2si\text{-}2s}$ gave a peak of 47 log units (statistical significance of $1\text{-}10^{-47}$) and Δlog $L_{3s\text{-}2s}$ method showed a peak of 150 log units (significance of $1\text{-}10^{-150}$). MLA with 1D-FES produced a likelihood value 193 log units better than two-state model. The lack of flexibility in simple kinetic models to capture the distributions of transition region population is the reason for the inferior performance of such models. Chapter 4 dissects this problem further by comparing the performance of different methods used for TPT calculations (section 4.4.3) on simulated SM-FRET photon trajectories corresponding to a diverse set of folding scenarios.

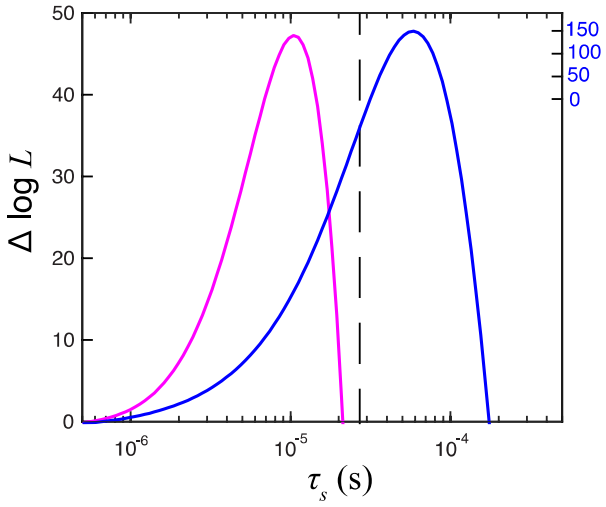


Figure 3-12 : Estimation of TPT from Δlog L curves. Δlog $L_{2si\text{-}2s}$ (magenta, left scale) and Δlog $L_{3s\text{-}2s}$ (blue, right scale) methods give TPT 10 μs and 58 μs respectively. The vertical dashed line signals the mean TPT obtained from the distribution in Figure 3-11.

3.8 Concluding remarks

All details about the conformational dynamics happening in a protein molecule within the interphoton time are embedded in the SM photon trajectory. The MLA method has the potential to extract these details, with the limitation being the model used to describe folding. We demonstrated the implementation of a free energy surface-based approach to extract the dynamic factors and the folding FES directly, thus minimizing the assumptions regarding the folding mechanism that is being introduced as part of the model. We also demonstrated the ability of the 1D-FES model being able to easily accommodate reaction-dependent diffusion coefficient, *D*, associated with using pair-distances as reaction coordinates[37,133]. Multiparametric FRET experiments can provide independent distance information between various parts of the protein. They can be utilized to address the question of *E* being an adequate reaction coordinate with the ability to capture the complex dynamics happening on the hyperdimensional landscape. We further resolved the TPT for a fast-folding protein using this approach. Interestingly, we found the TPT for a small single-domain protein to be longer than previously reported using other methods. Chapter 4 will explore this difference in TPT in more detail by utilizing simulated photon trajectories.

Chapter 4

4 Protein folding dynamics as diffusion on a free energy surface: rate equation terms, reactive transition paths and analysis of single-molecule photon trajectories

Theory suggests that protein molecules exploit a multitude of microscopic pathways during their search along the hyperdimensional energy landscape to get to the functional native conformation, resulting in a wide distribution of folding mechanisms[7,20,24,134–137]. An important approach to validate these results involves observing the structural evolution of protein molecules from one conformation to another individually as they undergo folding-unfolding transitions. The transition path time along a one-dimensional FES corresponds to the segment of the protein folding trajectory where the (un)folding transition takes place crossing the free energy barrier and contains invaluable information regarding the mechanism of self-assembly of protein molecules. TPT consists of the reactive excursions along the free energy barrier and includes information on the timescale of the local motions that drive folding reactions. Atomistic molecular dynamics simulations have successfully resolved protein folding mechanisms and have provided detailed mechanistic information about the folding process[9–11].

However, conventional bulk experiments are limited in their scope to study transition paths as these are properties of individual molecules and cannot be synchronized. Single-molecule techniques have been found helpful in studying these barrier crossing dynamics in proteins[65,92,138–141]. Recent advances in single-molecule Förster resonance energy transfer techniques and improved methods for the analysis of data have made it possible to obtain transition path times for multiple small proteins[51,54,142]. Traditionally protein folding models used for getting transition path times are based on two-state folding descriptions. Two-state folding behaviour implies a free energy barrier separating the native ensemble and the unfolded ensemble, which arises from the requirement of the barrier top being sufficiently high energy (>3 RT) so that the population distribution along the barrier is considered negligible. Experimental identification of downhill folders, proteins that undergo folding gradually without crossing a significant free energy barrier in a thermodynamically downhill process, challenges the validity of extending two-state-like kinetics to proteins indiscriminately[25,26,143] . Moreover, experiments have revealed that seemingly two-state folding proteins with small free energy barriers have tunable folding thermodynamics according to the experimental conditions[144,145]. This suggests that the use of models based on two-state-like descriptions can add inherent bias to the results of folding experiments. Maximum likelihood-based approach estimates TPT from the lifetime

of the intermediate in a three-state description of protein folding, which gives a better likelihood than the corresponding instantaneous transition model (two-state description). This analysis involves simplification of folding description into a simple two or three-state-based model. Careful evaluation of the effect of these approximations on the estimated TPT is crucial.

Furthermore, such simplistic approaches only report on mean TPT and cannot provide any meaningful information about the shape of the FES. While TPT is predicted to be insensitive to the height of the free energy barrier[52,142], the effect of the shape of the FES on the overall TPT is not included in such analysis. Direct analysis of transition path ensembles by measuring the TPT as a function of the shape and magnitude of the barrier can help dissect the role of shape of the FES and the barrier height in determining TPT. Simulated trajectories offer a unique opportunity to directly measure transition path time without making assumptions about the kinetics of folding. We perform a detailed investigation of the contribution of the free energy barrier and the curvature of the barrier in determining the overall transition path time. While single-molecule force microscopy methods have been utilized to obtain shapes of transition paths, free energy surfaces, and even energy landscapes[92,138,139,146,147], there have been very few attempts at getting the FES from SM-FRET photon trajectories[148]. We then test the performance of implementing GS-MLA in combination with simple kinetic models of protein folding by using a diverse set of simulated single-molecule photon trajectories corresponding to various folding scenarios.

4.1 Stochastic simulations of folding dynamics on Landau free energy surface

4.1.1 Landau one-dimensional free energy surface

The rate of a process described as diffusion on a free energy surface projected onto a single order parameter can be described using a Kramers-like rate expression:

$$k = \left(\frac{D\beta\omega\omega^{\ddagger}}{2\pi}\right)\exp(-\beta\Delta G^{\ddagger}) \quad \text{equation (1)}$$

where D is the intramolecular diffusion coefficient, ω^2 is the curvature of the well (e.g. folded and unfolded), $-(\omega^{\ddagger})^2$ is the inverted curvature at the barrier top, and $\Delta G^{\ddagger}$ is the free energy barrier on the projected free energy surface. The terms in such rate expression are empirical (not derivable from first principles), and they all contain key system-specific information about the underlying mechanisms. One way to experimentally access such information is through methods that can resolve the reactive transition paths[65]. To better understand the relationships between these rate terms and the transition paths we carried

stochastic kinetic simulations on 1D free energy surfaces with different shapes. In this respect, the Landau free energy surface model that we introduced earlier (section 2.4.2) is particularly convenient, because it permits to control the barrier and shape of the free energy surface through specific parameters. In addition, we set all of the free energy surfaces in this study to iso-stability conditions to ensure that the barrier height is identical in the folding and unfolding directions and hence facilitate comparison. Particularly, we generated free energy surfaces with barrier heights ranging from 10 k_BT to a minimalistic barrier of 0.4 k_BT, that is, a range that covers the barriers estimated for two-state single-domain proteins[149] and fast folders[78,150] (Fig. 1A). Importantly, these surfaces have exactly the same shape, with the barrier placed at 0.65 along the reaction coordinate (ϕ = 0.65); a position that is consistent with the experimental average phi-value of over 800 mutations on 24 single-domain proteins[87]. We also generated free energy surfaces with different levels of asymmetry (relative position of the barrier and curvature of the minima) (Fig. 1B). The asymmetry range that we explore here (ϕ from 0.5 to 0.9) mimics the range in β_T values observed on two-state folding proteins as determined from the ratio of the slopes of the chevron plot[151].

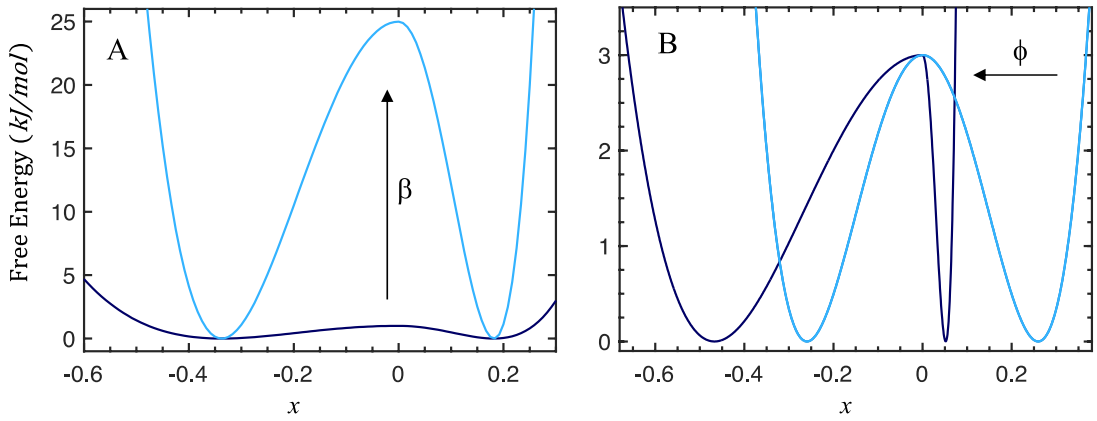


Figure 4-1 : The Landau 1D free energy surfaces used in this study. Left) 1D free energy surfaces corresponding to the two extremes of the range we used to investigate the relationship between barrier height and transition paths. The light blue surface represents the highest barrier we explored (10 k_BT), and the navy blue the lowest (0.4 k_BT). All the free energy surfaces are generated with ϕ = 0.65 (same shape). Right) 1D free energy surfaces corresponding to the two extremes of the range we used to investigate the effects of the position of the barrier along the reaction coordinate. The light blue represents a surface with ϕ = 0.5 (symmetric) and the navy blue a surface with ϕ = 0.9 (highest asymmetry).

4.1.2 Stochastic kinetic simulations.

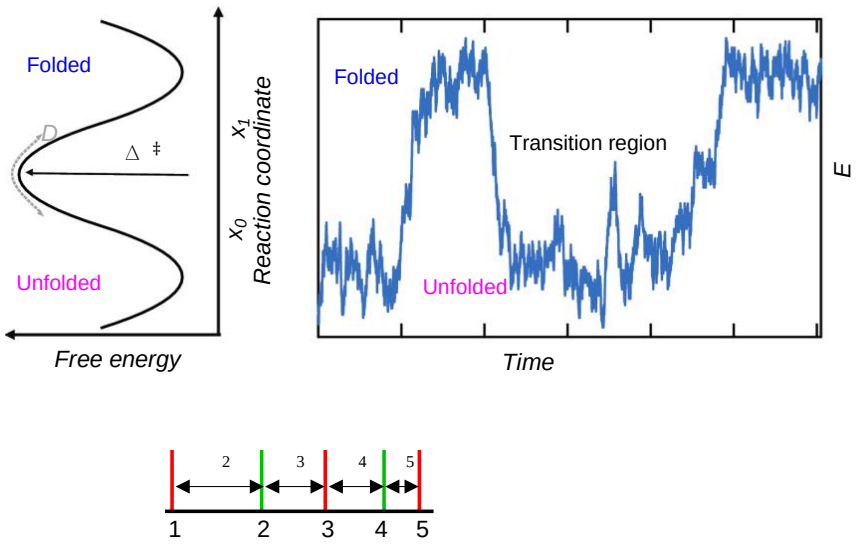


Figure 4-2 : Schematic of stochastic simulations of folding dynamics on Landau FES (A) Landau FES is generated (B) Molecular trajectory for free diffusive protein folding on the FES is generated (C) Photon trajectories corresponding to the molecular trajectory is generated from FRET efficiency and an exponential distribution of interphoton times

We carried out stochastic diffusive kinetic simulations on the discretized version of the free energy surfaces. Particularly, we used a grid of 121 points, which we confirmed was sufficient resolution to smoothly reproduce the shape of the surface (even for the most asymmetric ones). The simulations were performed with a version of the Gillespie algorithm[152,153] at constant time intervals, following the same procedure explained before[125]. Briefly, we defined three possible outcomes starting from microstate *i* after an elapse of time Δt: moving forward to i+1, moving backwards to i-1 or staying at i. According to this move set, the time-dependent probability of each of the microstates is obtained from the relationships

$$p(i \to i+1) = \Delta t \frac{1}{2}\left(\frac{p_{i+1}}{p_i} D + D\right)$$

$$p(i \to i-1) = \Delta t \frac{1}{2}\left(\frac{p_{i-1}}{p_i} D + D\right)$$

$$p(i \to i) = 1 - [p(i \to i-1) + p(i \to i+1)]$$

when $\Delta t \ll (\Delta x)^2/D$. The timescale for moving along the surface is determined by the intramolecular diffusion coefficient *D* and the free energy gradient between neighbouring microstates.

Simulations with increasing barrier height were performed using proportionally faster *D* and smaller Δ*t* to keep the total simulation time constant (20 s) while ensuring we could obtain sufficient sampling statistics, and enough number of transitions per burst on the photon trajectory simulations. The parameters used for each stochastic simulation are given in Table 1.

$\Delta G^{\ddagger}$ (kJ·mol^{-1})	**log *D*** ($x^2 \cdot s^{-1}$)	λ_1 (s^{-1})	**Δ*t*** (ns)	**Count-rate** (ms^{-1})
1	2.08	4,377	20	221
3	2.08	3,648	20	221
6	2.38	3,741	8	500
9	2.68	3,288	3	1000
15	2.78	620.5	1	1250
20	2.91	148.4	0.5	1764
25	3.00	30.2	0.2	****

Table 1 : Parameters used for the stochastic kinetic simulations of the free energy surfaces with barriers of different height, and for the simulations of time-stamped photon trajectories. ϕ = 0.65 was used for all simulations to change the barrier while keeping the exact same shape of the free energy surface.

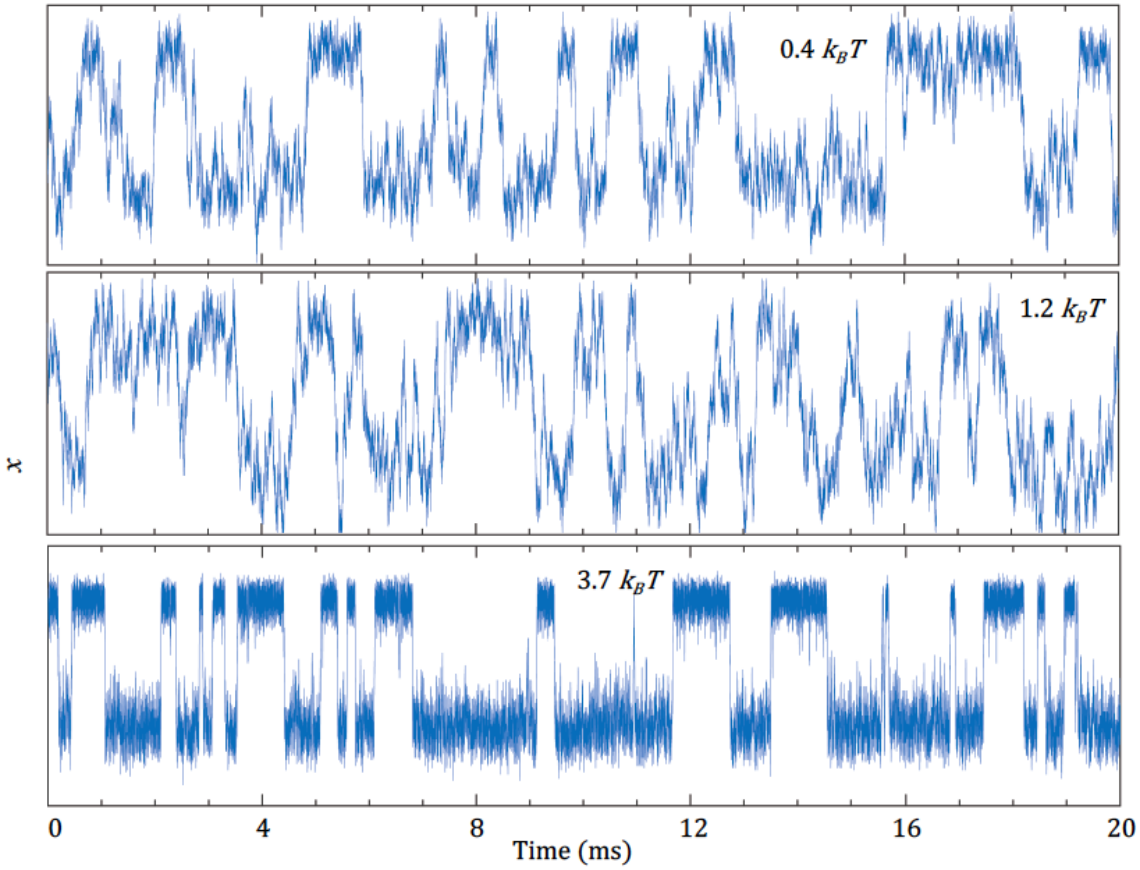


Figure 4-3 : Examples of molecular trajectories generated. 20 ms segment of the full-length trajectory (20 s)

4.2 Analysis of transition paths from stochastic kinetic trajectories.

For condensed phase reactions, such as those involving proteins and other biomolecules, a transition path is the segment of a reactive diffusive trajectory ($U \rightarrow F$ or vice versa) that corresponds to the actual crossing of the barrier separating the two minima. Such definition, of course, makes transition paths dependent on how one defines the effective transition region, which is somewhat arbitrary, even when the free energy surface is known *a priori*. This problem is related to the better studied one of identifying the transition state ensemble (TSE). In this regard, the TSE is considered the set of points in configuration space that share an equal probability of reaching the reactant region and the product region when propagated forward in time. An alternative definition explored by Hummer, builds upon the concept of TSE regions as localities common to most reactive trajectories, and hence as those points in configuration space with the highest probability that equilibrium trajectories passing through them are indeed reactive[52]. Transition paths are then identified as trajectory segments that exit from a reactant region and reach a product region without crossing back into the initial reactant region, and vice versa.

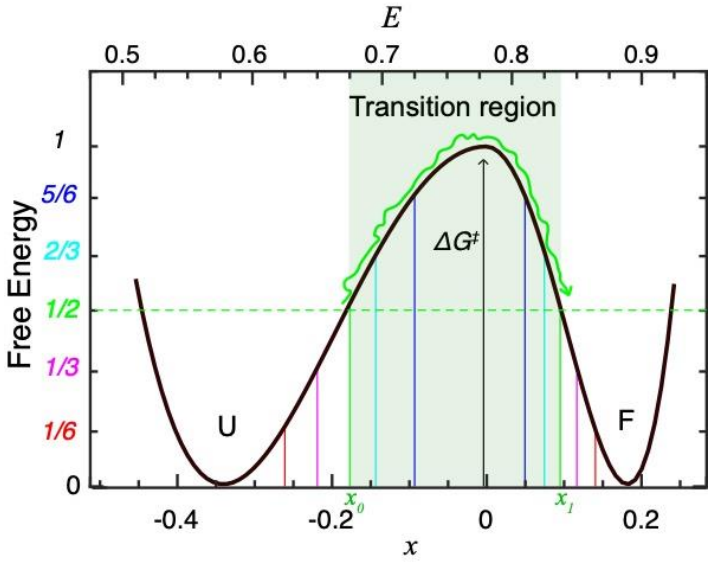


Figure 4-4 : Boundaries for the transition region. The transition region is defined by its boundaries (x_0 and x_1) on the reaction coordinate. Here five examples are given in which the position of x_0 and x_1 are determined as a fraction of the distance between the minimum and the barrier top (x = 0). The halfway definition of the transition region is highlighted as a green swath with a schematic representation of a possible transition path as example. The diagram shows the reaction coordinate (x) on the bottom, and the corresponding changes in FRET efficiency for the simulations of photon trajectories on top

Here we used the Hummer approach to simulate transition paths. Namely, we performed stochastic kinetic simulations on the 1D free energy surface (assuming it is projected on a good reaction coordinate), starting from a point on the reactant basin of attraction that is on route to the barrier (x_0), and propagating the trajectory forward in time. We then identified the trajectories that successfully traversed the transition region and crossed over to an equivalent point on the product basin of attraction (x_1) without revisiting x_0. Similar simulations were carried out in the opposite direction (reversing x_0 and x_1). This procedure is straightforward for a known 1D free energy surface, but it still requires that the boundaries of the transition region ($[x_0, x_1]$) are defined. We thus carried out simulations for different definitions of the transition region, that is, different $[x_0, x_1]$, but ensuring symmetry so that the relative distance from the minimum to the barrier was the same on both sides (Figure 4-4).

For each definition, thousands of trajectories were generated. Analysis of these trajectories rendered the distribution of transition path times in each direction (folding and unfolding) as well as the commitment probability, which we define as the ratio of reactive transitions relative to the total number of trajectory trials. For comparative purposes, the transition paths for each simulation were rescaled by the ratio between a reference diffusion coefficient and the diffusion coefficient used in the specific simulation. Transition path times were then expressed relative to the transition path time obtained for the halfway definition of the transition region and on the free energy surface with the highest barrier (ϕ = 0.65 and 10 k_BT).

