CONFORMATIONAL TRANSITION OF FORMIN DISORDERED REGION UPON PROFILIN BINDING

ZHAO CHEN

(B.Sci., Lanzhou University)

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

2014
DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

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

Zhao Chen

ZHAO CHEN

24 Jan 2014
Acknowledgments

I would like to express my deepest thanks and appreciation to my two thesis supervisors, Dr. Christopher Hogue and Dr. Boon Chuan Low, for all the support and advice on my research project. I am fortunate to join Chris’ lab and learn computational structural biology from him, who constantly inspires me with his broad knowledge and new ideas. I deeply thank Chris for his supervision, advice and guidance through all these years, without which this thesis would not have been possible. I truly give my thanks to Dr. Low Boon Chuan for his supervision and suggestions during my last semester of candidature. I would like to thank my qualifying examination committee members, Yang Dai Wen, Adam Yuan, G. V. Shivashankar, who gave me precious advice and suggestions on my project. I also want to thank my department, DBS, for the financial support and administration.

I gratefully thank all my fellow lab-mates. Thanks to Chengcheng Liu and Mingxi Yao for their discussions and suggestions on formin simulation using TraDES. I would like to thank Soumya Mohanty and Xuanhao Liao for spending their time in teaching me how to perform experiments such as vector cloning, protein expression and purification, ITC and fluorescence spectroscopy. I would like to thank Yuan Xin, a graduate student in Mechanobiology Institute, for discussions and suggestions on profilin purification and the help in constructing the PLP column. I would also like to thank Zhang Bo, Arun Chandramohan, Narendra Suhas Jagannathan and Wenwei Xiang. It is a real pleasure to work with you.
Finally, I wish to express my infinite gratitude to my parents for their unconditional support, sacrifice, understanding and encouragement through all these years.
# Table of Contents

Summary .......................................................................................................................... vii

List of Tables .................................................................................................................... ix

List of Figures ................................................................................................................... x

List of Symbols and Abbreviation ....................................................................................... xii

Chapter 1 Introduction ...................................................................................................... 1

1.1 Intrinsically Disordered Proteins (IDPs) ................................................................. 3
   1.1.1 New revelation of protein structure-function relation ...................................... 3
   1.1.2 Structural features of IDPs .............................................................................. 5
   1.1.3 Diverse roles of IDPs .................................................................................... 6
   1.1.4 Sequence preference of IDPs .......................................................................... 17
   1.1.5 Experimental characterization of IDPs .......................................................... 20

1.2 Molecular Simulations ............................................................................................... 22

1.3 Formin ......................................................................................................................... 26
   1.3.1 Formin domain organization ....................................................................... 26
   1.3.2 Disordered FH1 domain ............................................................................... 29
   1.3.3 Interaction of profilin and formin .................................................................. 34
   1.3.4 Formins in actin filament assembly ............................................................... 35
   1.3.5 Molecular motors ......................................................................................... 36
   1.3.6 Formins as motors ...................................................................................... 39

1.4 Aims and Objectives ................................................................................................. 41

Chapter 2 Monte Carlo Simulation of the Disordered Formin FH1 Domain.............. 44

2.1 Background .............................................................................................................. 44

2.2 Results ...................................................................................................................... 48
2.2.1 Disordered FH1 domain ................................................................. 48
2.2.2 Polyproline tracks in the FH1 domain and the pattern for profilin binding
sites .......................................................................................................... 50
2.2.3 Classification of free and bound profilin binding sites ...................... 53
2.2.4 Formin FH1 domain opens up and elongates upon profilin binding ...... 56
2.2.5 End-to-end distance ........................................................................ 58
2.2.6 The cooperativity of the neighbor sites closest to the bound site ........ 62

2.3 Discussion ............................................................................................ 65
2.3.1 Radius of gyration .......................................................................... 65
2.3.2 End-to-end distance ......................................................................... 68
2.3.3 Addition of profilin bound site reduces the contribution of individual binding
site in end-to-end distance expansion ....................................................... 70
2.3.4 Relationship between conformational elongation and formin motor function
.................................................................................................................. 71