4.3 Simulating time-stamped photon trajectories.

We simulated the outcome of single-molecule FRET experiments by performing stochastic simulations of donor/acceptor photon emission events as a function of time and the position of the molecule diffusing on the 1D free energy surface. The donor and acceptor emission rates were determined by the "experimental" count rate, which we set within realistic values for current SM-FRET experiments, and the probability of emitting a donor or an acceptor photon, as defined by the FRET efficiency (E) of each microstate on the 1D surface ($p_A=E$ and $p_D=1-E$). For simulations on the surfaces with the lowest barriers, we used the same count-rate we obtained experimentally on the protein *gpW*, which folds/unfolds over a barrier of ~1 RT[143]. For simulations with higher barriers, we progressively increased the diffusion coefficient (D) to obtain approximately the same overall relaxation rate (similar numbers of folding/unfolding events in the simulated trajectory), and we adjusted Δt (diffusive simulations, see the previous section) and photon count-rate accordingly to ensure comparable conditions (all the relevant parameters are given in Table 1).

We followed the same procedure we used before[125] to generate a stochastic trajectory of inter-photon times according to an exponential distribution with $\mu = 1/CR$, where CR is the total photon (A+D) count-rate. After generating the photon emission times, each photon

was then coloured (assigned an acceptor or donor tag) based on probabilities defined by the FRET efficiency of the microstate that the molecule was occupying at the exact time the photon was emitted. The result of this procedure is a strip of time-stamped donor/acceptor photons emitted stochastically by the molecule as it is moving diffusively along the 1D free energy surface during the entire simulation. This long strip of photons (4-25 million) was then divided into segments of 300 to 2,000 photons to represent the number of photons obtained per burst/trajectory that are typically obtained in single-molecule FRET experiments.

To determine the FRET efficiency as a function of the order parameter (*E(x)),* we used a linear function defined according to the empirical parameters E_U and E_F, corresponding to the FRET efficiency values at α_U (unfolded minimum) and α_F (folded minimum), respectively. The FRET efficiency as a function of the order parameter was then simply defined as,

$$E(x) = E_{TS} - x \text{ for } x \leq 0$$

$$E(x) = E_{TS} + x \text{ for } x > 0$$

where $E_{TS} = E_U + \phi(E_F - E_U)$ is the FRET efficiency at the barrier maximum ($x = 0$). For all the photon trajectory simulations in this dissertation we used $E_U = 0.574$ and $E_F = 0.885$, which are the *E* values we obtained from the analysis of SM-FRET experiments on the fast folder gpW. These values are also in good agreement with values reported by other authors on single-domain proteins at their chemical denaturation midpoint[47,51,96,154]

Count rate and *D* for lower barrier simulations were chosen according to the corresponding parameters obtained on analysis of experimental SM-FRET data on protein, *gpW*, with a 1-dimensional FES model. Count rates for simulations with higher barriers were chosen to be able to resolve the fastest exchange happening during the folding transition and are given in Table 1.

4.4 Direct measurement of transition path times and commitment probabilities

The analysis of transition paths relies on the definition of its boundaries ($[x_0, x_1]$), which is somewhat arbitrary. A simple strategy, if the free energy surface is known, is to set the transition region boundaries halfway between the reactant minimum and the barrier top; but there are no particularly strong arguments for using such definition beyond its convenience. However, the actual boundaries should have a significant impact on both the average time it takes a molecule to traverse the path and its commitment probability. We have explored this issue performing stochastic kinetic simulations with different

boundaries as shown in Figure 4-4 and on all the free energy surfaces of this study. Figure 4-5 shows how the average transition path time and commitment probability change depending on the transition region boundaries for an exemplary free energy surface (i.e. ϕ = 0.65 and 1.2 k_BT).

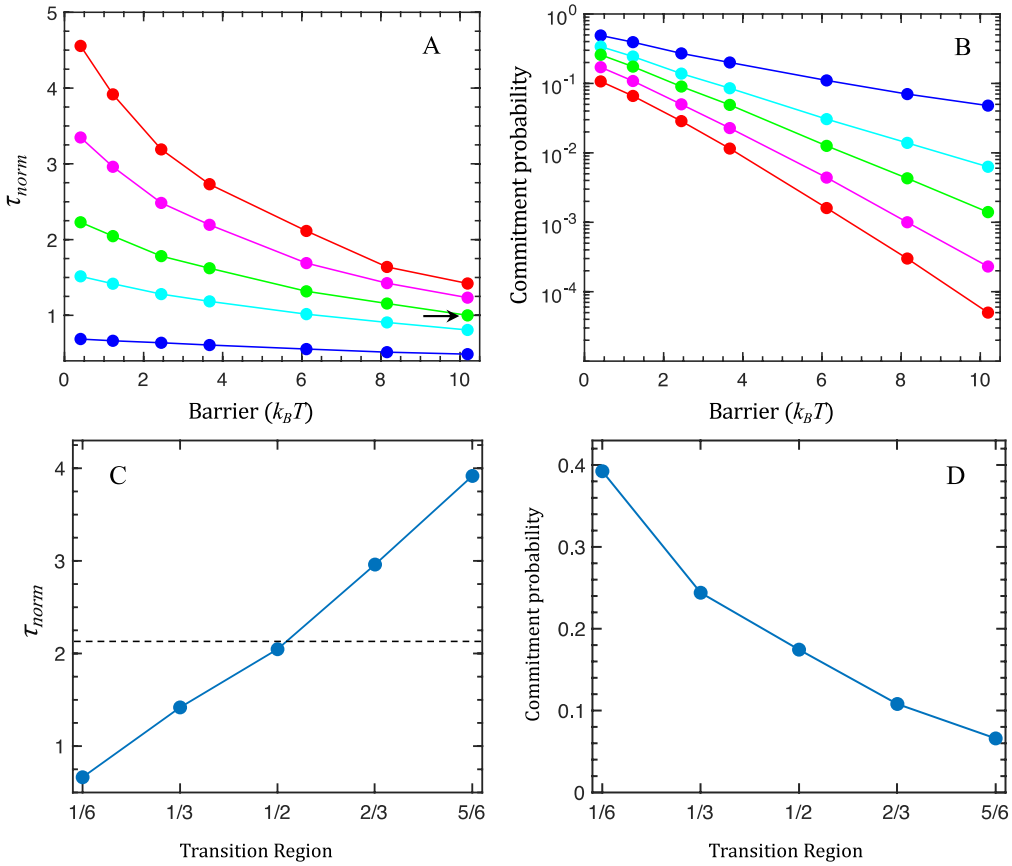


Figure 4-5 : TPT (A) and commitment probabilities (B) for different definitions of transition region following the colour code in Figure 4-4 as a function of free energy barrier. Dependence of the TPT (C) and commitment probability (D) on the boundaries for the transition region for barrier of 1.2 k_BT. The dashed horizontal line indicates the inverse of the second slowest non-zero eigenvalue of the rate matrix ($1/\lambda_2$).

As expected, the average transition path time (TPT) is strongly influenced by the definition of transition region, changing by nearly 6-fold between a narrow region (5/6) and a broad region (1/6). The same trend was found on all other free energy surfaces. However, we also observed that the sensitivity to the transition region boundaries is less marked at higher barriers, being reduced to about 3-fold for the 10 k_BT barrier. This result is consistent with previous theoretical studies which found the shape of the transition path to be roughly

insensitive to its precise definition for transitions that take place over barriers that are significantly higher than k_BT[155]. The commitment probability follows the inverse behaviour, increasing as the transition region becomes narrower. This is again expected as longer paths necessarily involve many more chances for the molecule to revert its trajectory. Notably, as the region becomes narrower, the commitment probability approaches 0.5, which is the value corresponding to the TSE (barrier top) in transition state theory[156].

These results highlight the importance of the choice of boundaries for the analysis of transition paths, particularly for the relatively small folding-unfolding free energy barriers expected for single-domain proteins[157]. The analysis of the eigen spectra of the diffusive rate matrix offers one useful point of reference. As explained in methods (section 2.4.3), the second slowest non-zero eigenvalue (λ_2) is closely related to the transition paths. As shown in Figure 4-5C, the inverse of this eigenvalue is very similar to the TPT calculated with the halfway transition region. However, λ_2 reflects the flux from either basin to the top rather than the crossing from one basin to the other. We can hence conclude that $1/\lambda_2$ represents a lower bound for the TPT, and therefore, that the transition region should ideally be somewhat broader than halfway. An alternative, more empirical way to address this question is to look for the transition region boundaries that produce the best scaling between the ratio $\tau_{kin}/(2\pi\langle\tau_{TP}\rangle)$ and the true barrier over the range of barriers that is most significant for protein (un)folding[47]. Here τ_{kin} is the overall relaxation time ($1/\lambda_1$), and $\langle\tau_{TP}\rangle$ is the TPT measured with the different transition region boundaries. This calculation for the data shows that the definition of transition region boundaries has negligible effect for the highest barriers. On the other hand, this calculation overestimates the height of the smaller barriers in general, and increasingly so the narrower are the boundaries of the transition region. For instance, the TPT calculated with the narrowest transition region (5/6) estimates a barrier of ~ 3 k_BT for the surface with a 0.4 k_BT barrier. From all these considerations, we decided to use the 1/3 transition region as the most accurate definition for the analysis of transition paths in the range of free energy barriers that we explore here.

4.4.1 Effects of the free energy barrier on the transition paths

The relationship between the transition path time and the free energy barrier has been subject of interest in recent years, with focus on developing procedures to estimate the height of the barrier using experimental measurements of the average transition path time [47]. Such analysis typically uses an analytical expression developed by A. Szabo[52,142] for a harmonic barrier > $2k_BT$, in which the average transition path time is approximated as

$$\langle\tau_{TP}\rangle \approx \ln\left[2e^{\gamma}\beta\Delta G^{\ddagger}\right]/D\beta\left(\omega^{\ddagger}\right)^2 \qquad \text{equation (2)}$$

where γ is Euler's constant (0.577…) and $(\omega^{\ddagger})^2$ is the curvature of the barrier. This expression indicates that the transition path increases with the barrier height (at constant curvature), but it does so weakly: with the logarithm of the barrier in contrast to the exponential dependence of the rate (equation 1). Based on this expression, one can estimate the barrier height from the rate and average transition path time (see previous section) by assuming that the diffusion coefficient is constant along the free energy surface and that the curvatures for both minima and barrier are similar.

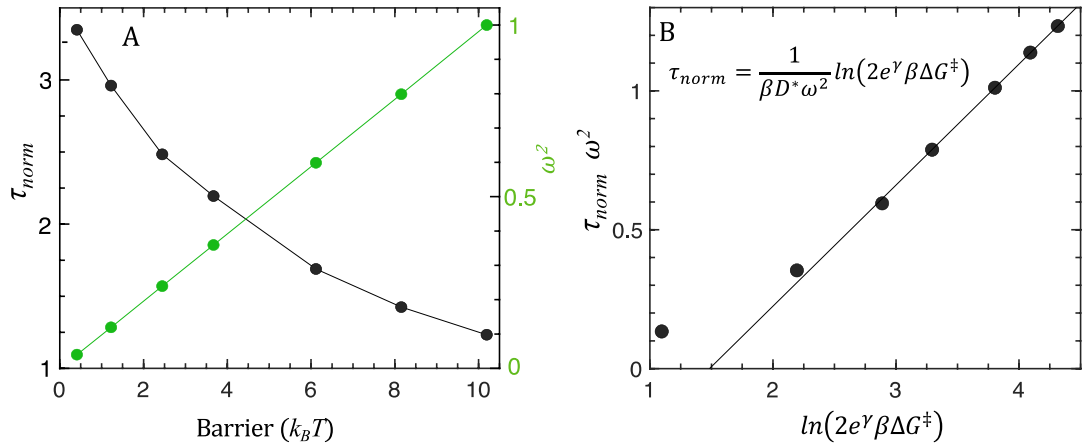


Figure 4-6 : A) Transition path time (black, left axis) and the curvature for free energy barrier (green, right axis) as a function of free energy barrier. The barrier curvature is shown normalized with respect to the highest barrier used in this study. B) Barrier dependence of the average transition path time multiplied by the barrier curvature

Our simulations for barriers that range from 10 to 0.4 k_BT show that the transition path time becomes progressively longer as the barrier decreases (Figure 4-5). The increase is quite significant, as the TPT is about 2.5-fold longer for the marginal barrier than for the highest. The data shown in Figure 4-6 corresponds to the 1/3 transition region, but we obtained similar results for other transition region boundaries. At first glance, this result seems counterintuitive, and the exact opposite of the trend predicted by equation 2. The analysis of the commitment probabilities sheds some light into the source of this behaviour. The commitment probability reflects how many of the attempts to cross the barrier are reactive. This probability decreases for broader transition regions (Figure 4-5, panels B and D) because the longer is the actual path, the higher is the probability that the molecule undergoes a trajectory reversal at some point. This effect is amplified by the gradient of the free energy surface, and hence, the commitment probability drops dramatically when the barrier raises (i.e., about 3 orders of magnitude from 0.4 to 10 k_BT). In other words, a steeper free energy gradient enhances the probability of a trajectory reversal occurring

during the climbing to the barrier top. This effect makes most trial trajectories unproductive, resulting on a slowdown of the rate. But, by the same token, the few trajectories that do indeed make it over the top are those that do not linger on the path, that is, those that minimize the transition path time.

From a practical standpoint, the key factor behind the apparent discrepancy between the results in Figure 4-6 and equation 2 is the curvature of the barrier. The free energy surfaces used in this analysis have identical shape: same position for the minima and the barrier top. However, barrier height and curvature are tightly interconnected and, therefore, as the barrier height increases so does its curvature: by ~20 fold between a barrier of 0.4 and one of 10 k_BT (green in Figure 4-6A). Figure 4-6B shows that, when the computed average transition path times are multiplied by the barrier curvature of the corresponding surfaces, the data follows equation 2 closely for barriers higher than 2 k_BT (third point in Figure 4-6B). The transition path time does diverge from equation 2 for the smallest barriers. Similar conclusions have been reached from theoretical analyses of the shape of transition paths[158,159]. These results indicate that faster folding proteins are also likely to exhibit longer transition paths, a conclusion of practical significance for the experimental analysis of protein (un)folding rates and transition paths. That is, the best candidates for the single-molecule analysis of folding transition paths are fast folding proteins, which not only undergo many more reactive transitions per unit of time, but their transition paths are longer and thus easier to resolve.

4.4.2 Asymmetry of the free energy surface and transition paths

Another important issue is whether transition paths contain useful information about the shape of the free energy surface. Theoretical analysis indicates that it is possible to extract the shape and height of the free energy surface from the shape of the transition paths, provided that they are resolved[155,159]. On the other hand, existing methods for the analysis of single-molecule experiments only produce the average transition path time, and hence need an independent estimate of the free energy surface[65,93]. Here we further explore the connection between the shape of the free energy surface and the transition path times using surfaces with different levels of asymmetry. The studied surfaces range from a perfectly symmetric one to a surface in which the barrier is at 90% of the total distance between *U* and *F* (Figure 4-1B). We then generated transition paths in the folding and unfolding directions (Figure 4-7). As expected from time reversal symmetry, the average transition path time is the same in both the forward (folding) and reverse (unfolding) directions. We also see the average transition path time is not sensitive to the asymmetry of the free energy surface, as it has been theoretically proposed[155,159]

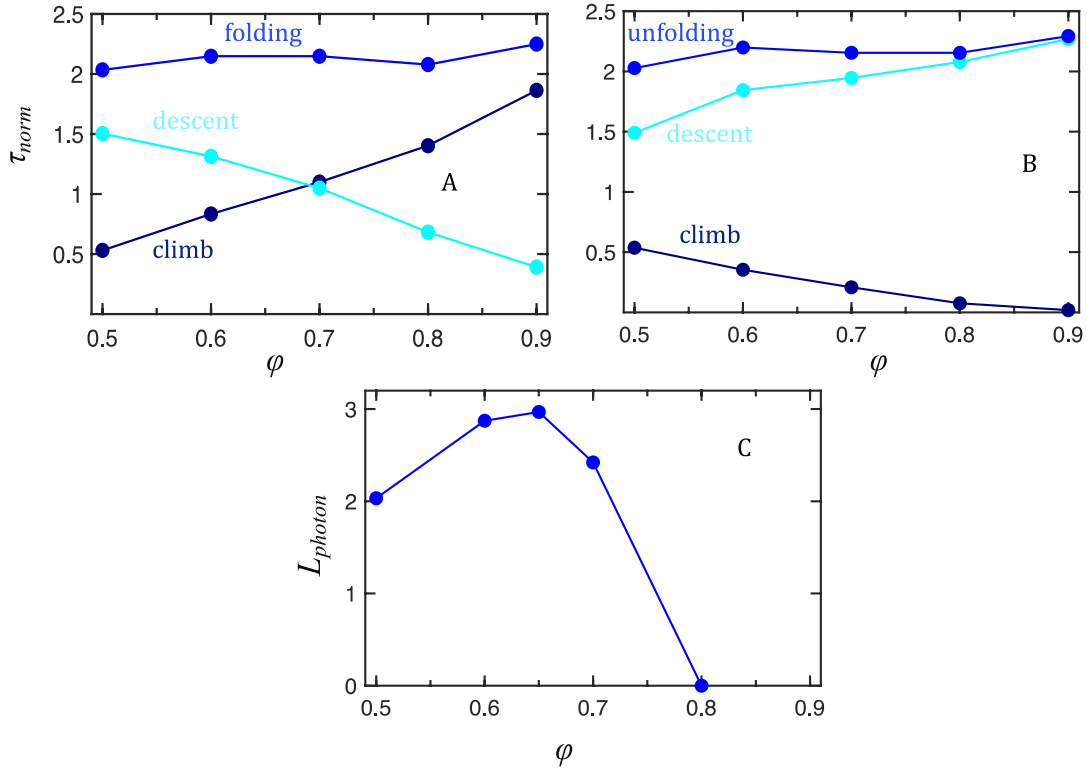


Figure 4-7 : Transition paths as a function of the asymmetry of the free energy surface (ϕ). (A) average transition path time in the folding direction ($U \rightarrow F$), showing the climbing (navy) and descent (cyan) segments of the path. (B) As in top panel but for transitions in the unfolding direction ($F \rightarrow U$). (C) likelihood per photon (multiplied by 10^4) obtained from the MLA of the photon trajectories simulated on a ϕ = 0.65 surface using a Landau free energy surface with fixed asymmetry.

However, there are very clear differences in the distribution of the transition path time between its two segments: the climb and the descent. The climb corresponds to the segment of the trajectory from the moment it enters the transition region (x_0) to the instant that it reaches the barrier top ($x = 0$) for the first time. The descent corresponds to the time it takes to reach x_1 for the first time starting from $x = 0$, but the molecule can return to $x = 0$ multiple times as long as it does not backtrack all the way to x_0. These segments are equivalent to $t_{TP}(x = 0|x_0)$ and $t_{FP}(x_1|x = 0)$ according to Makarov's definition[155], respectively. The two segments are identical for folding and unfolding transitions paths but in reverse for the latter: swapping x_0 and x_1. For the symmetric case it is interesting to note that the descent takes about 3-times longer even though is downhill. This is again because the productive climbs are those that minimize the path (or else they return to the originating basin and are not considered) whereas the descent results in much more lingering. However, the important result is that the separation between climb and descent segments is very sensitive

to the free energy surface asymmetry and produces distinctive profiles for both forward (folding) and reverse (unfolding) transitions. The implication is that the ratio between the times for the climb and descent segments of transition paths can provide key information about the shape of the free energy surface for non-trivial cases in which the barrier curvature is not symmetric (and hence the rate is not adequately described by equation 1). This finding has potentially great significance for the single-molecule characterization of protein (un)folding reactions via the maximum likelihood analysis of photon trajectories. The reason is that obtaining estimates of the average times for each segment is much easier than resolving the full path. We further explore this issue in the following sections.

4.4.3 Comparing different methods of transition path time calculations.

The maximum likelihood analysis of photon trajectories is the one approach available to estimate the mean transition path time from single-molecule FRET experiments[51]. One of its limitations is the need *a priori* of a kinetic model to compute and maximize the likelihood[123,124]. In its simplest version, the limitation is circumvented by comparing the maximized likelihoods obtained with two models: a two-state kinetic model and a simplified three-state model with a virtual intermediate. The two-state model assumes that the transitions are instantaneous and hence unresolved, whereas the lifetime of the virtual intermediate in the three-state model represents a finite (resolved) transition path. The comparison between the likelihoods ($\Delta \log L$) as a function of the intermediate lifetime provides an estimate of whether the transition paths are resolved, and the average transition path time if they are[47,51]. More technical details for the likelihood analysis with both models are given in the methods section (section 2.3)..

We analyse the photon trajectories generated from the diffusive stochastic kinetics on simple and well-defined 1D free energy surfaces, that is, assuming that the FRET efficiency operates as a good reaction coordinate. It provides a direct point of reference for the analysis of experiments. For this purpose, we generated stochastic diffusive kinetic trajectories on 1D free energy surfaces with asymmetry of $\phi = 0.65$ and with barriers ranging from 0.4 to 8.1 k_BT (Figure 4-1A). We then simulated photon emission trajectories associated to the intrinsic dynamics on the surface using a simple linear function $E(x)$ defined based on E_U and E_F values (the FRET efficiency at the two surface minima), which we chose to be consistent with typical SM-FRET experiments on En-HD at the chemical denaturation midpoint. We then performed MLA with the two-state and three-state models as described in methods. The results of these analyses are summarized in Figure 4-8.

The first point to note is that whereas $\log L$ is proportional to the total number of photons included in the analysis, $\Delta \log L$ is only sensitive to the much smaller fraction of photons emitted during the actual transition paths. In the original implementation this issue is tackled by analyzing only the small segments of photon trajectories that contain a transition (identified independently using a heuristic criterion, such as the detection of a sharp FRET

change in the binned photon data). This also requires that the ε_U and ε_F values are determined independently and then fixed during the procedure[47,51] because the short segments do not have enough information to set the minima. We tried this procedure using different ways to estimate ε_U and ε_F and performed a direct analysis of all the photon data without pre-selection and fitting globally ε_U, ε_F, k_u, k_f using the two-state model (section 2.3). This procedure is much simpler to implement as it eliminates the need for heuristic decisions to identify the segments corresponding to the transitions, which may not be so obvious when the transition path times are just slightly longer than the inter-photon times or there is some degree of acceptor blinking/bleaching[112]. The only issue against performing a global analysis is the added computational cost. However, the analysis is trivially parallelizable by distributing the photon trajectories among multiple threads/cores, so we strongly recommend performing this analysis using all the available photons.

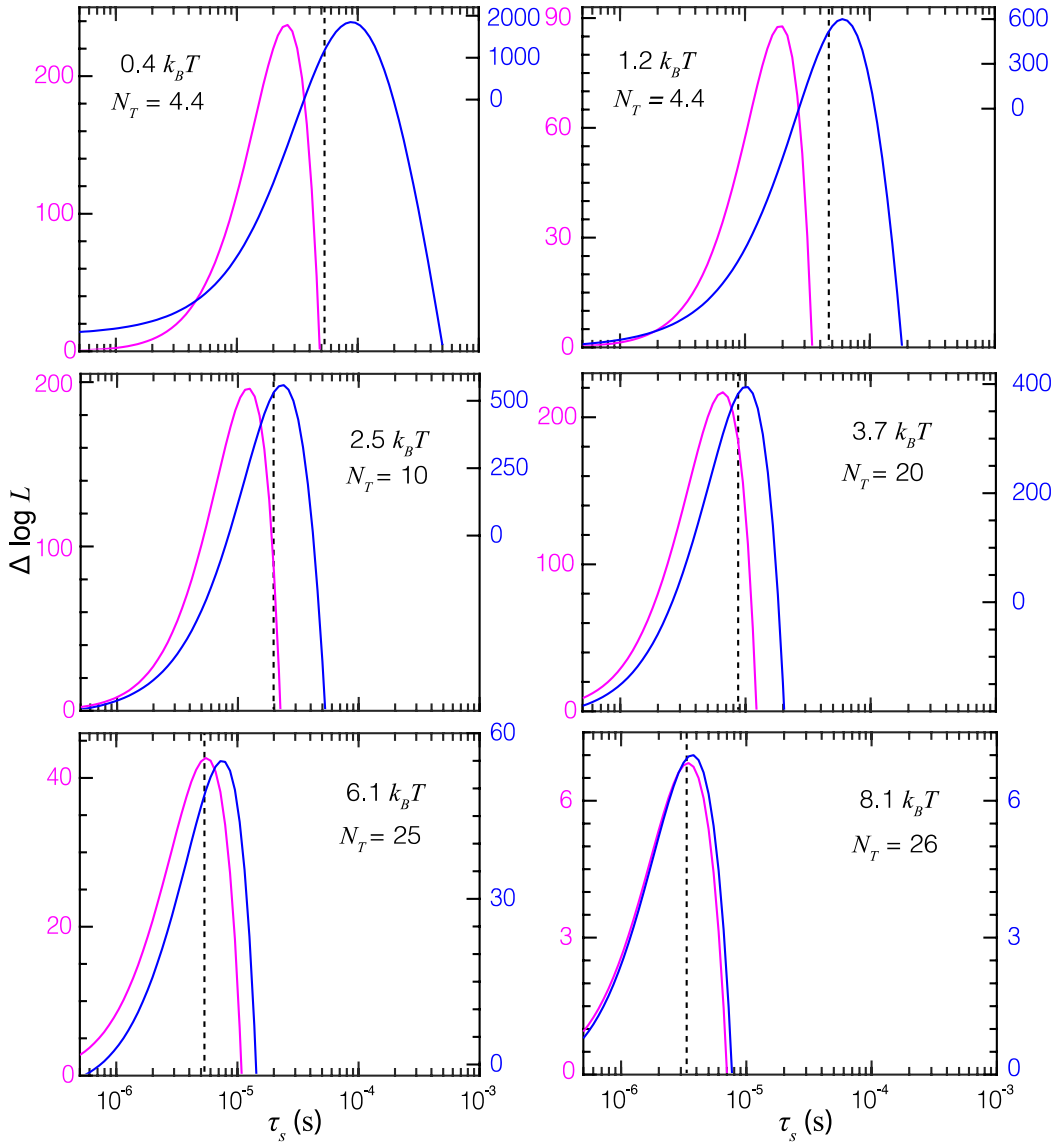


Figure 4-8 : Δlog L estimates of the experimental (not normalized) average TPT obtained from photon trajectory simulations on free energy surfaces with varying barrier. Results for simulations with increasing barrier are shown from top left to bottom right with the barrier height indicated in k_BT: Δlog $L_{2si\text{-}2s}$ (magenta, left scale) and Δlog $L_{3s\text{-}2s}$ (blue, right scale) methods. The vertical dashed line signals the average TPT directly measured from the stochastic kinetic simulations using 1/3 boundaries for the transition region. N_T is the total number of million photons used in the analysis

$\Delta G^{\ddagger}(k_BT)$	$\Delta\log L_{\text{FES-2s}}$	$\Delta\log L_{\text{3s-2s}}$	$\Delta\log L_{\text{2si-2s}}$
0.4	$4.76 \cdot 10^{-4}$	$4.18 \cdot 10^{-4}$	$5.4 \cdot 10^{-5}$
1.2	$1.76 \cdot 10^{-4}$	$1.36 \cdot 10^{-4}$	$1.99 \cdot 10^{-5}$
2.5	$7.38 \cdot 10^{-5}$	$5.6 \cdot 10^{-5}$	$1.96 \cdot 10^{-5}$
3.7	$3.02 \cdot 10^{-5}$	$2.0 \cdot 10^{-5}$	$1.08 \cdot 10^{-5}$
6.1	$5.21 \cdot 10^{-6}$	$2.2 \cdot 10^{-6}$	$1.7 \cdot 10^{-6}$
8.2	$7.72 \cdot 10^{-7}$	$2.7 \cdot 10^{-7}$	$3.0 \cdot 10^{-9}$

Table 2 : Likelihood values compared between different methods expressed per photon

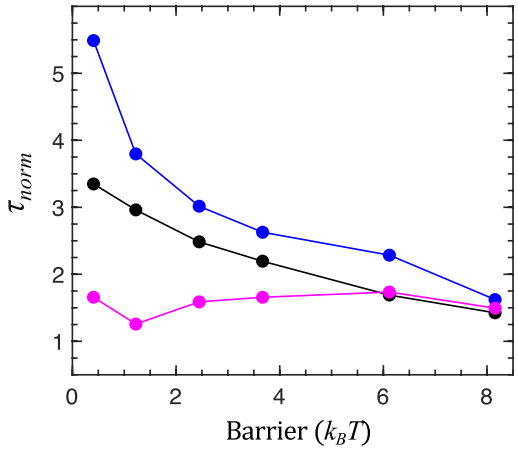


Figure 4-9 : TPT as a function of the free energy surface barrier (normalized as before). TPT directly measured is shown in black (dashed line Figure 4-8). TPT from $\Delta\log L_{\text{2si-2s}}$ and $\Delta\log L_{\text{3s-2s}}$ are shown in magenta and blue, respectively (τ_s corresponding to the peaks in Figure 4-8)

As for the performance of the original $\Delta\log L_{\text{2si-2s}}$ analysis, we find that it reliably produces a maximum for all the free energy surfaces, at least up to a barrier of ~8 k_BT (magenta). Therefore, this simple procedure does indeed detect transition paths only ~5 times longer than the average inter-photon time (compare Figure 4-8 with count-rates in Table 1). For the highest barrier, the maximum in $\Delta\log L_{\text{2si-2s}}$ is nearly identical to the actual TPT

obtained from the kinetic simulations and 1/3 boundaries. As the barrier decreases, however, the estimated and actual TPT diverge, with the $\Delta\log L_{2si\text{-}2s}$ maximum underestimating the TPT by up to 2.5-fold for the lowest barriers (Figure 4-9). This result is consistent with the findings of Taumoefolau and Best on surfaces projected from coarse-grained simulations, which tended to have barriers on the lower end of the range we explore here[160]. Notably, the absolute $\Delta\log L_{2si\text{-}2s}$ value at the maximum follows the opposite trend, with the marginal 0.4 k_BT barrier producing a maximum at ~240 (i.e. statistical significance of $1\text{-}10^{-240}$), whereas it is only ~7 (significance of $1\text{-}10^{-7}$) for the 8.1 k_BT barrier (magenta scale in Figure 4-8). The difference in likelihood as a function of the barrier is even more drastic when is corrected by the number of photons used in the analysis (Table 2). However, we note that this difference is simply a reflection of the number of "significant" photons that are contained in each dataset, that is, the photons that are emitted during the transition paths. A simple way to correct for the number of "significant" photons is to multiply the $\Delta\log L_{2si\text{-}2s}$ per photon (from Table 2) by $\exp(\Delta G^{\ddagger})$. Such calculation shows that the $\Delta\log L_{2si\text{-}2s}$ per "significant" photon is almost 5-fold higher for the tallest than the shallowest barrier, explaining why this method performs better the more two-state-like is the free energy surface, whereas it increasingly underestimates the TPT when the barrier drops below 4 k_BT (Figure 4-9).

We also tried an alternative procedure in which the analysis with the three-state model is performed with fixed k_u and k_f (values from the two-state analysis), but ε_U and ε_F are adjusted to allow accommodation for the virtual intermediate ($\Delta\log L_{3s\text{-}2s}$, blue curves in Figure 4-8 and Figure 4-9). This analysis generally produced much higher $\Delta\log L$ maxima (blue in Figure 4-8), and longer TPT for lower barriers. The latter is expected since lower barriers are more populated and hence more consistent with the presence of an intermediate. On the other hand, the $\Delta\log L$ as a function of the intermediate lifetime is a broader curve, and its maximum overestimates the TPT, particularly for the intermediate to small barriers, although it is comparatively closer to the real value, overall. Nevertheless, the most important result that emerges from this comparison is that, together, the two procedures set a useful range and their mean provides a good estimate of the actual TPT, even when the barrier is very small.

4.5 MLA with a one-dimensional FES model

The combination of the two procedures outlined in the previous section can lead to fairly accurate estimates of the TPT. However, such analysis does not provide any information about the underlying free energy surface, nor even accurate estimates of the barrier height since barrier curvature and height are directly linked (see Fig. 4). In principle, if the rate of a process can be effectively described as diffusion on a free energy surface (e.g. equation 1), the photon trajectories from single-molecules should contain all of the information about the free energy surface and the diffusion coefficient. The key question is whether, or to what extent, such information can be accurately extracted via ML analysis. The methodology to implement such an approach has been discussed in detail in chapter 2

(section 2.4) and has been demonstrated for En-HD (chapter 3, section 3.7).We investigated the performance of the MLA of simulated photon trajectories implemented with the variable-shape free energy surface model that we used to generate the original 1D free energy surfaces.

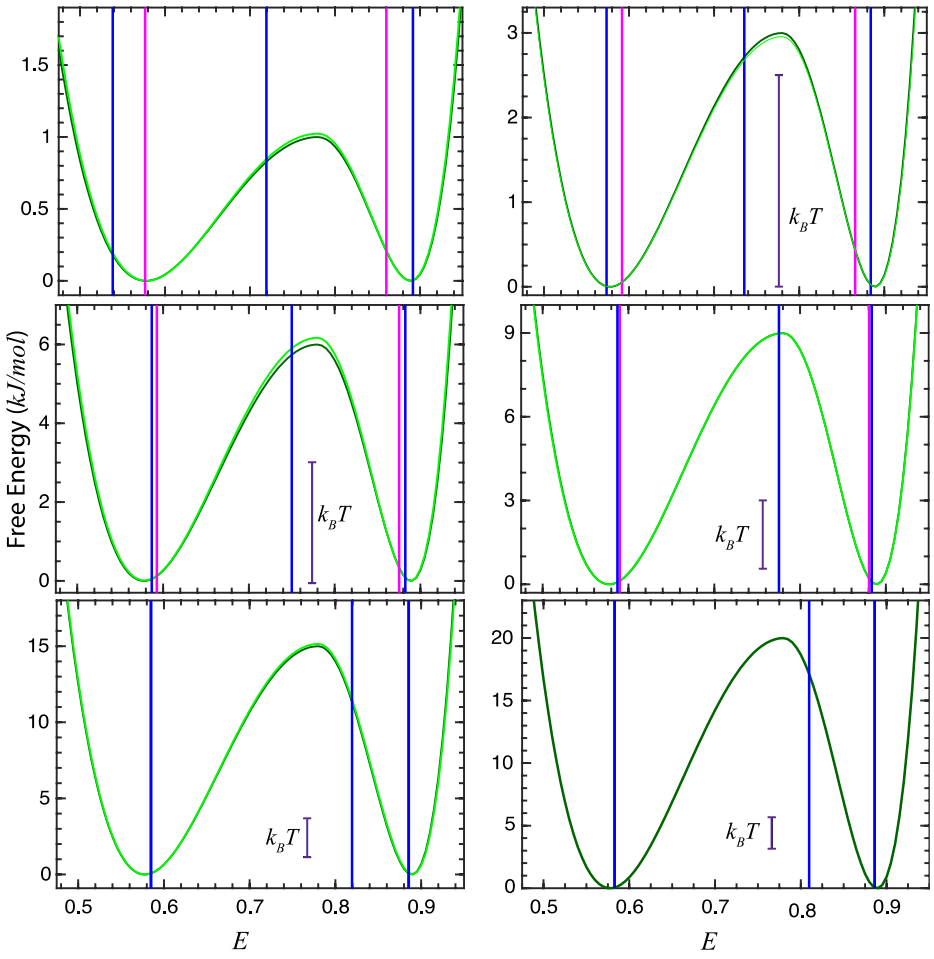


Figure 4-10 : Free energy surfaces corresponding to different folding scenarios showing E for folded state (E_F) and unfolded state (E_U) obtained from two-state (magenta lines) and E_F, E_U and E_S from three-state kinetic models (blue vertical lines). FES in light and dark green shows the original FES used for the simulation and the FES retrieved from MLA analysis of the simulated photon trajectories respectively

The first observation from the FES-MLA is that the same photon trajectory data that we used for the analysis with $\Delta\log L_{\text{2s-2si}}$ and $\Delta\log L_{\text{2s-3s}}$ methods contains sufficient information to accurately retrieve the original free energy surface, regardless of how high is the barrier that separates the two minima (Figure 4-10). $\Delta\log L$ of the FES-MLA relative to the two-state fit ($\Delta\log L_{\text{FES-2s}}$) is much higher than $\Delta\log L_{\text{2si-2s}}$, or even $\Delta\log L_{\text{3s-2s}}$, for all

the barriers (Table 2). This result confirms that there is enough information in the photon trajectories to accurately define the barrier height and dynamics. Moreover, the FES-MLA is also perfectly capable to resolve the differences between the climb and descent segments of transition paths from the photon trajectories, and hence, it can determine the degree of asymmetry (shape) of the free energy surface.

Importantly, the comparison of the results from FES-MLA and the two-state/three-state MLA provides an empirical reference to ascertain whether a given photon trajectory dataset can define the shape of the free energy surface relative to the presence of a discrete intermediate (the Landau MLA has the same number of fitting parameters than the restricted three-state model with single k_S). This is particularly useful to discriminate between dynamics on a free energy surface with a shallow barrier and the relaxation of an intermediate state separated from U and F by barriers. For instance, the $\Delta\log L$ results summarized in table 1 indicate that, with our simulated protein dynamics, count-rates, and FRET efficiency values, about 120,000 photons are already sufficient to distinguish (with statistical significance of $1\text{-}10^{-12}$) the dynamics on a 0.4 k_BT surface from a relaxation through a discrete intermediate. Table 1 shows that such a shallow barrier is the most challenging scenario to discriminate, whereas resolving the diffusive transition paths becomes comparatively easier the higher is the barrier (provided the count-rate is at least 4-5-fold faster than the TPT).

The comparative analysis reveals another by-product of fitting the photon trajectories to discrete kinetic models. Namely, we find that the two-state model does not place ε_F and ε_U where the surface minima are, but closer together (magenta lines in Figure 4-10). The discrepancy is minimal for the scenarios with the highest barriers, but it grows proportionally to the barrier shallowness. This effect seems to reflect dynamic averaging with the growing population of the microstates at the barrier. To test this idea, we performed the two-state fitting with the actual ε_F and ε_U fixed to the known E_F and E_U values, which returned many orders of magnitude lower likelihoods, indicating that the effect does indeed reflect dynamic averaging from excursions along the barrier slope (the population at the top of a 0.4 k_BT barrier is 2/3 of that at the minimum). Because the shift in ε_F and ε_U reflects the true dynamics of the system, it is preferable to fit these parameters jointly with k_u and k_f, rather than trying to fix them to expected values, as we suggested above. In fact, the observation of such shift in the two-state analysis could be used as a first indication that the TPT might be resolvable. The three-state model partially corrects the dynamic shift in ε_F and ε_U by introducing an intermediate. However, the three-state model experiences a similar effect in the placement of ε_S relative to the barrier top, which is shifted towards U for the lower barriers, and towards F for the highest ones (blue lines in Figure 4-10). We note that these issues arise from the use of discrete kinetic models. Hence, we conclude that an MLA procedure that directly uses a model of diffusion on a simple 1D free surface, which interprets such dynamic effects to extract the shape of transition paths and underlying free energy surface, is much preferable, as it has been proposed before by our group [17], and more recently by others[160]. This is particularly true if the shape of the

idealized 1D surface (barrier height and asymmetry) can be controlled with specific parameters, as the Landau model we introduce here.

4.6 Conclusion

There has been a growing interest during the last years in the analysis of the reactive transition paths of protein (un)folding and other biomolecular reactions driven by large conformational changes[140,141,161]. This is so, because such transition paths contain essential information to unravel the complex mechanisms that determine the rates of these processes[65]. A major motivating factor has been the recent developments in single-molecule fluorescence and force techniques, which have opened the opportunity to address these questions experimentally for the very first time. Here we tackled the two related questions of i) how to extract transition path information from single-molecule experiments, and ii) how to interpret such information in mechanistic terms. We start by assuming that the rates and transition paths of protein (un)folding and related processes can be described as diffusion on a simple free energy surface that results from the projection of the hyper-dimensional energy landscape onto a single reaction coordinate. We describe such free energy surface analytically using a Landau quartic polynomial, which produces an idealized 1D surface with two minima separated by a barrier of controllable height and symmetry.

Using stochastic simulations of diffusion on such surfaces, we closely looked into the direct relationships between the shape of the free surface and the transition paths (shape and time). From this analysis we extract three main conclusions:

a) The choice of boundaries defining the transition region in the surface has strong impact on the resulting transition path times (TPT) and commitment probabilities. However, we conclude that setting the boundaries at 1/3 of the distance between the minimum and the barrier top is the definition of transition paths that best captures the dynamics on the surface, particularly for barriers in the range that is most relevant to protein folding reactions (0-10 k_BT).

b) For free energy surfaces with the same shape (symmetry), the TPT is inversely proportional to the barrier height, being longest when the barrier is marginal. We show that this behaviour, which arises because barrier curvature and height are tightly linked, is in fact consistent with the known expression $\langle \tau_{TP} \rangle \approx \ln [2e^{\gamma} \beta \Delta G^{\ddagger}] / D\beta \left(\omega^{\ddagger}\right)^2$. Practically, this result implies that it should be easiest to resolve the transition paths for the fastest folding proteins because their transition paths will be both longer and much more frequent.

c) Whereas the average TPT is essentially insensitive to the surface asymmetry, as it has also been indicated by others[155,159], we find that the climb and descent segments of the transition path are exquisitely sensitive to the barrier asymmetry. Therefore, resolving the climb and descent fractions of the TPT could provide enough information to estimate the level of asymmetry of the surface (e.g. how close is the barrier to the folded minimum).

We then investigated the performance of various maximum likelihood approaches for extracting transition path information from single-molecule photon trajectories. Our results show that the approach for determining the mean TPT from MLA of two-state and a restricted three-state kinetic models underestimates the TPT, particularly for barriers smaller than 4 k_BT, consistently with the recent results from a similar analysis based on coarse-grained molecular simulations[160]. The major factor behind such underestimation is that the two-state MLA shifts the two minima closer together in response to the growing population at the barrier, which makes the transition path effectively shorter. In response to that observation, we introduce here a variant of the approach in which the three-state MLA is allowed to accommodate the position of the minima. We find that this three-state variant overcompensates for the effect, resulting in an overestimation of the TPT. However, we propose that the two MLA variants combined provide a rather accurate mean TPT over the entire range of barriers explored here.

However, we do find that the MLA implemented with a free energy diffusive representation of the kinetics (MLA-FES) offers a much superior performance. Such analysis produces much higher likelihoods than the discrete kinetic models with equal number of parameters, indicating that there is sufficient information in the simulated photon trajectories to extract the diffusive dynamics on the free energy surface. We then confirm that this analysis obtains the correct barrier throughout the 0-10 k_BT range, as well as the asymmetry of the free energy surface. For a real experiment, in which the free energy surface is not known *a priori*, the researcher can compare the likelihoods of FES and three-state models as direct indicator of the statistical significance of the FES analysis (e.g. Table 2). Accordingly, we strongly recommend performing the MLA of experimental SM-FRET photon trajectories with free energy surface models discussed in chapter 2 (section 2.4). Furthermore, such analysis can easily accommodate more complex scenarios: a diffusion coefficient that changes along the reaction coordinate, as it has been postulated by theory [7] and observed in simulations [37,133,162]; or the use of higher dimensional surfaces to reproduce the results from experiments that monitor several distances in the protein (multiparametric and/or three-color FRET experiments).