2.4 Methods ............................................................................................... 76
2.4.1 Generation of structure ensembles of mDia1-FH1 domain ............... 76
2.4.2 Alignment of [PDB:2V8F] to mDia1-FH1 conformers ....................... 78
2.4.3 Merging and checking clashes .......................................................... 78
2.4.4 Measurement of $R_{gyr}$ and end-to-end distances ............................ 79

Chapter 3 Thermodynamic Study of the FH1 Domain and Profilin Binding 81

3.1 Background .......................................................................................... 81
3.2 Calorimetry Theory ............................................................................. 83
3.3 Thermodynamics of Binding ............................................................... 85
3.4 Materials and Methods ....................................................................... 88
3.4.1 Peptide and protein .................................................................... 88
3.4.2 Instrumentation ................................................................. 88
3.4.3 Isothermal titration calorimetry ........................................ 90
3.5 Results and Discussion .......................................................... 91

Chapter 4 Terbium Chelate as a Luminescent Probe for Studying Formin FH1 Conformational Change ......................................................... 95
4.1 Background ........................................................................... 95
4.2 Results and Discussion ........................................................... 97
  4.2.1 Overlap of donor and acceptor spectra .............................. 97
  4.2.2 LRET study ................................................................... 99
  4.2.3 Measurement of luminescence decay ............................... 102
4.3 Conclusion ........................................................................... 105
4.4 Materials and Methods ......................................................... 106
  4.4.1 Peptide and profilin .......................................................... 106
  4.4.2 Terbium titration ............................................................. 106
  4.4.3 LRET ............................................................................ 106
  4.4.4 Measurements of decay ................................................. 107

Chapter 5 Conclusions and Future Research ................................. 108
Bibliography ............................................................................. 113
Summary

Formin is found to be an active motor for actin filament assembly during the filopodia elongation process. The FH1 domain of formin is predicted to be disordered and plays an essential role in motor function while models trying to explain formin protein behavior are all centered on FH2 domain till now. To provide insight into the mechanism by which the disordered FH1 domain switches formin from FH2 mediated capping capacity to FH1-FH2 coordinated accelerator role, conformations of disordered formin protein FH1 domain were simulated using a trajectory directed ensemble sampling software based on the Monte Carlo method. Statistical analysis of four million FH1 conformations showed the disordered FH1 domain elongates after binding to profilin. Since the binding motif on FH1 domain is composed of five consecutive prolines, it naturally forms polyproline II helix so that the conformation of the motif is relatively rigid comparing to the linker between motifs. The conformational transition of FH1 domain can be attributed to the linker residues in this case. FH1 domain of formin gradually opens up as more profilin binds to it and the distribution of the radius of gyration for FH1 structures shifts from small compact conformations towards extended ones with every newly added profilin at a steady pace of one nanometer. This finding showed a similarity between the additive effects of profilin bound polyproline motif on conformational elongation and motor acceleration rate. Both displayed a positive correlation with the number of profilin bound motifs inside the FH1 domain. Henceforth, a mechanical jack model for formin motor is a possible explanation for this profilin bound motif number based acceleration performance. From the calculation of single site length, each
binding site, counting from the linker in the middle, is elongated upon profilin binding. This finding implies each single binding site has the capability to stretch and will elongate when binding occurs. To determine the thermal dynamic profile of profilin-FH1 binding, isothermal titration calorimetry was performed. It turned out to be an exothermal reaction and released 6.6 thousand calories of Gibbs free energy per mole. This finding unveiled the possibility that the binding energy derived from profilin-FH1 binding could be exploited for formin motor function. Besides conformational change a mechanical jack also requires energy source, while the argument on the role of ATP in formin motor function is still going on, results from this study now shows the potential of FH1 domain to provide energy from binding event itself. To further support the conformational elongation effect of polyproline motif on FH1 domain upon profilin binding, luminescent decay experiments were conducted. The result is in accordance with simulation finding on neighbor cooperativity that a bound profilin induces the neighbor residues to adopt an elongated conformation as well.
List of Tables