Chapter 5

5 Investigating the role of topology and secondary structural elements in folding dynamics

5.1 Introduction

Protein folding is the process by which protein molecules adopt their biologically functional native structure with distinctive secondary and tertiary structures and specific overall topology. Roles of various determinants of folding like secondary structural elements and topology in the overall folding mechanism is one of the fundamental questions that has been evaluated since the question about how proteins fold was raised [163,164]. Deciphering the molecular processes that govern how proteins fold into the functional form constitutes a problem of enormous relevance.

Experimental and theoretical results from simulations have shown that many proteins, particularly small fast-folding proteins, fold over a minimally rough energy funnel as their sequences have been evolved to have reduced the level of energetic frustration. Hence, it is believed that the topology of the native fold plays an important role in determining the formation of transition state ensemble and thus the mechanism of folding. However, notable exceptions to this have been found with multiple proteins with the same final fold but significantly different mechanisms, suggesting that the topology alone cannot be used as a determinant for folding mechanism[69,165]. There are studies suggesting that a protein folding rates depend on properties that are evident from its native structure. This leads to a protein folding model in which the size and stability of the polypeptide segments organized in the transition state are largely independent of the protein length but are related to the topological complexity of the native state[166,167]. It has also been shown that equilibrium protein folding cooperativity can be explained as a combination of weak propensities of peptide chains to form secondary structures and stronger propensities of tertiary interactions to stabilize the secondary structures. The rate of formation of various secondary structural elements like α-helix and β-strand is significantly different. It is also observed that helical proteins tend to fold faster than β-sheet proteins in general. The composition of secondary structural elements has been utilized to predict the folding rate of large proteins[168].

This study aims to study how the interplay between the topology and the secondary structure defines the overall folding mechanism. Four proteins: En-HD, FBP11-WW1, Nedd4-WW4, and gpW were chosen for this study. En-HD consists of three α-helices, the FBP11-WW1 and Nedd4-WW4 are WW domain proteins and have 3-stranded β-sheet, and

gpW has two α-helices and one β-hairpin. Each domain represents one of three major protein secondary structural classes: all-helix, all-beta, and α+β, providing a relatively good sampling of the secondary structural patterns observed in nature. All four proteins have their secondary structural elements arranged in an antiparallel fashion and share similar overall topology. Folding of these single domain proteins involve spatial rearrangement of three antiparallel structural elements: helix-helix-helix (En-HD), strand-strand-strand (FPB11-WW1 and Nedd4-WW4), and helix-hairpin-helix (gpW). Ultrafast kinetic measurements using laser Temperature jump from our group have identified up to an 8-fold difference in folding rate constants of 2 proteins, Ned4-WW4 and FBP11-WW1, which have similar size, secondary structure, and tertiary structure[169,170]. This is remarkable as it suggests a larger change in diffusion coefficient among two proteins of the same size and structure. Most of the fast folders known are proteins with helical secondary structures. WW domain proteins belong to the family of fastest folding β-sheet proteins and hence are interesting candidates for resolving dynamics. Prior experimental results on these proteins indicate that these are fast-folding proteins and fold in a downhill fashion or by crossing a minimal free energy barrier. This suggests comparable energetics, and hence these proteins are ideal for studying the effect of secondary structure on protein dynamics. Bulk experiments done on these proteins have shown that they (un)fold in sub-ms timescales at

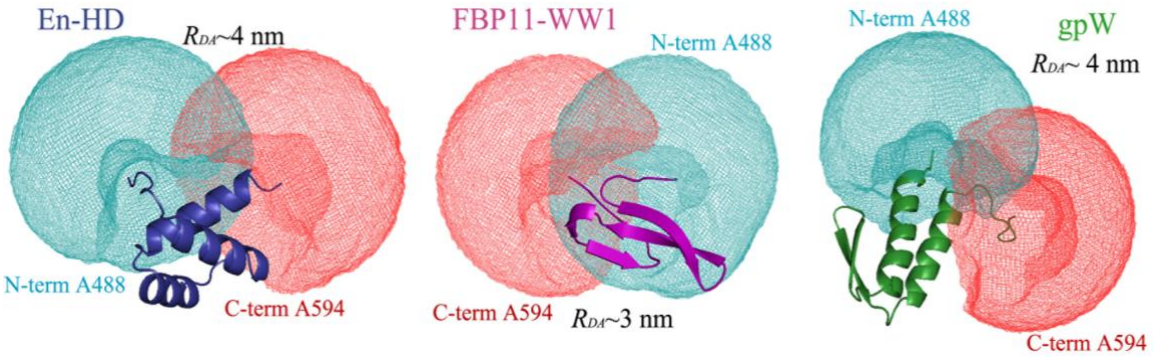


Figure 5-1 : Cartoon representation of the fluorophore labelled in the folded structure of three fast-folding domains. The mesh representation shows the 3D anisotropic orientations of the fluorophores indicating the site-specific labelling. The proteins are labelled at their flexible C-terminal and N-terminal ends with Alexa Fluor-488 and Alexa Fluor-594 as FRET pair, as shown in the figure. Figure also indicates the calculated average end to end distance for the FRET pair (R_0 ~ 5 nm). FBP11-WW1 and Nedd4-WW4 shares the same structure.

room temperature and chemical denaturation midpoint[129,130,143,170]. These rates are optimal for being studied by SM-FRET experiments under free diffusion conditions.

A simple primary analysis of SM data based on binning will be utilized to produce FEH at different binning times. This will be used to obtain a rough idea about the populations present under experimental conditions. We then will make use of the MLA method in

combination with the 1D-FES model to obtain the free energy surfaces for each protein. Free energy barrier, rate of folding and unfolding and intramolecular diffusion coefficient will be compared for all four proteins to gain insights into the nature of folding. The ability to directly obtain the free energy surface and the dynamic factors hold the key to dissect the role of fold topology and formation of secondary structure in the overall folding mechanism. Obtaining similar intramolecular diffusion coefficients would be suggestive of folding dynamics determined predominantly by the spatial arrangement of three-preformed secondary structural elements, thus making the folding mechanism to be driven primarily by topology. Observation of significantly different intramolecular diffusion coefficient, on the other hand is suggestive of the rate of formation secondary structural elements dictating the folding mechanism as it is already known the rate of formation of different structural elements are different.

All four proteins were produced with cysteine residues at the N and C-terminal[1]. Proteins were then fluorescently labelled with FRET acceptor-donor fluorophore pairs (Alexa-488 and Alexa-594) by making use of cysteine-maleimide chemistry as described in chapter 2 (detailed procedures are described in appendix). SM measurements were carried out on the home-built SM-FRET set-up as per the procedures described in earlier chapters. Protein concentration for all the experiments was kept at 20 pM to ensure SM conditions. Laser power was kept at 150 μW, and a photoprotection cocktail with a final concentration of 1mM (*S*)-Trolox methyl ether and 10 mM cysteamine was used in the working buffer. Working buffer for both WW domain proteins was 20 mM MOPS buffer at pH 8.0 and 20-mM phosphate buffer at pH 7.4 was used for gpW. The data used to demonstrate the methodology in chapter 2 is used for discussions for En-HD.

The methodology of utilizing MLA with the 1D-FES model established in chapter 2 was used to analyses all data discussed in this chapter. Data analysis followed the same procedure, starting with the identification of photon bursts, screening the bursts to remove the bursts with photophysical artifacts, followed by MLA analysis with a 1D-FES model to extract the free energy surface and the intramolecular diffusion coefficient. The MLA method along with a simple two-state kinetic model was also done for extracting folding-unfolding rates for comparative studies.

5.2 Results and Discussion

All proteins considered in our study are relatively small, single domain proteins. Distance estimated between the FRET acceptor-donor pair estimated from the folded structure for WW is ~3 nm and is 4 nm for gpW and En-HD. The labelling sites used in the constructs used in this study were selected to have similar donor-acceptor average distance. Among

[1] SM-FRET data on protein gpW was collected by Dr. Abhigyan Sengupta. Data analysis was done in collaboration with Dr. Sengupta and the manuscript publishing the results on gpW is under preparation.

these proteins, En-HD and gpW have similar sizes (61 amino acid residues) and are larger than the two WW domain proteins (43 amino acid residues). Resolving folded and unfolded conformations of proteins through the change in FRET efficiency gets challenging as the size of the protein decreases (while the end-to-end distance in folded state is small, the rotational fluctuation of C and N-terminal results in higher *E* for unfolded state, resulting in a smaller range of FRET change).

Midpoint denaturation condition was used for the collection of SM data to maximize the number of folding-unfolding transitions. En-HD and gpW show bimodal FEH, determination of denaturation midpoint for these proteins were done by collecting SM data at different concentration of denaturant and selecting the concentration with roughly equal folding-unfolding rates from a two-state model (C_m was also confirmed from bulk fluorescence unfolding curves). Both the WW domain proteins show unimodal FEH at different concentrations of denaturant. A gradual shift of the mean FRET of the population to lower FRET values with increasing denaturant concentration was found. The two-state model performs poorly when the *E* of the states are not clearly defined (unimodal FEH), and hence near midpoint denaturant concentration was approximated to be when the mean *E* of the unimodal distribution was halfway between the *E* for the most unfolded and most folded conditions and symmetric distribution of the population was observed in FEH.

FRET efficiency distribution of bursts identified for FBP11-WW1 and Nedd4-WW4 shows unimodal population distribution for different concentrations of GdnCl. Mean *E* of the population shifts to the left (lower *E*) with increasing concentration of GdnCl, indicating a shift of population to the unfolded state. We see a steady decrease in *E* with increasing denaturant concentration. This is a commonly observed phenomenon for multiple proteins by various other groups and is a result of the expansion of the unfolded state in the presence of high concentrations of denaturant [171–173]. It is important to note that the total change in the mean *E* for the lowest and highest denaturant concentration for both WW domains appears to be close to 20%, but this is because of the expanded unfolded state. The contrast between the signal for the unfolded state and folded state, i.e., change in *E*, to resolve dynamics along the folding reaction coordinate is low. This ends up being a limitation for resolving the conformational dynamics of faster folding WW domains.

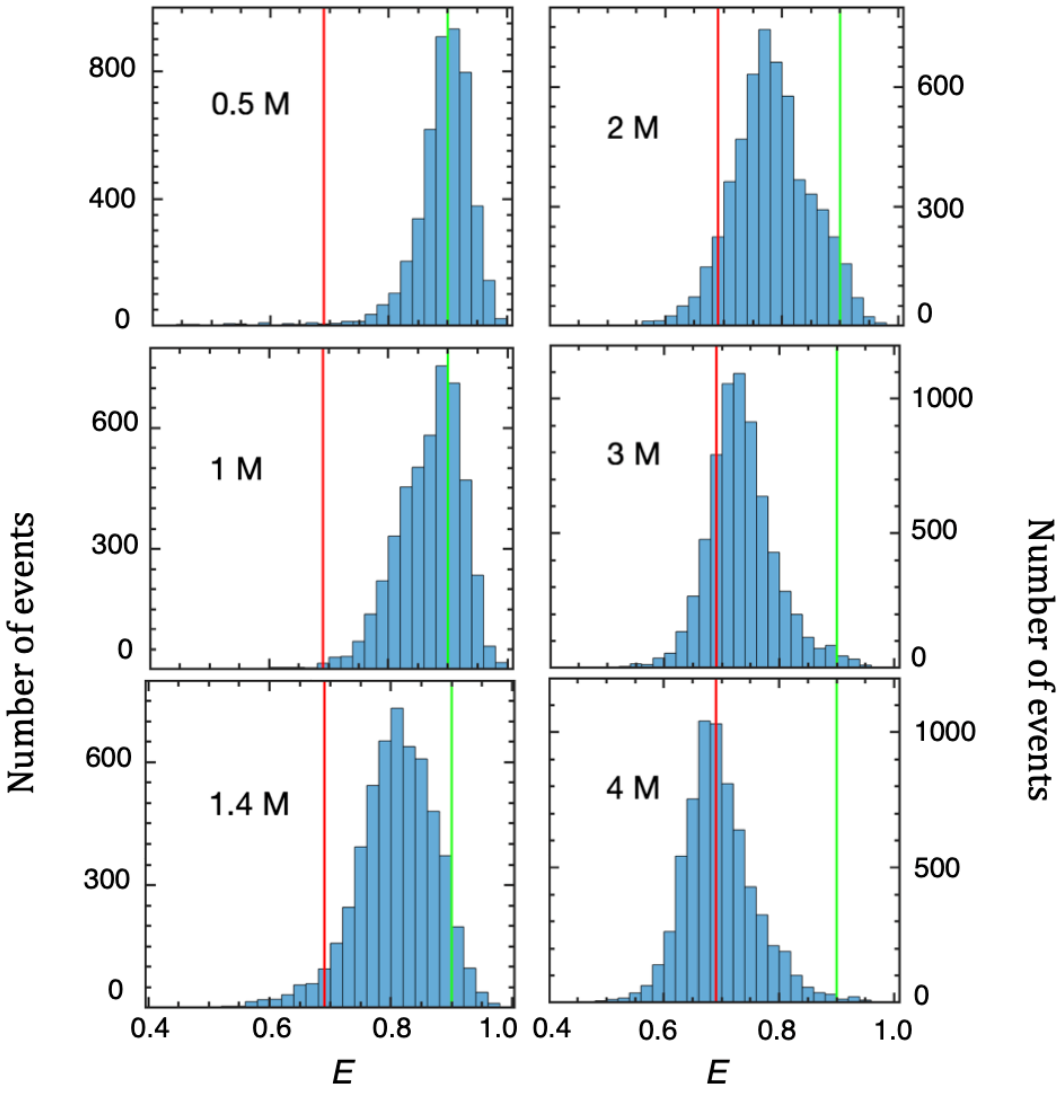


Figure 5-2 : Chemical denaturation at SM conditions for Nedd4-WW4. Histograms show the distribution of *E* for the bursts identified using algorithm described earlier for different concentration of GdnCl (mean of the duration of burst is ~0.6 ms). Red and green vertical lines (0.69 and 0.90) indicate the shift of mean *E* for the lowest and highest concentration used. FEH for 1.4 M GdnCl (~C_m) is further analysed with different binning times.

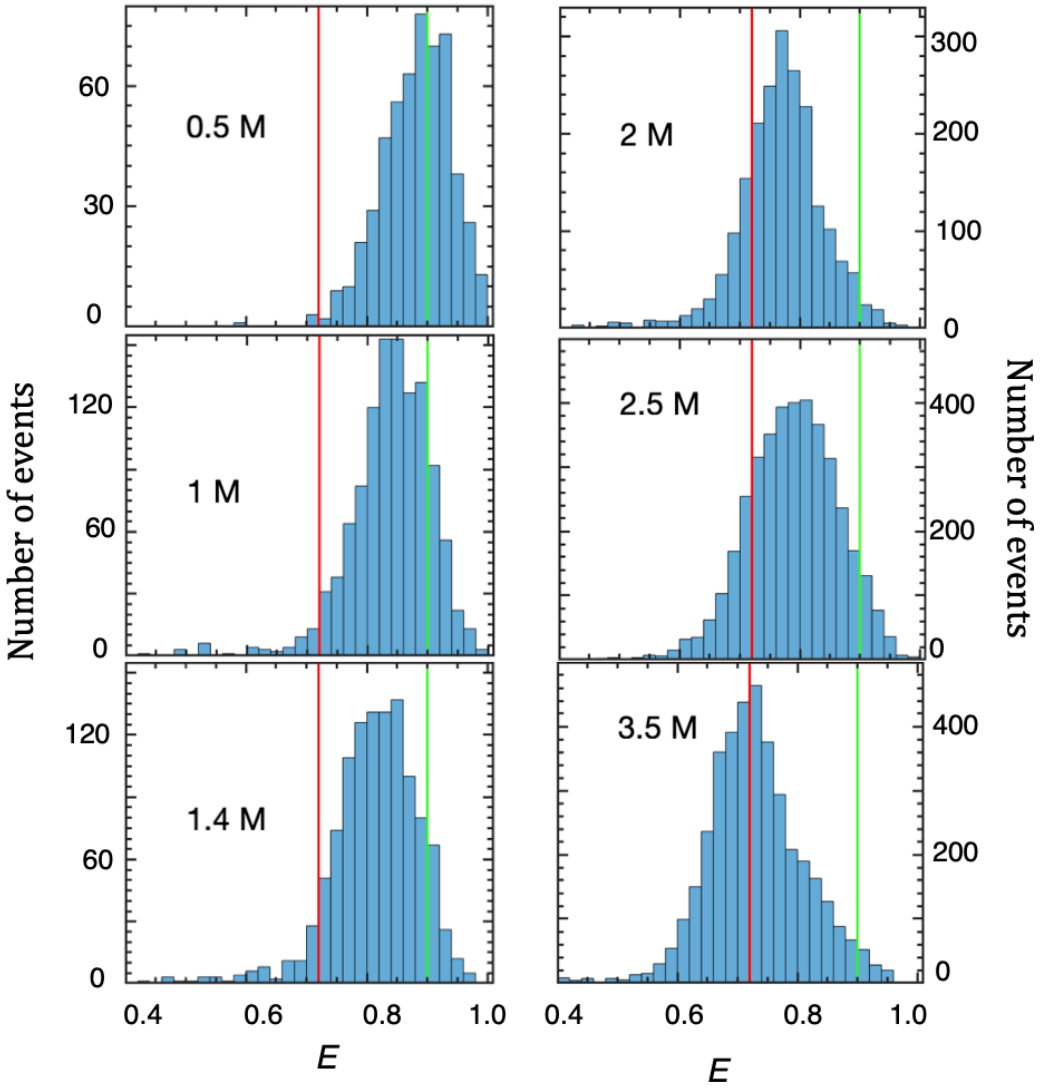


Figure 5-3 : Chemical denaturation at SM conditions for FBP11-WW1. Histograms shows the distribution of *E* for the bursts identified using algorithm described earlier for different concentration of GdnCl (mean of the duration of burst is ~0.6 ms). Red and green vertical lines (0.72 and 0.90) indicate the shift of mean *E* for the lowest and highest concentration used. FEH for 2.4 M GdnCl (~C_m) is further analysed with different binning times.

The most common way of representing protein folding kinetic data as a function of the concentration of denaturant is by generating a chevron plot (Figure 1-7). Chevron plot is usually generated from kinetic experiments where relaxation rates following a rapid perturbation of the system are monitored[83,174]. Examples are stopped-flow and continuous-flow experiments where folding rates are observed after rapidly diluting protein in a concentrated denaturant solution to a lower denaturant concentration. Unfolding rates are obtained from a similar approach starting with protein in a low denaturant concentration. T-jump and pressure-jump experiments are also used to improve the time resolution of the system. A commonality between all these methods is that relaxation following a rapid perturbation is followed, thus providing the folding rate of the unfolding rate at any given point. Single-molecule experiments provide a method to obtain the chevron plot from equilibrium conditions without introducing a perturbation to the system. It further extends the effective range of the chevron plot by providing unfolding rates in conditions where the folded state is preferred (low denaturant concentration) and folding rates at unfolding conditions (high denaturant concentration).

SM-FRET data for Nedd44-WW4 at different concentrations of GdnCl was analysed using

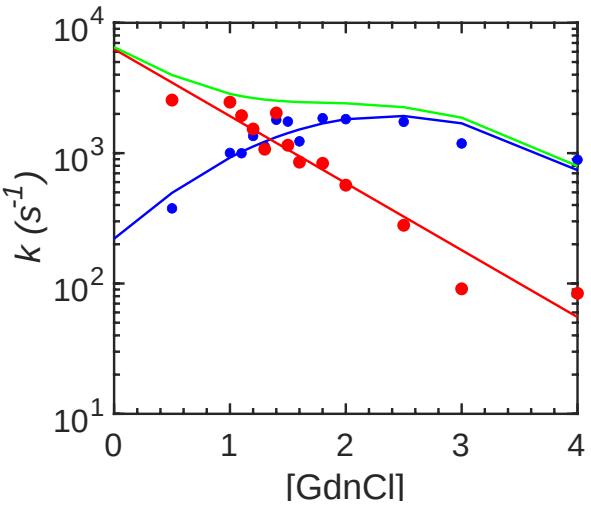


Figure 5-4: Chevron for Nedd4-WW4 obtained from 2-state fit to single-molecule data. Blue and red circles show unfolding and folding rates directly obtained from the fit. Green curve shows sum of the fits to the rates

MLA with a two-state model. Unfolding and folding rates thus obtained from the two-state model were then used to generate the chevron plot (Figure 5-4). We see a relatively flat chevron with the rates minimally dependent on the concentration of denaturant. Deviation from linearity of the folding and unfolding arms of the chevron indicates that the folding here is a minimally cooperative process and suggests a downhill folding scenario[95].