Table 1.1 Diverse roles of IDPs/IDRs................................................................. 12
Table 1.2 Formins with FH1 domain identifier in UniProt................................. 30
Table 2.1 $R_{gyr}$ and end-to-end distance simulation result for mDia1 FH1 domain......................................................................................................................... 59
Table 2.2 Mann-Whitney-Wilcoxon Test (unpaired) for $R_{gyr}$......................... 60
Table 2.3 Mann-Whitney-Wilcoxon Test (unpaired) for single site distance distributions of conformers bound with only one profilin. ....................... 61
List of Figures

Figure 1.1 Energy landscapes of folded proteins (A) and IDPs (B) .................. 24
Figure 1.2 Schematic illustration of formin domain organization. ............... 28
Figure 1.3 Schematic illustration of the direct transfer polymerization model.. ........................................................................................................... 38

Figure 2.1 Prediction results of the mDia1 FH1 domain [UniProt O08808] using disorder prediction software. ........................................... 49
Figure 2.2 Distribution of the length of a polyproline track. ....................... 51
Figure 2.3 Amino acid composition in the upstream and downstream positions surrounding the polyproline tracks............................................ 52
Figure 2.4 Determination of the root mean square deviation (RMSD) threshold for a good alignment. ........................................................................ 55
Figure 2.5 Determination of the clash tolerance for sound profilin docking. 55
Figure 2.6 Radius of gyration (R_{gyr}) A) and End-to-end distance (NCdist) B) distribution of conformers with 0 ~ 6 profilin binding sites. ............. 57
Figure 2.7 Comparison of the distance distributions of single sites. Site B and site G are the sites on the two terminals................................................ 64
Figure 2.8 Comparison of mDia1-FH1 conformers with (A) and without (B) profilin binding capability................................................................. 67
Figure 2.9 Illustration of excluded volume effect on the FH1 domain......... 69
Figure 2.10 Scatter plot of R_{gyr} vs end-to-end distance of FH1 structural ensembles ....................................................................................... 72
Figure 2.11 The proposed cooperative "jack model" for formin FH1 in the actin filament elongation................................................................. 75
Figure 2.12 Illustration of the workflow..................................................... 77
Figure 3.1 Schematic diagram of the Gibbs energy change during the interaction of ligand and protein. ............................................................... 87

Figure 3.2 Schematic diagram of an ITC instrument. A reference cell and a sample cell are held in an adiabatic shield. .................................................. 90

Figure 3.3 Titration of mouse 2a profilin bound with mDia1-FH1.. ............ 92

Figure 4.1 Stick representation of the X-ray structure of Tb(III) bound LBT. The metal-ligating residues are shown in yellow......................................... 98

Figure 4.2 Spectra of mDia1 FH1-labeled with either terbium or AF555...... 99

Figure 4.3 Emission spectra of donor and acceptor labeled peptides. ........... 101

Figure 4.4 Acceptor lifetime at 565 nm of acceptor-only (o), and donor-acceptor (♦) groups................................................................. 102