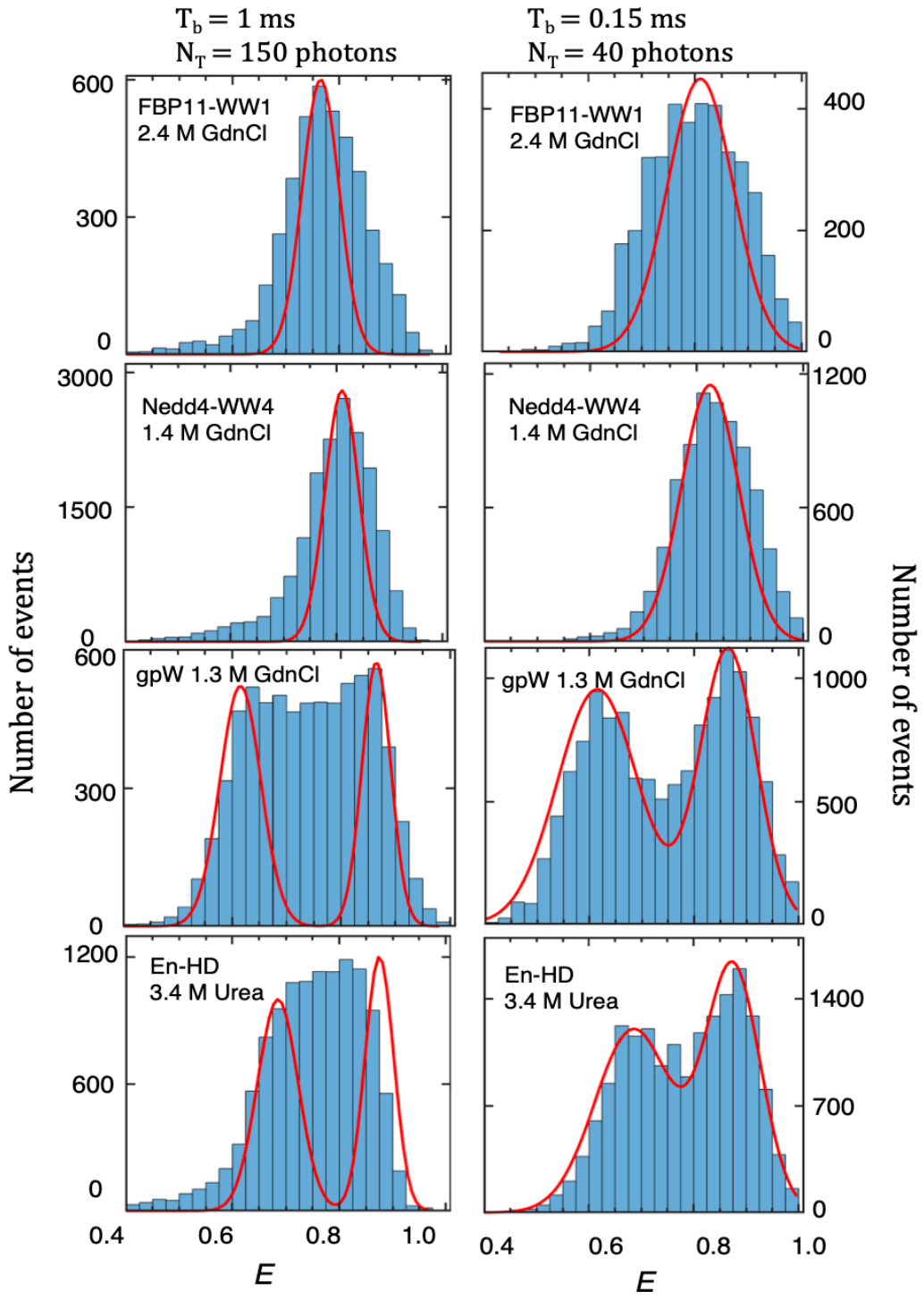


Figure 5-5 : FEH for all four proteins at different binning time (T_b) at ~C_m. Photon threshold of 150 photons and 40 photons were used for 1 ms (left column) and 0.15 ms (right column) respectively. Red curve shows the shot noise limited distribution around mean *E*. Broadening suggests dynamics happening within the timescale of binning.

To further compare the dynamics, simple time binning was done for near midpoint denaturant concentration for all four proteins. FEH was generated for 1 ms and 0.15 ms binning with 150 photons and 40 photons threshold, respectively (Figure 5-5). FEH was compared with the shot noise limited curves to check for dynamic averaging. FEH for En-HD and gpW produced a unimodal distribution significantly broader than shot noise limited FEH for 1 ms binning. A shorter bin time separated the unimodal distribution into mostly shot noise limited bimodal distribution of folded and unfolded ensemble. This suggests the timescale of folding dynamics to be comparable with the 0.15 ms binning time. Both WW domains showed FEH with broader than shot noise determined unimodal distribution for longer binning time (1 ms) as well as the shorter binning time (0.15 ms).

A unimodal FEH for a fast-folding protein can be interpreted in two ways, first scenario suggests the protein is undergoing dynamic exchange between the conformational ensembles, separated by an energy barrier, faster than the binning time used. The second scenario involves the downhill folding proteins where the protein undergoes folding without crossing a free energy barrier through a gradual ordering of unfolded ensemble to the folded ensemble, resulting in a unique population distribution at any given instance along the folding reaction. We performed MLA analysis for the bursts identified for all four proteins with a simple two-state kinetic model to obtain folding-unfolding relaxation rates. Relaxation rates obtained for gpW, En-HD, Ned4-WW4 and FBP1-WW1 were 3765 s^{-1}, 4780 s^{-1}, 2200 s^{-1} and 2397 s^{-1} respectively. This is an interesting observation as the proteins gpW and En-HD, which clearly show a bimodal distribution with a binning time of 0.15 ms, have faster relaxation rates. In contrast, both the WW domains which had unimodal FEH distributions under the same binning time show slower dynamics when analysed with a two-state model.

However, it is also important to note that relaxation rates from a two-state model are highly dependent on the ability of the model to identify the mean E of the two states accurately, folded and unfolded ensemble. As a result, the two-state analysis of unimodal distributions are trickier compared to bimodal distributions. Moreover, the two-state model implicitly assumes the protein to be populating two distinct states separated by a substantial barrier. This assumption falls apart for downhill folding proteins and the relaxation rates obtained from such a model fail to explain the dynamics happening. These limitations lead us to the need for a more inclusive model with minimal assumptions about the folding mechanism embedded in it.

A solution to this issue of measuring conformational dynamics for downhill folders is to measure the dynamics along a barrierless free energy surface. Obtaining the free energy surface from SM-FRET data would also confirm whether the proteins under discussion do indeed fold without crossing a free energy barrier. Such an approach can give information about not just the folding-unfolding relaxations but also about various other dynamics that could be taking place along the reaction coordinate. The practice of separating dynamics along a 1-dimensional FES to folding-unfolding relaxation rate is an approach developed

for proteins folding over a significant free energy barrier. Downhill folding proteins are expected to have comparable timescales for various exchanges along the FES.

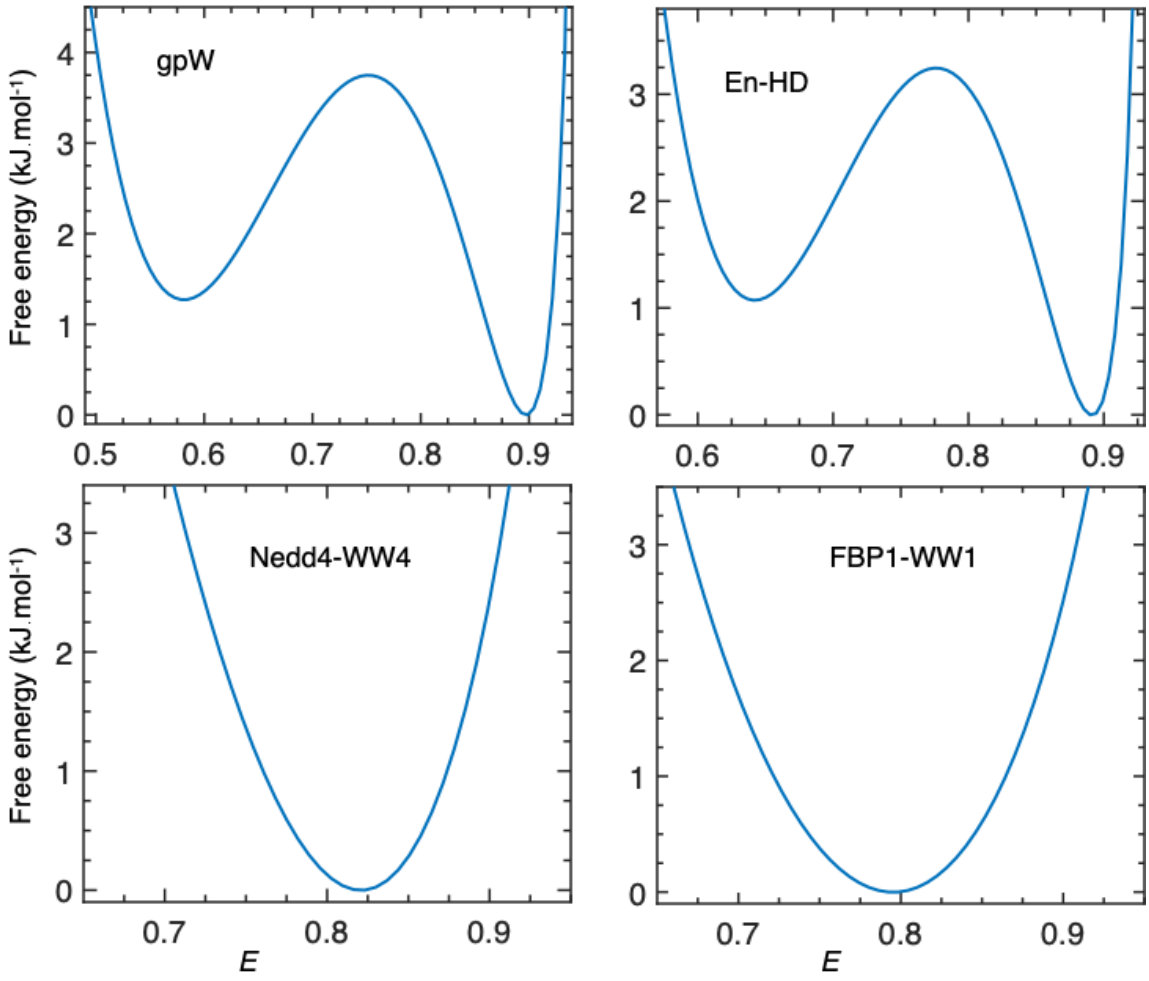


Figure 5-6: Free energy surface obtained from MLA analysis with the 1D-FES model at C_m for the four proteins under study

1D-FES model along with MLA, was able to retrieve free energy surface for all proteins. MLA method offers an opportunity to compare the performance of models for a given data set as the likelihood values can be compared for different models. The likelihood obtained for the 1D-FES model for each protein was significantly better than the value obtained with a two-state kinetic model. Table 3 shows the difference in likelihood in log units, the values are calculated per million photons as the total number of photons (amount of single-molecule data) collected for each protein were different. En-HD and gpW showed an FES with a minimal barrier (< 2 RT) while both WW domains showed strictly barrierless FES.

These FESs demonstrate the capacity of our 1D-FES model to extract FESs corresponding to diverse folding scenarios. This model, regardless of its simplicity (5 parameter model with FES defined by just 2 parameters) was able to retrieve the shapes and dynamics of four proteins under study.

	$L_{1DFES} - L_{2s}$ (per 10^6 photons)	log D	λ_1 (s^{-1})	λ_2 (s^{-1})	λ_3 (s^{-1})
gpW	134	2.57	3434	29939	61577
En-HD	82	2.63	4557	34834	71632
FBP11-WW1	356	1.2	1088	2532	4275
Nedd4-WW4	298	1.0	1256	2579	4026

Table 3 : 1D-FES analysis for 4 proteins. Likelihood values are expressed in log units

The rate matrix used for defining the diffusive folding process on the 1D-FES contains information about various dynamic processes taking place along the reaction coordinate. The slowest relaxation rate (λ_1) corresponds to the folding-unfolding relaxation rates while the second slowest rate (λ_2) corresponds to the depopulation of the barrier and the third slowest rate (λ_3) corresponds to the rate of interconversion among *U*/*F* ensemble (refer to section 2.4.3). We see En-HD to have the fastest relaxation rates and WW domains to be the slowest. gpW shows rates slower than En-HD, but still much faster than the all-beta WW domains.

FBP11-WW1 and Nedd4-WW4 fold without crossing any free energy barrier in a completely downhill fashion. An implication of such folding is that the timescale of various relaxation processes along the free energy surfaces being comparable. This is evident from the values of $\lambda_1$and λ_2 for the WW domains (Table 3). The rates are different only within a factor of two. In fact, the separation of timescales into folding time and transition path time, is not a useful approach for these proteins. A more meaningful approach to obtain the overall relaxation rates for barrierless scenarios is by inspecting the amplitudes of various terms of the rate matrix, **K (**section 2.4.3**)** and estimating an overall relaxation rate with contributions from all terms of the rate matrix. Figure 5-7A shows the kinetic amplitudes of the three slowest non-zero eigenvalues obtained from the rate matrix. Figure 5-7B shows the relaxation simulated according to the terms of the rate matrix. Overall relaxation rate obtained from this decay was 1192 s^{-1}. A similar analysis for En-HD and gpW gave overall relaxation rates 5751 s^{-1} and 4108 s^{-1} respectively. This exercise of analysing the overall relaxation rate for downhill folding scenario is performed to demonstrate the ability of 1D-FES model to capture the complex folding dynamics observed for downhill folders. However, the parameter of significance with respect to identifying the dominant factor among secondary structure and topology is intramolecular diffusion coefficient, *D*.

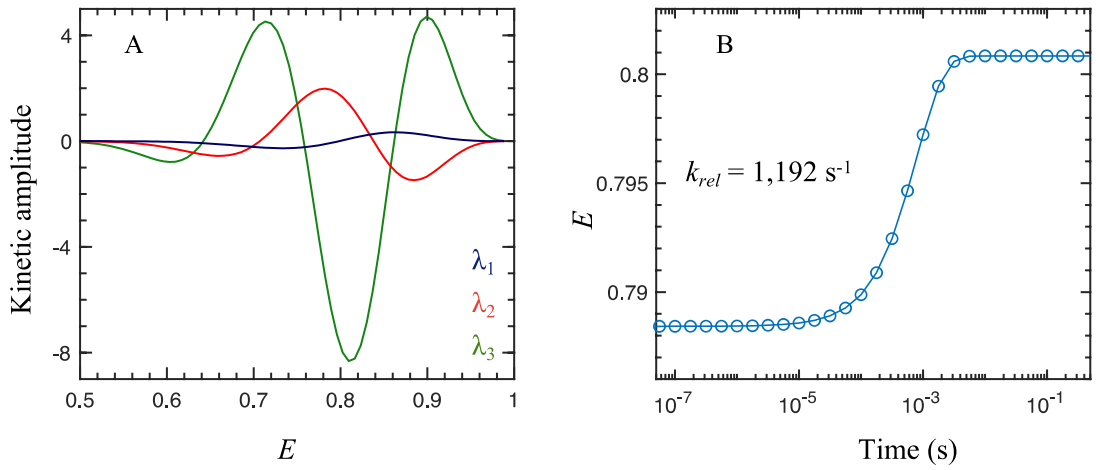


Figure 5-7 : Rate analysis for Nedd4-WW4 (A) Kinetic amplitudes of the three slowest non-zero eigenvalues obtained from the rate matrix (B) Relaxation kinetics simulated with the rates from the matrix **K**

D is the parameter that essentially determines the timescale of the motion along the FES (Table 3). We find that En-HD, composed of 3 helices, has the largest *D* followed by gpW, consisting of 2 helices and one β-hairpin while the WW domains which are all-beta proteins have the slowest intramolecular diffusion coefficient. The rate of formation of α-helices is known to be faster than that of β-sheet. *D* for both the WW domains, with an antiparallel arrangement of sheet-sheet-sheet, were found to be different from the other two proteins, En-HD and gpW, which have an antiparallel arrangement of helix-helix-helix and helix-hairpin-helix respectively. This indicates the secondary structural elements to be playing the determinant role in the folding mechanism, among the four proteins studied here.

5.3 Conclusion

We employed MLA with the 1D-FES model to probe the effect of topology and secondary structural elements on the folding mechanism. While all three major structural classes probed showed folding under minimal or no fee energy barrier, the dynamic factors including the intramolecular diffusion coefficient (*D*) and various relaxation processes along the FES were found to be different. A comparison of *D* among three classes of secondary structures suggests the secondary structural elements dictating the global folding motions and mechanisms for the proteins under consideration, with all-helical protein showing the fastest *D*, followed by helix-hairpin and all-beta being the slowest.

Chapter 6

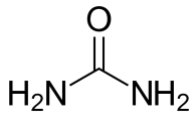
6 Effects of denaturants and ionic strength on the folding dynamics of Engrailed-HD.

6.1 Introduction

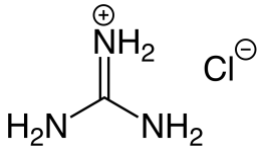
Different terms involved in the folding rate expression contain complex energetics, with entropic and dynamic contributions, and cannot be easily rationalized from first principles. To fundamentally understand such rates and the terms involved, it is essential to empirically characterize how they depend on physical-chemical factors (temperature, cosolvents, ionic strength) and protein structure and sequence. However, it has been impossible to directly deconstruct the effects of these factors on all the rate terms by experiment. As a result, experimentalists have followed traditional assumptions from chemical reaction rate theory to account for the effects of different factors on the rate. The folding rate expression is usually broadly divided into two terms, the exponential term depending on the free energy barrier for folding and a preexponential term. Effects of temperature, cosolvents/denaturants, structure, and mutations are ascribed to the barrier (exponential term). At the same time, solvent viscosity and ionic strength are usually assigned to be determinants of the preexponential term[175]. A detailed analysis of the effect of these factors on different terms of rate expression is long overdue.

Using chemical denaturants to unfold proteins is a commonly used technique in protein biochemistry research [176]. Early experiments aimed at studying protein stability involved using chemical denaturants to access the unfolded state and measuring the difference between the stability of the native state and unfolded state[177,178]. Many of the mutational studies also use chemical denaturants to measure the difference in stability induced by mutations. Urea and guanidinium hydrochloride (GdnCl) are two commonly used chemical denaturants to study protein stability. The mode of action of urea and guanidinium hydrochloride in denaturing proteins is still indeterminate despite both chemicals being used frequently in protein folding studies and has been a topic of discussion for many decades now[179–182]. Guanidinium hydrochloride and urea are weekly interacting molecules with very small free energies of interaction between the denaturant and protein residues, and it takes molar concentration of both to denature proteins. Moreover, there is a scarcity of structures available in the presence of denaturants. These have made it challenging to study the mechanism of their action. Whether the unfolding results from direct interaction with the protein chain or inducing changes in the solvent environment is still unclear.

Even though there are considerable structural similarities between guanidinium hydrochloride and urea, they both are different in their chemistry, with the former being a salt and latter a polar molecule. It doesn't surprise that they can have very different effects on the protein (un)folding process due to the difference in ionic strength, especially in the case of proteins that are stabilized by multiple electrostatic interactions.



Urea



Guanidinium hydrochloride

Protein stability and kinetics are strongly influenced by the solvent ionic strength. This is expected as proteins are mixed charge polymers. Simple monovalent salts, like sodium chloride, can have stabilizing or destabilizing effect on protein stability depending on the charge distribution within the biomolecule[183–185]. Debye-Huckel relation, which describes the screening of Coulombic interaction through the distribution of counterions formed around a point charge, is utilized to explain this behaviour. Proteins with net destabilizing electrostatic interactions see an increase in the stability of native state in presence of salt because of effective screening of charges. However, probing the effect of ionic strength on different terms of rate equation (particularly preexponential factor) along with direct estimation of protein stability is required to better understand their mechanism of action.

Separating the effects of denaturants and ionic strength on kinetics and dynamics have been challenging as traditional experimental probes cannot measure them directly. Implementation of the 1D-FES method to analyse SM-FRET data allows us to obtain the folding rates and dynamics at any given condition directly along with the corresponding FES. This can help in dissecting the effect each denaturant, urea and GdnCl, and ionic strength have on the stability and kinetics of protein folding.

We carried out SM measurements on the home-built SM-FRET set-up as per the procedures described in earlier chapters in the presence of different denaturants and salt. The methodology of utilizing MLA with the 1D-FES model established in chapter 2 was used to analyse all data discussed in this chapter. Data analysis followed the same procedure, starting with the identification of photon bursts, screening the bursts to remove the bursts with photophysical artifacts, followed by MLA analysis with a 1D-FES model to extract the free energy surface and the intramolecular diffusion coefficient. The MLA method along with a simple two-state kinetic model was also done for extracting folding-unfolding rates for comparative studies.

6.2 Results and Discussion

FEH obtained for midpoint denaturation condition in urea and GdnCl was significantly different, suggesting a difference in the timescale of dynamics (Figure 6-1). While the FEH for urea showed two populations not well separated on the FRET efficiency axis, FEH in GdnCl showed two relatively well-separated populations. *E* for both folded and unfolded population was found to be lower in GdnCl, suggesting an expansion of the conformational ensembles. MLA with a simple two-state model showed that the dynamics of En-HD in GdnCl is 2.5 times slower compared to that of urea with relaxation rates 2010 s^{-1} and 4786 $s^{-1,}$ respectively, confirming the difference in their dynamics.

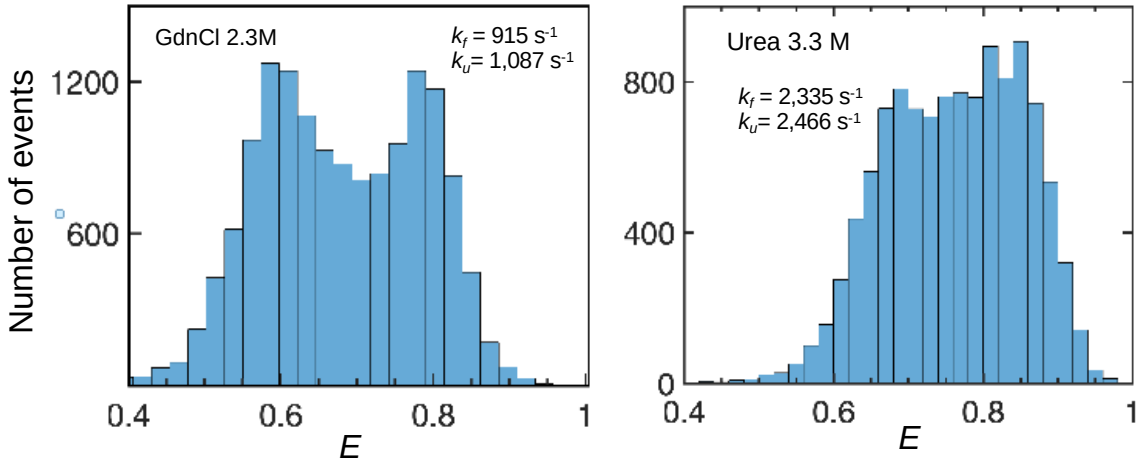


Figure 6-1 : FEH for bursts collected for En-HD at midpoint denaturation condition (C_m) in GdnCl and urea. Mean residence time for the bursts shown here is ~0.6 ms

MLA with 1D-FES model was carried out to understand whether the origin of this difference in rate is because of a difference in thermodynamic stability or because of a change in the preexponential factor of the rate equation, a unique ability of FES based approaches. The FESs and the dynamic factors were obtained for both conditions(Figure 6-2 and Table 4). Interestingly, there was a difference in the timescale of various relaxation processes and the intramolecular diffusion coefficient from the model (Table 4) apart from the difference in the barrier.

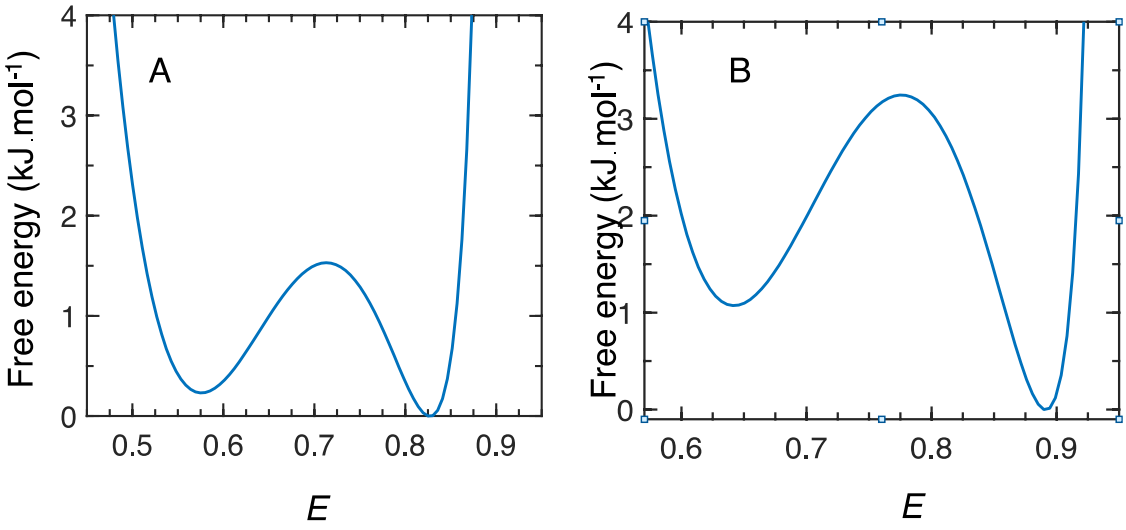


Figure 6-2: FES obtained from MLA with 1D-FES model for En-HD at midpoint denaturation concentration (A) GdnCl (B) Urea

	$\Delta G_F^{\ddagger}$ (kJ mol^{-1})	$\Delta G_U^{\ddagger}$ (kJ mol^{-1})	log D	λ_1 (s^{-1})	λ_2(s^{-1})	λ_3(s^{-1})
GdnCl	1.3	1.5	1.99	1603	8657	17616
Urea	2.1	3.1	2.63	4557	34834	71632

Table 4 : Parameters from 1D-FES analysis. λ_1 λ_2 and λ_3 are the slowest three relaxation rates obtained from the rate matrix (section 2.4.3).

Folding and unfolding barrier obtained from the FES was 0.4 *RT* and 0.6 *RT* higher in urea than in GdnCl. This difference in the barrier accounts for a 1.5-fold and 1.8-fold decrease in the folding and unfolding rates in urea if the effect of denaturant was only on the barrier. However, λ_1 obtained from the model (corresponds to the relaxation rate between the folded and unfolded well, refer to section 2.4.3) was shown to be 2.8-fold faster in urea. This suggests that the denaturants have a significant effect on the preexponential factor of the rate equation. This is also reflected in the values of *D*, with the protein in GdnCl showing a slower intramolecular diffusion coefficient (4.4-fold slower) for the chain dynamics. The difference in λ_2 and λ_3, which corresponds to the barrier depopulation process and the dynamics within the free energy minima, respectively, also confirm this slowdown of dynamics in GdnCl.

En-HD is positively charged at the experimental conditions (10 positive charges). The most notable difference between the chemical nature of urea and GdnCl is the difference in ionic strength between them. GdnCl is a salt and introduces ionic species in the solution, while urea is a polar molecule. To account for the ionic strength introduced into the solution with GdnCl, we collected SM-FRET data in the presence of 2.3M NaCl at urea midpoint concentrations (C_m in GdnCl is 2.3 M).

A relatively high concentration of urea, 7 M, had to be used to achieve near C_m conditions in the presence of 2.3 M NaCl (Figure 6-3). It is important to note that a solution with 7 M urea with 2.3 M NaCl has higher solvent viscosity, which itself influences the dynamics. Concentration of urea was not increased beyond 7 M as SM-FRET measurements in such concentrated solutions have practical limitations (increased scattering resulting in poor signal to noise ratio and crystallization in the solution effectively changing the ionic strength). The increase in the stability of the protein can be due to the effective electrostatic screening effect in high salt concentration, stabilizing the positive charge of the protein. FEH in the presence of 2.3 M salt at near ~ C_m in urea clearly shows two populations similar to what was obtained at GdnCl midpoint conditions, hinting at a slowdown of the dynamics (Figure 6-3A). We also notice an expansion of structural ensembles as evident from the lower *E* in the presence of salt.

A comparison of the FESs at urea midpoint with and without salt showed minimal difference in the free energy barrier (Figure 6-3 : 0.2 *RT* higher for unfolding and 0.1 *RT* lower for unfolding in presence of 2.3 M NaCl, this difference is due to 7 M urea being slightly lower than C_m thus making the conditions slightly folding). Meanwhile, the intramolecular diffusion coefficient, *D*, was 6-fold slower in presence of 2.3 M salt (Table 5). Similarly, a 6.4-fold decrease was found in λ_1. This suggest that the slowdown of dynamics is entirely coming from the preexponential factor. The preexponential factor for protein folding has contributions from the solvent viscosity, apart from the internal friction. Relative viscosity of 7 M urea solution with 2.3 M salt is only 1.6-fold higher than 3.3 M urea solution[186]. Hence it is the increased internal friction that predominantly contributed to the slowdown of dynamics of En-HD in high ionic strength solutions. The increase in the stability and a slowdown of dynamics along with the expansion of conformations observed in presence of high ionic strength solutions is an interesting observation. While the effective electrostatic screening in presence of salt stabilizes the native fold, the charge screening also resulted in slower rates. This suggests that the charge distribution present on En-HD accelerates the protein chain dynamics and screening of this charge increases the internal friction.

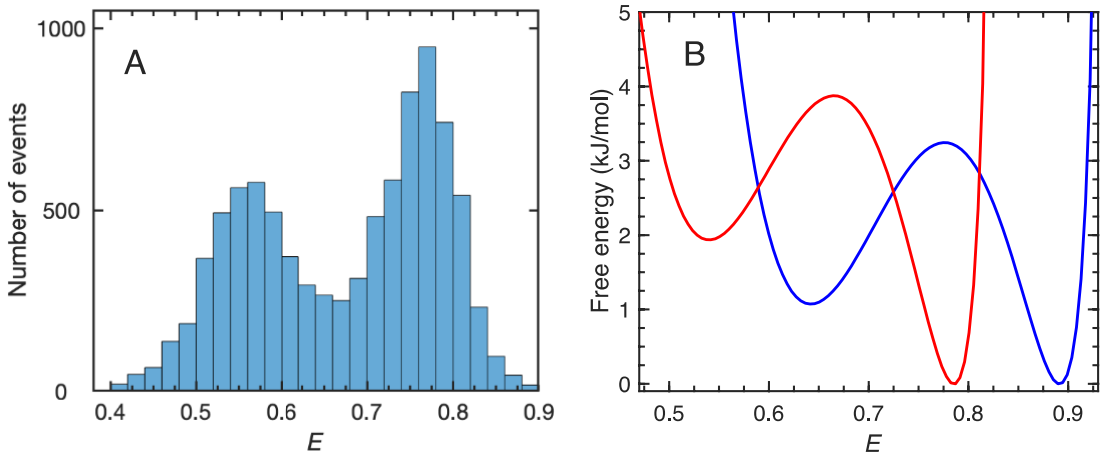


Figure 6-3: SM-FRET analysis at 7 M urea (~C_m) in presence of 2.3 M NaCl (A) FEH for the bursts collected (B) FES obtained from 1D-FES model in 7 M urea + 2.3 NaCl (red). Blue curve shows the FES at urea C_m discussed earlier (3.3 M urea) for comparison

NaCl (M)	log *D*	λ_1 (s^{-1})	λ_2 (s^{-1})	λ_3 (s^{-1})
0.1	2.63	4557	34834	71632
2.3	1.85	714	5833	12009

Table 5 : Parameters from 1D-FES analysis near C_m in urea at different NaCl concentration

6.3 Conclusion

En-HD shows different dynamics at midpoint denaturation concentration of urea and GdnCl (salt), two commonly used chemical denaturants in protein folding studies. ML analysis with 1D-FES model identified the origin of this change to be from the preexponential factor of the folding rate equation. The protein shows slower dynamics in GdnCl, possibly because of the high ionic strength of GdnCl solutions. A slowdown of En-HD dynamics was observed in the presence of NaCl under urea denaturation midpoint. This confirms that the reason for the difference in the dynamics in urea and GdnCl is the difference in ionic strength of both solutions. Interestingly, the slowdown in the dynamics is accompanied by an increase in the stability of the protein. The increase in the stability of the protein in presence of high ionic strength is explained by the effective electrostatic screening of the positive charge on the protein. Charge screening stabilizes the native fold, thus increasing the stability.

1D-FES model with MLA showed differences in the dynamic factors including intra molecular diffusion coefficient, D. The origin of the difference in the relaxation rates was thus determined to be from the preexponential term of the rate expression. We propose conducting additional SM-FRET experiments on En-HD with FRET acceptor-donor pair introduced between different helices. The additional distance information would enable monitoring the changes in the packing of helices.

Bibliography

1. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662–666 (1958).

2. Finkelstein, A. V. 50+ Years of Protein Folding. *Biochemistry (Moscow)* **83**, (2018).

3. Dill, K. A. & MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science (80-.).* **338**, 1042–1046 (2012).

4. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).

5. Seringhaus, M. & Gerstein, M. Chemistry nobel rich in structure [5]. *Science* **315**, 40–41 (2007).

6. Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D. & Voelz, V. A. The protein folding problem: when will it be solved? *Current Opinion in Structural Biology* **17**, 342–346 (2007).

7. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Bioinforma.* **21**, 167–195 (1995).

8. Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 7524–7528 (1987).

9. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science (80-.).* **334**, 517–520 (2011).

10. Shaw, D. E. *et al.* Atomic-level characterization of the structural dynamics of proteins. *Science (80-.).* **330**, 341–346 (2010).

11. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17845–17850 (2012).

12. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).

13. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science (80-.).* **309**, 1868–1871 (2005).

14. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588**, 203–204 (2020).

15. Service, R. Google's DeepMind aces protein folding. *Science (80-.).* (2018). doi:10.1126/science.aaw2747

16. Uversky, V. N. Intrinsically disordered proteins and their 'Mysterious' (meta)physics. *Frontiers in Physics* **7**, 10 (2019).

17. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1804**, 1231–1264 (2010).

18. Peng, Z. *et al.* Exceptionally abundant exceptions: Comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **72**, 137–151 (2014).

19. Uversky, V. N. Intrinsic Disorder-based Protein Interactions and their Modulators. *Curr. Pharm. Des.* **19**, 4191–4213 (2013).

20. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).

21. Wolynes, P. G. Chapter 3. The Protein Folding Energy Landscape: A Primer. in *Protein Folding, Misfolding and Aggregation* 49–69 (Royal Society of Chemistry, 2008). doi:10.1039/9781847558282-00049

22. Shakhnovich, E. Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chemical Reviews* **106**, 1559–1588 (2006).

23. Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci.* **84**, (1987).

24. Dill, K. A. & Chan, H. S. From levinthal to pathways to funnels. *Nature Structural Biology* **4**, 10–19 (1997).

25. Sadqi, M. *et al.* Atom-by-atom analysis of global downhill protein folding. *Nature* **442**, 317–321 (2006).

26. Garcia-Mira, M. M., Sadqi, M., Fischer, N., Sanchez-Ruiz, J. M. & Muñoz, V.

Experimental identification of downhill protein folding. *Science (80-.).* **298**, 2191–2195 (2002).

27. Wolynes, P. G. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **119**, 218–230 (2015).

28. Davidson, A. R. & Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci.* **91**, (1994).

29. LaBean, T. H., Kauffman, S. A. & Butt, T. R. Libraries of random-sequence polypeptides produced with high yield as carboxy-terminal fusions with ubiquitin. *Mol. Divers.* **1**, 29–38 (1995).

30. Taketomi, H., Ueda, Y. & Gō, N. STUDIES ON PROTEIN FOLDING, UNFOLDING AND FLUCTUATIONS BY COMPUTER SIMULATION: I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **7**, 445–459 (1975).

31. Schafer, N. P., Kim, B. L., Zheng, W. & Wolynes, P. G. Learning to fold proteins using energy landscape theory. *Israel Journal of Chemistry* **54**, 1311–1337 (2014).

32. Ferreiro, D. U., Hegler, J. A., Komives, E. A. & Wolynes, P. G. On the role of frustration in the energy landscapes of allosteric proteins. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 3499–3503 (2011).

33. Ferreiro, D. U., Hegler, J. A., Komives, E. A. & Wolynes, P. G. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19819–19824 (2007).

34. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860–5868 (1996).

35. Best, R. B. & Hummer, G. Diffusion models of protein folding. *Phys. Chem. Chem. Phys.* **13**, 16902–16911 (2011).

36. Best, R. B. & Hummer, G. Diffusive model of protein folding dynamics with Kramers turnover in rate. *Phys. Rev. Lett.* **96**, (2006).

37. Best, R. B. & Hummer, G. Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1088–1093 (2010).

38. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways,

and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Bioinforma.* **21**, 167–195 (1995).

39. Kubelka, J., Hofrichter, J. & Eaton, W. A. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* **14**, 76–88 (2004).

40. Li, P. *et al.* Dynamics of one-state downhill protein folding. *Proc. Natl. Acad. Sci.* **106**, 103–108 (2009).

41. Schuler, B. Single-molecule FRET of protein structure and dynamics - a primer. *J. Nanobiotechnology* **11**, S2 (2013).

42. Neuman, K. C. & Nagy, A. Single-molecule force spectroscopy: Optical tweezers, magnetic tweezers and atomic force microscopy. *Nature Methods* **5**, 491–505 (2008).

43. Greenleaf, W. J., Woodside, M. T. & Block, S. M. High-resolution, single-molecule measurements of biomolecular motion. *Annual Review of Biophysics and Biomolecular Structure* **36**, 171–190 (2007).

44. Jacob, M. & Schmid, F. X. Protein folding as a diffusional process. *Biochemistry* **38**, 13773–13779 (1999).

45. Hänggi, P., Talkner, P. & Borkovec, M. Reaction-rate theory: Fifty years after Kramers. *Rev. Mod. Phys.* **62**, 251–341 (1990).

46. Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7**, 284–304 (1940).

47. Chung, H. S. & Eaton, W. A. Single-molecule fluorescence probes dynamics of barrier crossing. *Nature* **502**, 685–688 (2013).

48. Liu, F., Nakaema, M. & Gruebele, M. The transition state transit time of WW domain folding is controlled by energy landscape roughness. *J. Chem. Phys.* **131**, 195101 (2009).

49. de Sancho, D., Sirur, A. & Best, R. B. Molecular origins of internal friction effects on protein-folding rates. *Nat. Commun.* **5**, (2014).

50. Zheng, W., De Sancho, D., Hoppe, T. & Best, R. B. Dependence of Internal Friction on Folding Mechanism. *J. Am. Chem. Soc.* **137**, 3283–3290 (2015).

51. Chung, H. S., McHale, K., Louis, J. M. & Eaton, W. A. Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science (80-.).* **335**, 981–984 (2012).

52. Hummer, G. From transition paths to transition states and rate coefficients. *J. Chem. Phys.* **120**, 516–523 (2004).

53. Shaw, D. E. *et al.* Atomic-Level Characterization. 341–347 (2010).

54. Chung, H. S., Piana-Agostinetti, S., Shaw, D. E. & Eaton, W. A. Structural origin of slow diffusion in protein folding. *Science (80-.).* **349**, 1504–1510 (2015).

55. Best, R. B. *Atomistic molecular simulations of protein folding*. *Current Opinion in Structural Biology* **22**, 52–61 (Elsevier Current Trends, 2012).

56. Muñoz, V., Sadqi, M., Naganathan, A. N. & de Sancho, D. Exploiting the downhill folding regime via experiment. *HFSP J.* **2**, 342–353 (2008).

57. Munoz, V., Cerminara, M., Muñoz, V. & Cerminara, M. When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. *Biochem. J.* **473**, 2545–2559 (2016).

58. Liu, F. *et al.* An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2369–2374 (2008).

59. Liu, F. & Gruebele, M. Downhill dynamics and the molecular rate of protein folding. *Chem. Phys. Lett.* **461**, 1–8 (2008).

60. Kjaergaard, M., Teilum, K. & Poulsen, F. M. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc. Natl. Acad. Sci.* **107**, 12535 LP – 12540 (2010).

61. Cerminara, M., Desai, T. M., Sadqi, M. & Muñoz, V. Downhill protein folding modules as scaffolds for broad-range ultrafast biosensors. *J. Am. Chem. Soc.* **134**, 8010–8013 (2012).

62. Prigozhin, M. B. & Gruebele, M. Microsecond folding experiments and simulations: a match is made. *Phys. Chem. Chem. Phys.* **15**, 3372 (2013).

63. Garvey, E. P. & Matthews, C. R. Effects of Multiple Replacements at a Single Position on the Folding and Stability of Dihydrofolate Reductase from Escherichia

coli. *Biochemistry* **28**, 2083–2093 (1989).

64. Goldenberg, D. P., Frieden, R. W., Haack, J. A. & Morrison, T. B. Mutational analysis of a protein-folding pathway. *Nature* **338**, 127–132 (1989).

65. Chung, H. S. & Eaton, W. A. Protein folding transition path times from single molecule FRET. *Curr. Opin. Struct. Biol.* **48**, 30–39 (2018).

66. Buchner, G. S., Murphy, R. D., Buchete, N. V. & Kubelka, J. Dynamics of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1814**, 1001–1020 (2011).

67. Eaton, W. A. *et al.* Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327–359 (2000).

68. A. Royer, C. Probing Protein Folding and Conformational Transitions with Fluorescence. *Chem. Rev.* **106**, 1769–1784 (2006).

69. Brockwell, D. J., Smith, D. A. & Radford, S. E. Protein folding mechanisms: New methods and emerging ideas. *Current Opinion in Structural Biology* **10**, 16–25 (2000).

70. Barth, A. & Zscherp, C. What vibrations tell us about proteins. *Quarterly Reviews of Biophysics* **35**, 369–430 (2002).

71. Palmer, A. G. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **104**, 3623–3640 (2004).

72. Mittermaier, A. & Kay, L. E. New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224–228 (2006).

73. Huang, G. S. & Oas, T. G. Submillisecond folding of monomeric λ repressor. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 6878–6882 (1995).

74. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L. & Oas, T. G. The energy landscape of a fast-folding protein mapped by Ala→Gly substitutions. *Nat. Struct. Biol.* **4**, 305–310 (1997).

75. Myers, J. K. & Oas, T. G. Mechanisms of Fast Protein Folding. *Annu. Rev. Biochem.* **71**, 783–815 (2002).

76. Gillespie, B. *et al.* NMR and temperature-jump measurements of de novo designed proteins demonstrate rapid folding in the absence of explicit selection for kinetics. *J. Mol. Biol.* **330**, 813–819 (2003).

77. Myers, J. K. Chemical Denaturation. in *Molecular Life Sciences* 75–80 (Springer New York, 2014). doi:10.1007/978-1-4614-1531-2_646

78. Naganathan, A. N., Doshi, U. & Muñoz, V. Protein Folding Kinetics: Barrier Effects in Chemical and Thermal Denaturation Experiments. *J. Am. Chem. Soc.* **129**, 5673–5682 (2007).

79. Roder, H. *et al.* Rapid mixing methods for exploring the kinetics of protein folding. *Methods* **34**, 15–27 (2004).

80. Shastry, M. C. R. C. R., Luck, S. D. & Roder, H. A Continuous-Flow Capillary Mixing Method to Monitor Reactions on the Microsecond Time Scale. *Biophys. J.* **74**, 2714–2721 (1998).

81. Jacob, M. *et al.* Microsecond folding of the cold shock protein measured by a pressure- jump technique. *Biochemistry* **38**, 2882–2891 (1999).

82. Clegg, R. M. & Maxfield, B. W. Chemical kinetic studies by a new small pressure perturbation method. *Rev. Sci. Instrum.* **47**, 1383–1393 (1976).

83. Nölting, B., Golbik, R. & Fersht, A. R. Submillisecond events in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 10668–10672 (1995).

84. Kubelka, J. Time-resolved methods in biophysics. 9. Laser temperature-jump methods for investigating biomolecular dynamics. *Photochem. Photobiol. Sci.* **8**, 499–512 (2009).

85. Hofrichter, J. Laser temperature-jump methods for studying folding dynamics. *Methods in molecular biology (Clifton, N.J.)* **168**, 159–191 (2001).

86. Weikl, T. R. & Dill, K. A. Transition-States in Protein Folding Kinetics: The Structural Interpretation of Φ values. *J. Mol. Biol.* **365**, 1578–1586 (2007).

87. Naganathana, A. N. & Muñoz, V. Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8611–8616 (2010).

88. Fersht, A. R. & Sato, S. Φ-value analysis and the nature of protein-folding transition

states. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 7976–7981 (2004).

89. Borgia, A., Williams, P. M. & Clarke, J. Single-molecule studies of protein folding. *Annual Review of Biochemistry* **77**, 101–125 (2008).

90. Ferreon, A. C. M. & Deniz, A. A. Protein folding at single-molecule resolution. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1814**, 1021–1029 (2011).

91. Chris, A. *et al.* Protein folding at single-molecule resolution. *Biochim. Biophys. acta. Proteins proteomics* **1814**, 1021–1029 (2011).

92. Yu, H. *et al.* Energy landscape analysis of native folding of the prion protein yields the diffusion constant, transition path time, and rates. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14452–14457 (2012).

93. Neupane, K. *et al.* Direct observation of transition paths during the folding of proteins and nucleic acids. *Sci. (American Assoc. Adv. Sci.* **352**, 239–242 (2016).

94. Greenleaf, W. J., Woodside, M. T. & Block, S. M. High-resolution, single-molecule measurements of biomolecular motion. *Annual Review of Biophysics and Biomolecular Structure* **36**, 171–190 (2007).

95. Huang, F., Sato, S., Sharpe, T. D., Ying, L. & Fersht, A. R. Distinguishing between cooperative and unimodal downhill protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 123–127 (2007).

96. Schuler, B., Lipman, E. A., Eaton, W. A., a. Lipman, E. & a. Eaton, W. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **419**, 743–747 (2002).

97. Deniz, A. A. *et al.* Single-molecule protein folding: Diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5179–5184 (2000).

98. Rounsevell, R., Forman, J. R. & Clarke, J. Atomic force microscopy: Mechanical unfolding of proteins. *Methods* **34**, 100–111 (2004).

99. Ng, S. P., Randles, L. G. & Clarke, J. Single molecule studies of protein folding using atomic force microscopy. *Methods Mol. Biol.* **350**, 139–167 (2007).

100. Neuman, K. C. & Nagy, A. Single-molecule force spectroscopy: Optical tweezers, magnetic tweezers and atomic force microscopy. *Nature Methods* **5**, 491–505

(2008).

101. Bustamante, C., Chemla, Y. R. & Moffitt, J. R. High-resolution dual-trap optical tweezers with differential detection: An introduction. *Cold Spring Harb. Protoc.* **4**, pdb.top60 (2009).

102. Ritchie, D. B. & Woodside, M. T. Probing the structural dynamics of proteins and nucleic acids with optical tweezers. *Current Opinion in Structural Biology* **34**, 43–51 (2015).

103. Ha, T. *et al.* Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 6264–8 (1996).

104. Chung, H. S. & Gopich, I. V. Fast single-molecule FRET spectroscopy: Theory and experiment. *Phys. Chem. Chem. Phys.* **16**, 18644–18657 (2014).

105. Chung, H. S. Transition Path Times Measured by Single-Molecule Spectroscopy. *J. Mol. Biol.* **430**, 409–423 (2018).

106. Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* **45**, 207–231 (2016).

107. Kapanidis, A. N. & Strick, T. Biology, one molecule at a time. *Trends Biochem. Sci.* **34**, 234–243 (2009).

108. Ha, T., Kozlov, A. G. & Lohman, T. M. Single-molecule views of protein movement on single-stranded DNA. *Annu. Rev. Biophys.* **41**, 295–319 (2012).

109. Henzler-Wildman, K. A. *et al.* Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**, 838–844 (2007).

110. Rahul, R., Hohng, S. & Ha, T. A Practical Guide to Single Molecule FRET. *Nat. Methods* **5**, 507–516 (2008).

111. Schuler, B., Eaton, W. a, a. Eaton, W. & Eaton, W. a. Protein folding studied by single molecule FRET. *Curr. Opin. Struct. Biol.* **18**, 16–26 (2008).

112. Campos, L. A. *et al.* A photoprotection strategy for microsecond-resolution single-molecule fluorescence spectroscopy. *Nat. Methods* **8**, 143–146 (2011).

113. Rasnik, I., McKinney, S. A. & Ha, T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods* **3**, 891–893 (2006).

114. Kong, X., Nir, E., Hamadani, K. & Weiss, S. Photobleaching pathways in single-molecule FRET experiments. *J. Am. Chem. Soc.* **129**, 4643–4654 (2007).

115. Lakowicz, J. R. *Principles of fluorescence spectroscopy*. *Principles of Fluorescence Spectroscopy* (Springer, 2006). doi:10.1007/978-0-387-46312-4

116. Stryer, L. & Haugland, R. P. Energy transfer: a spectroscopic ruler. *Proc. Natl. Acad. Sci. U. S. A.* **58**, (1967).

117. Ha, T. *et al.* Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci.* **93**, (1996).

118. Tyagi, S. & Lemke, E. A. Genetically Encoded Click Chemistry for Single-Molecule FRET of Proteins. in *Methods in Cell Biology* **113**, 169–187 (Academic Press Inc., 2013).

119. Schuler, B. & Eaton, W. A. Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.* **18**, 16–26 (2008).

120. Lemke, E. A. Site-specific labeling of proteins for single-molecule fret measurements using genetically encoded ketone functionalities. in *Methods in Molecular Biology* **751**, 3–15 (Humana Press Inc., 2011).

121. Gopich, I. V. & Szabo, A. FRET efficiency distributions of multistate single molecules. *J. Phys. Chem. B* **114**, 15221–15226 (2010).

122. Chung, H. S. *et al.* Extracting rate coefficients from single-molecule photon trajectories and FRET efficiency histograms for a fast-folding protein. *J. Phys. Chem. A* **115**, 3642–3656 (2011).

123. Gopich, I. V & Szabo, A. Single-Molecule FRET with Diffusion and Conformational Dynamics. *J. Phys. Chem. B* **111**, 12925–12932 (2007).

124. Gopich, I. V. & Szabo, A. Decoding the pattern of photon colors in single-molecule FRET. *J. Phys. Chem. B* **113**, 10965–10973 (2009).

125. Ramanathan, R. & Muñoz, V. A Method for extracting the free energy surface and conformational dynamics of fast-folding proteins from single molecule photon

trajectories. *J. Phys. Chem. B* **119**, 7944–7956 (2015).

126. Sancho, D. De *et al.* Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys. Chem. Chem. Phys.* **13**, 17030–17043 (2011).

127. Muñoz, V. & Sanchez-ruiz, J. M. Exploring Protein-Folding Ensembles: A Variable-Barrier Model for the Analysis of Equilibrium Unfolding Experiments. *Proc. Natl. Acad. Sci. - PNAS* **101**, 17646–17651 (2004).

128. Religa, T. L. *et al.* The Helix-Turn-Helix Motif as an Ultrafast Independently Folding Domain: The Pathway of Folding of Engrailed Homeodomain. *Proc. Natl. Acad. Sci. - PNAS* **104**, 9272–9277 (2007).

129. Mayor, U. *et al.* The complete folding pathway of a protein from nanoseconds to microseconds. *Nature* **421**, 863–867 (2003).

130. Mayor, U., Johnson, C. M., Daggett, V. & Fersht, A. R. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13518–13522 (2000).

131. Gianni, S. *et al.* Unifying Features in Protein-Folding Mechanisms. *Proc. Natl. Acad. Sci. - PNAS* **100**, 13286–13291 (2003).

132. Naganathan, A. N. & Muñoz, V. Thermodynamics of Downhill Folding: Multi-Probe Analysis of PDD, a Protein that Folds Over a Marginal Free Energy Barrier. *J. Phys. Chem. B* **118**, 8982–8994 (2014).

133. Möglich, A., Joder, K. & Kiefhaber, T. End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12394–12399 (2006).

134. Harrison, S. C. & Durbin, R. Is there a single pathway for the folding of a polypeptide chain? *Proc. Natl. Acad. Sci. U. S. A.* **82**, 4028–4030 (1985).

135. Karplus, M. & Weaver, D. L. Diffusion–collision model for protein folding. *Biopolymers* **18**, 1421–1437 (1979).

136. Udgaonkar, J. B. Multiple routes and structural heterogeneity in protein folding. *Annual Review of Biophysics* **37**, 489–510 (2008).

137. Dinner, A. R., Šalib, A., Smitha, L. J., Dobsona, C. M. & Karplus, M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in Biochemical Sciences* **25**, 331–339 (2000).

138. Hoffer, N. Q., Neupane, K., Pyo, A. G. T. T. & Woodside, M. T. Measuring the average shape of transition paths during the folding of a single biological molecule. *Proc. Natl. Acad. Sci. - PNAS* **116**, 8125–8130 (2019).

139. Cossio, P., Hummer, G. & Szabo, A. Transition paths in single-molecule force spectroscopy. *J. Chem. Phys.* **148**, 123309 (2018).

140. Kim, J.-Y. Y. & Chung, H. S. Disordered proteins follow diverse transition paths as they fold and bind to a partner. *Science (80-.).* **368**, 1253–1257 (2020).

141. Sturzenegger, F. *et al.* Transition path times of coupled folding and binding reveal the formation of an encounter complex. *Nat. Commun.* **9**, 1–11 (2018).

142. Chung, H. S., Louis, J. M., Eaton, W. A. & Fersht, A. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci.* **106**, 11837–11844 (2009).

143. Fung, A. *et al.* Expanding the realm of ultrafast protein folding: gpW, a midsize natural single-domain with α+β topology that folds downhill. *J. Am. Chem. Soc.* **130**, 7489–7495 (2008).

144. Liu, F. & Gruebele, M. Tuning λ6-85 Towards Downhill Folding at its Melting Temperature. *J. Mol. Biol.* **370**, 574–584 (2007).

145. Liu, F., Gao, Y. G. & Gruebele, M. A Survey of λ Repressor Fragments from Two-State to Downhill Folding. *J. Mol. Biol.* **397**, 789–798 (2010).

146. Neupane, K. *et al.* Transition path times for nucleic acid folding determined from energy-landscape analysis of single-molecule trajectories. *Phys. Rev. Lett.* **109**, 068102 (2012).

147. Kirmizialtin, S., Huang, L. & Makarov, D. E. Topography of the free-energy landscape probed via mechanical unfolding of proteins. *J. Chem. Phys.* **122**, 234915 (2005).

148. Liu, J. *et al.* Exploring one-state downhill protein folding in single molecules. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 179–184 (2012).

149. Akmal, A. & Muñoz, V. The nature of the free energy barriers to two-state folding. *Proteins Struct. Funct. Genet.* **57**, 142–152 (2004).

150. Muñoz, V. Conformational Dynamics and Ensembles in Protein Folding. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 395–412 (2007).

151. Jackson, S. E. How do small single-domain proteins fold? *Fold. Des.* **3**, R81–R91 (1998).

152. Gillespie, D. T. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55 (2007).

153. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. in *Journal of Physical Chemistry* **81**, 2340–2361 (American Chemical Society, 1977).

154. Campos, L. A. *et al.* Gradual disordering of the native state on a slow two-state folding protein monitored by single-molecule fluorescence spectroscopy and NMR. *J. Phys. Chem. B* **117**, 13120–13131 (2013).

155. Makarov, D. E. Shapes of dominant transition paths from single-molecule force spectroscopy. *J. Chem. Phys.* **143**, 194103 (2015).

156. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350 (1998).

157. Campos, L. A., Sadqi, M. & Munõz, V. Lessons about Protein Folding and Binding from Archetypal Folds. *Acc. Chem. Res.* **53**, 2180–2188 (2020).

158. Chaudhury, S. & Makarov, D. E. A harmonic transition state approximation for the duration of reactive events in complex molecular rearrangements. *J. Chem. Phys.* **133**, 034118 (2010).

159. Kim, W. K. & Netz, R. R. The mean shape of transition and first-passage paths. *J. Chem. Phys.* **143**, 224108 (2015).

160. Taumoefolau, G. H. & Best, R. B. Estimating transition path times and shapes from single-molecule photon trajectories: A simulation analysis. *J. Chem. Phys.* **154**, 115101 (2021).

161. Kim, J.-Y. Y., Meng, F., Yoo, J. & Chung, H. S. Diffusion-limited association of disordered protein by non-native electrostatic interactions. *Nat. Commun.* **9**, 1–10 (2018).

162. Hummer, G. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.* **7**, 34 (2005).

163. Clementi, C., García, A. E., Onuchic, J. N., García, A. E. & Onuchic, J. N. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L. *J. Mol. Biol.* **326**, 933–954 (2003).

164. Cheung, M. S., Finke, J. M., Callahan, B. & Onuchic, J. N. Exploring the Interplay between Topology and Secondary Structural Formation in the Protein Folding Problem. *J. Phys. Chem. B* **107**, 11193–11200 (2003).

165. Gunasekaran, K., Eyles, S. J., Hagler, A. T. & Gierasch, L. M. Keeping it in the family: Folding studies of related proteins. *Current Opinion in Structural Biology* **11**, 83–93 (2001).

166. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).

167. Baker, D. A surprising simplicity to protein folding. *Nature* **405**, 39–42 (2000).

168. Fleming, P. J., Gong, H., Rose, G. D. & Jenkins, T. C. Secondary structure determines protein topology. *Protein Sci.* **15**, 1829–1834 (2006).

169. Iglesias-Bexiga, M. *et al.* Protein Folding Cooperativity and Thermodynamic Barriers of the Simplest β-Sheet Fold: A Survey of WW Domains. *J. Phys. Chem. B* **122**, 11058–11071 (2018).

170. Szczepaniak, M. *et al.* Ultrafast folding kinetics of WW domains reveal how the amino acid sequence determines the speed limit to protein folding. *Proc. Natl. Acad. Sci. - PNAS* **116**, 8137–8142 (2019).

171. Sherman, E. & Haran, G. Coil-globule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11539–11543 (2006).

172. Merchant, K. A., Best, R. B., Louis, J. M., Gopich, I. V. & Eaton, W. A. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy any molecular simulations. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1528–1533 (2007).

173. Guo, L., Chowdhury, P., Glasscock, J. M. & Gai, F. Denaturant-induced Expansion

and Compaction of a Multi-domain Protein: IgG. *J. Mol. Biol.* **384**, 1029–1036 (2008).

174. Eaton, W. A., Thompson, P. A., Chan, C. K., Hagen, S. J. & Hofrichter, J. Fast events in protein folding. *Structure* **4**, 1133–1139 (1996).

175. Plaxco, K. W. & Baker, D. Limited internal friction in the rate-limiting step of a two-state protein folding reaction. *Proc. Natl. Acad. Sci.* **95**, 13591 LP – 13596 (1998).

176. Hopkins, F. G. Denaturation of proteins by urea and related substances. *Nature* **126**, 328–330 (1930).

177. Johnson, C. M. & Fersht, A. R. Protein Stability as a Function of Denaturant Concentration: The Thermal Stability of Bamase in the Presence of Urea. *Biochemistry* **34**, 6795–6804 (1995).

178. Mayr, L. M. & Schmid, F. X. Stabilization of a Protein by Guanidinium Chloride. *Biochemistry* **32**, 7994–7998 (1993).

179. Monera, O. D., Kay, C. M. & Hodges, R. S. Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions. *Protein Sci.* **3**, 1984–1991 (1994).

180. Lim, W. K., Rösgen, J. & Englander, S. W. Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2595–2600 (2009).

181. Bennion, B. J. & Daggett, V. The molecular basis for the chemical denaturation of proteins by urea. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5142–5147 (2003).

182. England, J. L. & Haran, G. Role of solvation effects in protein denaturation: From thermodynamics to single molecules and back. *Annu. Rev. Phys. Chem.* **62**, 257–277 (2011).

183. Record, M. T., Zhang, W. & Anderson, C. F. Analysis of effects of salts and uncharged solutes on protein and nucleic acid equilibria and processes: A practical guide to recognizing and interpreting polyelectrolyte effects, hofmeister effects, and osmotic effects of salts. *Adv. Protein Chem.* **51**, 281–353 (1998).

184. Dominy, B. N., Perl, D., Schmid, F. X. & Brooks, C. L. The effects of ionic strength on protein stability: The cold shock protein family. *J. Mol. Biol.* **319**, 541–554

(2002).

185. Von Hippel, P. H. & Wong, K. Y. Neutral salts: The generality of their effects on the stability of macromolecular conformations. *Science (80-.).* **145**, 577–580 (1964).

186. Halonen, S., Kangas, T., Haataja, M. & Lassi, U. Urea-Water-Solution Properties: Density, Viscosity, and Surface Tension in an Under-Saturated Solution. *Emiss. Control Sci. Technol.* **3**, 161–170 (2017).

7 Appendix

7.1 Single molecule FRET microscope: Optical set-up

The Single-molecule FRET microscope is constructed based on the inverted microscope,

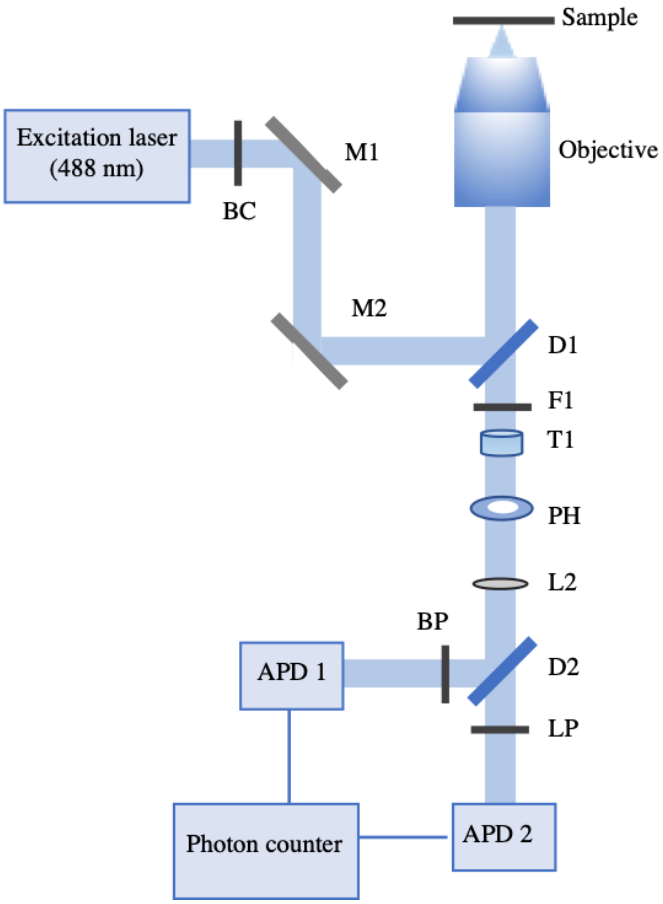


Figure 7-1: Schematic of 2-colour SM-FRET optical set-up

Nikon Ti-U. A continuous wave laser working at 488 nm is used as excitation source. The excitation laser is delivered by a single mode transmission optical fiber and passed through a collimator, BC which is directed into the microscope using mirrors M1 and M2. Light then passes through dichroic mirror, D1 which is then focused to the sample by a 100 X oil-immersion objective (N.A 1.49). Fluorescence emitted from samples is collected by the same objective and the dichroic mirror, D1 and the long pass filter, F1 to remove the scattered and reflected excitation wavelength from emission. The fluorescence is then

focused by a tube lens, T1 and passes through a 200 μm pinhole, PH, to remove the out of focus light. Emission is focused by a lens, L2 and passed through a second dichroic mirror, D2 which separates the fluorescence to acceptor and donor channels which then passes through additional bandpass filter, BP, and long pass filter, LP to reduce the leak through and scattered light. Two APDs collect the emitted photons and are registered by a multichannel photon counter.

Optical components

1. Inverted microscope: Nikon, Eclipse Ti-U.

2. Oil-immersion objective: Nikon CFI Apochromat TIRF 100XC Oil

3. Beam collimator: OZ optics, calibrated from 400 nm to 700 nm (DTS0060).

4. Polarization maintaining single mode optical fiber: Thorlabs, single mode transmission from 400 to 680 nm.

5. Excitation laser: Coherent Sapphire 488 LP, continuous wave laser system working at 488 nm.

6. Avalanche single photon detector (APD): Excelitas SPCM, with 180 μm activate area

7. Optics:

    1) Dichroic mirror: Chroma, cutting edge at 490 nm.

    2) Long pass filters: Chroma, cutting edge at 505 nm.

    3) Long pass filters: Chroma, cutting edge at 590 nm.

    4) Band pass filter: Chroma, 97% transmission from 502 nm to 545 nm.

8. Photon counter: HydraHarp 400, multichannel picosecond event timer

Alignment of Single-molecule FRET

Excitation pathway:

1) Attach the optic fiber from the excitation laser to the FC adaptor on the beam collimator. Make sure the tip of the optic fiber and the fiber port is clean. Take proper precaution to avoid exposure as the laser light exiting fiber is of high power. Roughly adjust the height and orientation of the mount to direct the light towards the microscope.

2) Turn on the laser and make sure the light is exiting through the center of the beam collimator and is not being clipped off. This can be done by using the XY adjusting knobs on the mount of the beam collimator. Use a card or a beam profiler to check the diameter of the beam all along the excitation pathway.

3) Direct the excitation laser to the high reflective mirror through the neutral density filter (Power can be adjusted by rotating the wheel mount of the filter). Ensure that the excitation light is parallel to the surface.

4) Adjust the mirror to ensure that light is entering the objective without an angle and is not being clipped off. This can be done by adjusting the mirror till light exiting the objective is seen as a perfect circle on the ceiling.

5) Ensure that excitation path is aligned with the emission path by using back reflection. This can be done by placing a mirror on the stage to reflect the light back towards the excitation path.

6) Use the alignment pinhole in front of the collimator to view the reflected beam and ensure that the back reflected light is passing through the center of the pinhole. Adjust the mirror angles if required. Ensure the path of light is straight by moving the pinhole parallel to the optical table.

7) Confirm the alignment with the camera mounted on to the eyepiece (make sure to remove excitation filter, F1 from the path of light while checking for back reflection). If the excitation path is aligned, the laser spot will expand and contract uniformly.

Emission pathway:

1) Place a sample holder with high concentration of a fluorophore on the stage (for eg. 1 μM Alexa 488) and focus the excitation light on to the surface of the coverslip (This is when the laser spot is the smallest on camera). Focus light into the sample by moving the objective 15 micron towards the sample using the fine adjustment focus knob of microscope

2) Direct the emission towards the side of microscope with excitation optics using the optical path selector knob. Measure the diameter of emitted light using the beam profiler and place the pinhole mount where it is the smallest (It can also directly be placed at the focal length of the tube lens, 20 mm for 1x magnification). Place the beam profiler in front of the pinhole and maximize the intensity of light by moving the pinhole in X and Y followed by Z direction in an iterative fashion.

3) Place the lens, L2 in the pinhole mount to focus the light onto the detectors.

4) Place the filter cube with dichroic mirror, D2, bandpass filter, BP, and long pass filter, LP centered along the path of light.

3) Use the beam profiler to measure the diameter of the light and place the APD detectors mounted on XYZ stage where the beam is the smallest

4) Insulate the detectors from background light by covering it with blackout curtains. Dilute fluorophore sample to nM concentration. Maximize the photon counts rate by adjusting the APD detector in X and Y direction followed by Z direction iteratively.

7.2 Properties of the fluorescent dyes

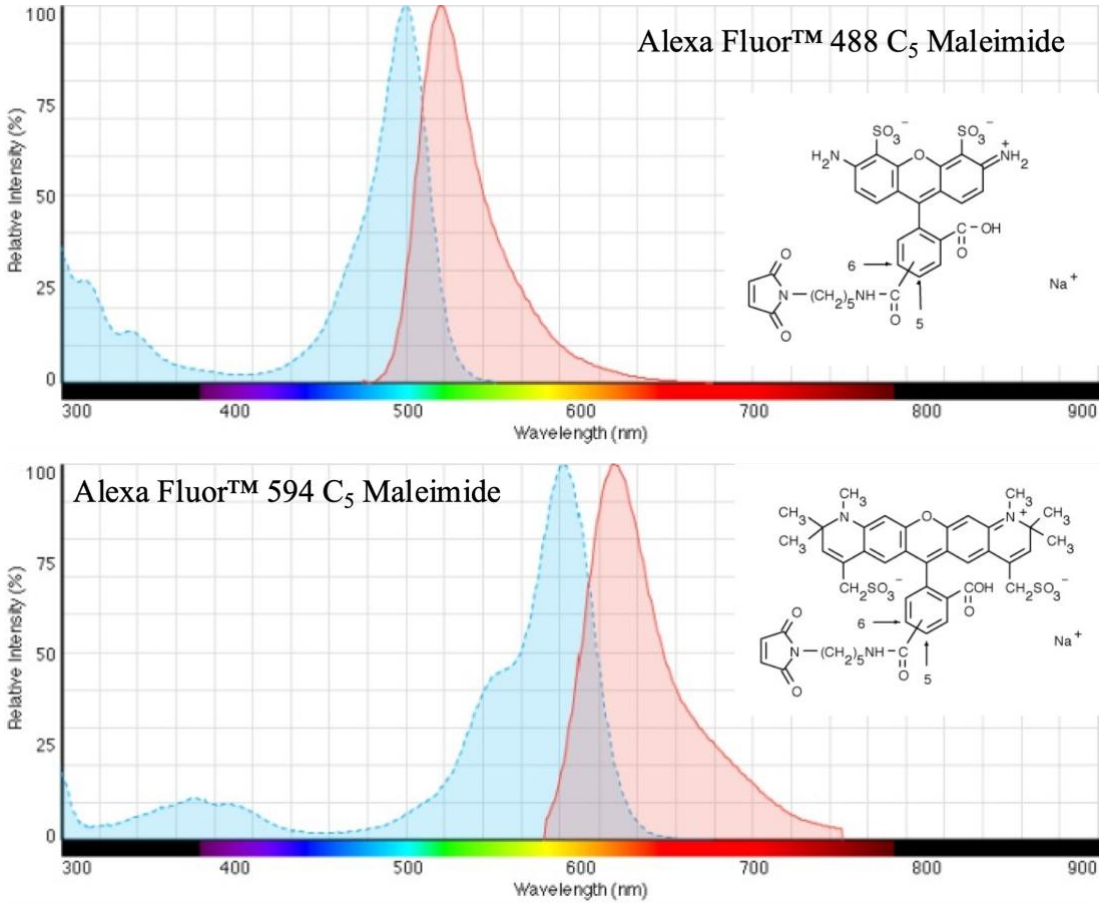


Figure 7-2 : Spectral properties and chemical structure of the organic fluorophores used as FRET acceptor-donor pair. Excitation spectra (blue) and emission spectra (red) of dyes are shown (As reported by Thermo Fischer).

7.3 Transformation of cells with recombinant plasmids

Reagent preparation

Prepare LB agar plates with ampicillin

Prepare a water bath at 42℃

Warm the SOC medium to room temperature

Keep the LB agar plates at 37℃ to remove excess moisture

Keep a shaker incubator ready at 37℃

1. Thaw one vial of One Shot® cells on ice per transformation.

2. Add 5–10 ng of DNA, in a volume of 1–5 µL to the cells and mix by tapping gently. Do not mix cells by pipetting.

3. Incubate the vial(s) on ice for 30 minutes.

4. Heat shock the cells by incubating the vial(s) for exactly 30 seconds in the 42°C water bath. Do not mix or shake.

5. Remove the vial(s) from the 42°C bath and quickly place on ice.

6. Add 250 µL of pre-warmed SOC medium to the

vial(s). (SOC is a rich medium; use proper sterile technique to avoid contamination.)

7. Secure the vial(s) in a microcentrifuge rack with tape.

Place the rack in a shaking incubator and shake the vial(s) at 37°C for 1 hour at 225 rpm.

8. Plate two different volumes of the transformation reaction onto LB plates containing the appropriate antibiotic for plasmid selection. Include 34 µg/mL chloramphenicol if using BL21(DE3)pLysS or BL21(DE3)pLysE cells. Select two volumes ranging from 20–200 µL to ensure well-spaced colonies on at least one plate.

The remaining transformation reaction may be stored at 4°C and plated out the next day, if needed.

9. Invert the plates and incubate at 37°C overnight.

Note: A single colony was collected and transferred into 15 mL LB medium and grown overnight at 37°C. Glycerol stocks were prepared from this for storing in -80°C.

7.4 Double cysteine labelling of En-HD

En-HD was labelled with Alexa Fluor™ 488 C5 Maleimide and Alexa Fluor™ 594 C5 Maleimide at the C-terminal and N-terminal following the protocol optimized in the group for double labelling of protein constructs with cysteines by making use of maleimide-thiol conjugation reaction. Both fluorophores were ordered from Thermo Fisher Scientific. Fresh batch of dyes were used for labelling reactions every time as the reactivity of dye was found to decrease days after opening the vial. An alternative was dissolving the full vial immediately after opening in MQ water/methanol and aliquoting into multiple vials which were stored in dark after removing the solvent through lyophilization.

- 100 μM of En-HD in 20 mM Sodium Phosphate buffer with 100 mM NaCl and 3 M Urea at pH 7.3 was incubated with 1 mM TCEP for one hour
- TCEP was removed by using a desalting PD10 column and the fractions collected were concentrated using Amicon centricon filter sets.
- on filter units.
- 100 μM of TCEP treated En-HD as mentioned above was then mixed with 70 μM Alexa 488 dye. The sample was incubated in total darkness for 2 h at room temperature.
- 130 μM of Alexa 594 dye was added to this solution and incubated at room temperature overnight.
- The reaction mixture was then passed through Mono S cation exchange columns to separate unreacted dyes from the labelled protein mixture
- The fractions with labelled protein were further purified by reverse phase chromatography on 0–95% water/acetonitrile gradient in the presence of 0.1% trifluoroacetic acid.
- Sample purity was confirmed by mass spectrometry.
- 10 µL aliquots of 40 µM double labelled En-HD were made in 30 % acetonitrile (reverse phase buffer) and stored in -80°C. One of these aliquots are taken and freshly thawed on ice right before experiments. It is important to use a fresh sample for single-molecule experiments as multiple freeze thaw have shown to introduce heterogeneity in En-HD samples.

7.5 SM-FRET data on En-HD at different temperatures

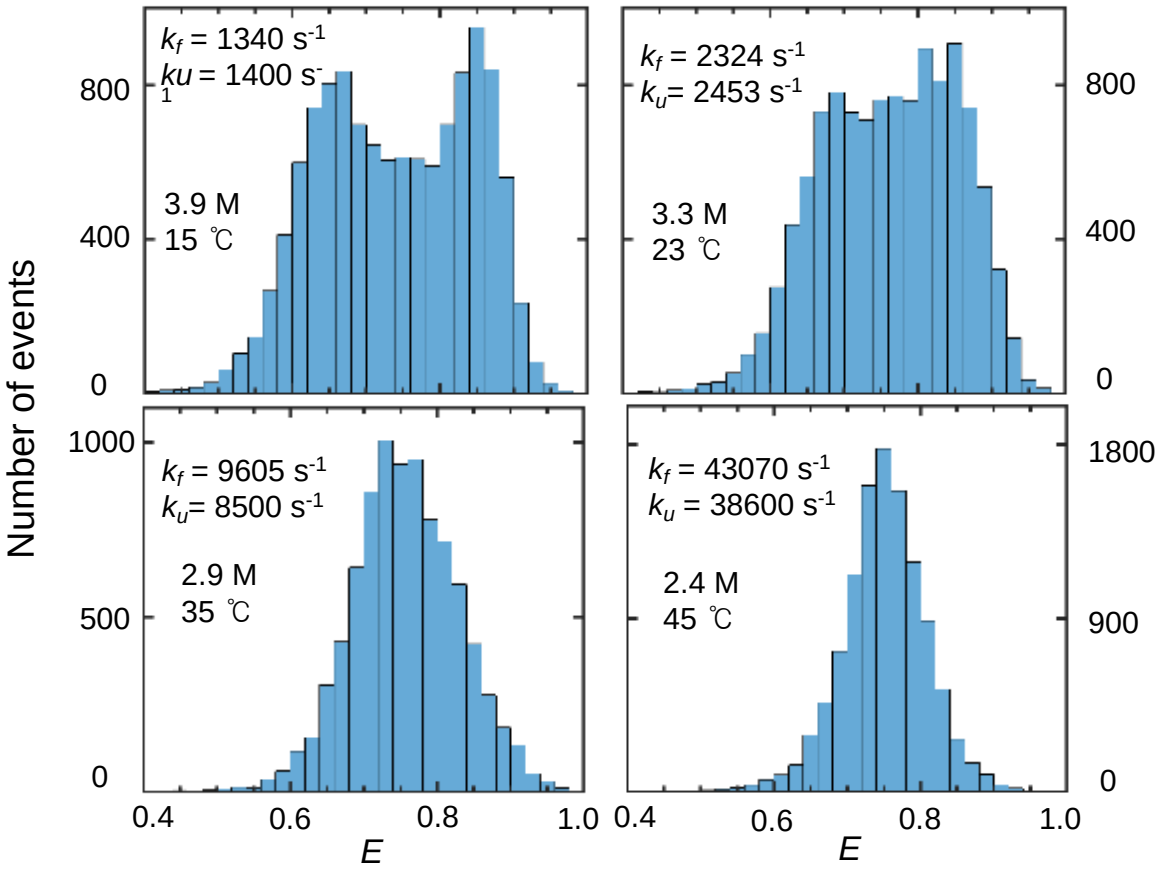


Figure 7-3 : FEH for bursts collected at urea midpoint at different temperatures for En-HD. The rates from MLA analysis with a two-state model are also shown. The dynamics at higher temperature was found to be too high to be resolved by the 1D-FES with the rate matrix formalism we employ.

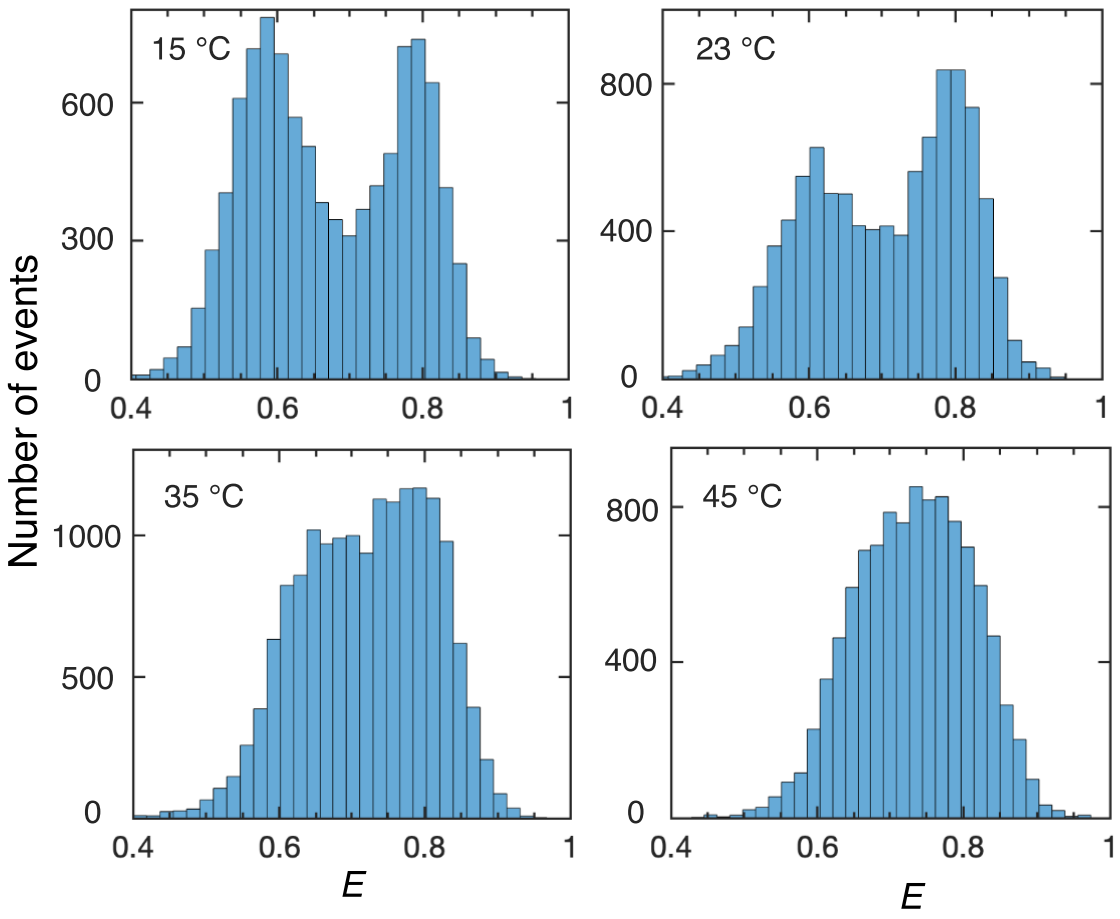


Figure 7-4 : 7-5 FEH for GdnCl midpoint condition at different temperature. Ionic strength was maintained at 2.3 M by adding NaCl for the decrease in the C_m for higher temperature.

T (°C)	GdnCl (M)	log *D*	$\lambda_1(s^{-1})$	$\lambda_2(s^{-1})$	$\lambda_3(s^{-1})$
15	2.6	2.04	1816	9752	19833
23	2.3	1.99	1321	8319	17044
35	2.1	2.11	1467	10784	22156
45	1.9	2.66	6634	39290	80322

Table 6 : Parameters from 1D-FES analysis

7.6 Bulk characterization of WW domains

Chemical denaturation experiments

All experiments were done in 20 mM MOPS buffer at pH 8. Stock solution of protein in 4.5M GdnCl was prepared. Concentration of protein stock was determined using UV-Vis absorption. GdnCl stock was made, and the concentration was determined using refractometer. Protein stock was diluted into a final concentration of 3 uM in MOPS buffer (solution a) and in GdnCl stock (solution b). Unfolding and folding curves are obtained by diluting solution a with solution b sequentially and vice versa respectively. Mixing time after each dilution was kept constant at 5 minutes. Protein was excited at 295 nm and emission was collected from 305-500 nm. Donor was excited at 460 nm and emission was collected from 460 nm - 660 nm (acceptor emission was integrated from 580-660 nm for calculation of FRET in bulk fluorescence experiments).

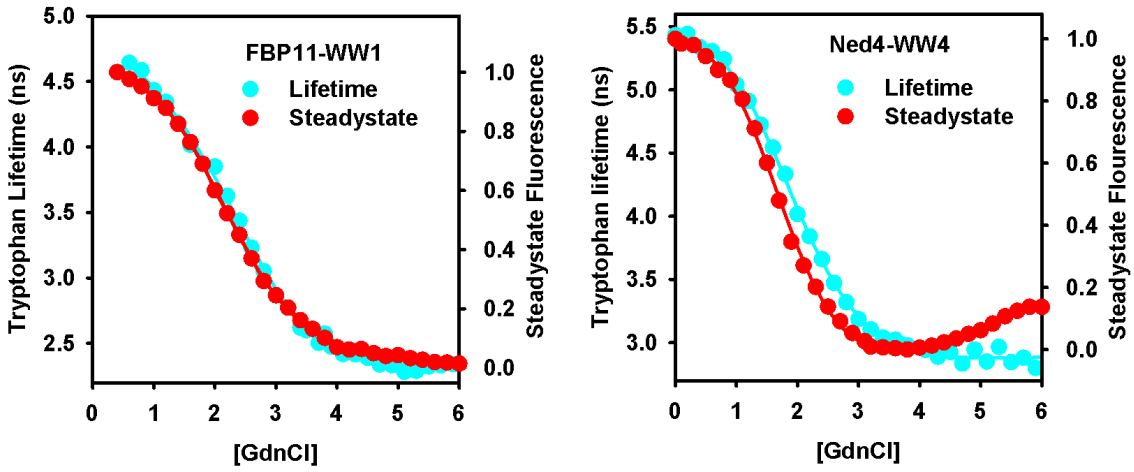


Figure 7-6 : Chemical denaturation curves for FBP11-WW1 and Ned4-WW4 monitored using steady state tryptophan fluorescence and tryptophan lifetime. Denaturation midpoint of 1.6 and 2.4 M found from global fit to two-state

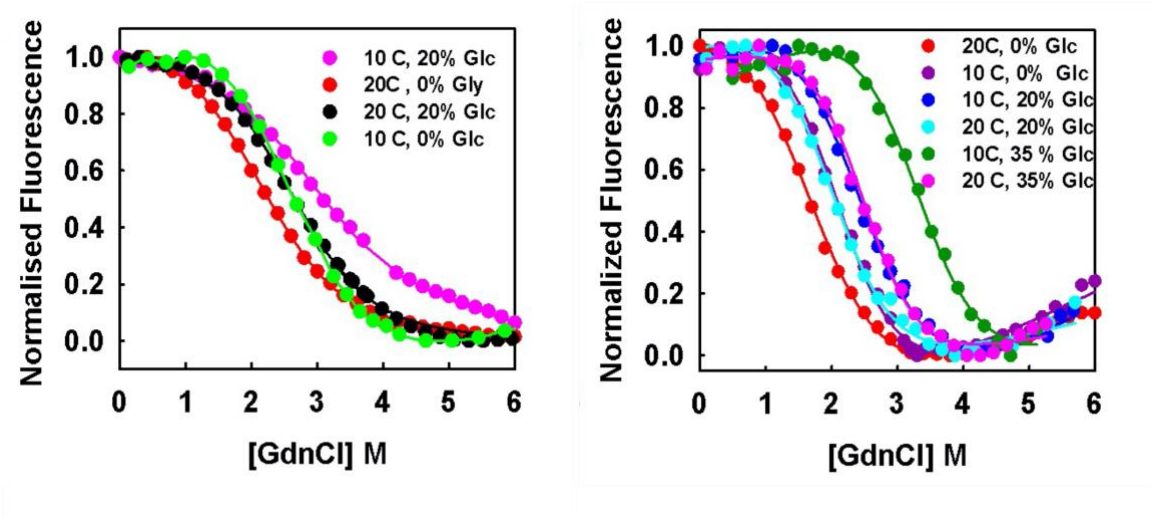


Figure 7-7: Chemical denaturation curves of FBP11-WW1and Ned44-WW4 at different conditions