Figure 4.5 Donor lifetime data at 545 nm of donor-only labeled mDia1-FH1 (♦), and donor with profilin binding to FH1 (o).................................. 103
**List of Symbols and Abbreviation**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>C terminal</td>
<td>Carboxyl terminal</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EGTA</td>
<td>Ethylene glycol tetraacetic acid</td>
</tr>
<tr>
<td>E. coli</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>FH1</td>
<td>Formin homology domain 1</td>
</tr>
<tr>
<td>FH2</td>
<td>Formin homology domain 2</td>
</tr>
<tr>
<td>FRET</td>
<td>Förster resonance energy transfer</td>
</tr>
<tr>
<td>g</td>
<td>Centrifugal force</td>
</tr>
<tr>
<td>HEPES</td>
<td>2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid</td>
</tr>
<tr>
<td>I</td>
<td>Luminescence intensity</td>
</tr>
<tr>
<td>IDPs</td>
<td>Intrinsically disordered proteins</td>
</tr>
<tr>
<td>IDR</td>
<td>Intrinsically disordered regions</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>K</td>
<td>Kelvin</td>
</tr>
<tr>
<td>$K_d$</td>
<td>Dissociation constant</td>
</tr>
<tr>
<td>LBT</td>
<td>Lanthanide binding tags</td>
</tr>
<tr>
<td>LRET</td>
<td>Luminescence resonance energy transfer</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular dynamics</td>
</tr>
<tr>
<td>N-terminal</td>
<td>Amino terminal</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PLP</td>
<td>Poly-L-proline</td>
</tr>
<tr>
<td>O.D.</td>
<td>Optical density</td>
</tr>
<tr>
<td>$R_g$</td>
<td>Radius of Gyration</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Square Deviation</td>
</tr>
<tr>
<td>$R$</td>
<td>The universal gas constant</td>
</tr>
<tr>
<td>t</td>
<td>Lifetime</td>
</tr>
<tr>
<td>Tb(III)</td>
<td>Trivalent terbium</td>
</tr>
<tr>
<td>w/v</td>
<td>Weight per volume</td>
</tr>
<tr>
<td>ºC</td>
<td>Degree Celcius</td>
</tr>
<tr>
<td>$\Delta G$</td>
<td>Change in Gibbs energy</td>
</tr>
<tr>
<td>$\Delta H$</td>
<td>Change in enthalpy</td>
</tr>
<tr>
<td>$\Delta S$</td>
<td>Change in entropy</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Proteins are generally considered to have folded globular structures and are the result of folding from the primary amino acid sequence to their tertiary structure. As the interest on protein structures grows, it is becoming clear that native protein structures can be grouped into three categories: the folded or the ordered state, random coil and molten globule state [1]. The discovery of intrinsic disordered proteins has challenged our current understanding of protein structure and folding. Contrary to the canonical view of protein structure and function, disordered state is devoid of a fixed folded structure but still capable of carrying out normal biological function by forming folded structures upon binding. Because of the random coil shape, disordered proteins have more flexibility than folded proteins and usually function as binding hubs, facilitating a lot of protein interactions inside the cell. However, crystallography fails in characterizing intrinsically disordered protein structure due to the absence of single folded structure. NMR is a popular method investigating disordered proteins nowadays with many conformational changes of disordered proteins revealed [6-9] but the large demand of highly purified protein makes it challenging since a great many IDPs are insoluble. Structural characterization of disordered proteins is in urgent need as considerable amount of protein sequences from various genome projects is predicted to be disordered and functional. Due to their functional importance, research on disordered proteins cannot be side stepped.

One such protein that is an integral part of the cell is Formin. Formin widely exists in different organisms and functions in nucleating and
polymerizing unbranched actin filament. Due to its active role in actin dynamics, formin is essential for variety of important cellular activities such as cytokinesis, filopodium formation, morphogenesis and endocytosis. Formins contain two characteristic conserved domains: the proline-rich formin homology domain 1 (FH1) that recruits profilin-actin for progressive filament elongation and formin homology 2 (FH2) that nucleates actin filaments by forming a ‘donut-shaped’ dimer which progressively stays at the barbed ends of filaments.

Experiments verify that formin accelerates actin filament elongation in the presence of profilin. Although many models have been suggested to explain the functional mechanism of formin, none of them explain how this acceleration happens and the role of FH1 in it. Because formin FH2 domain possess a well-defined 3D structure, it was the main focus of formin related research. From stepwise progression movement along the barbed end of actin filament to the internal rotational movement between FH2 domain and the actin filament it attaches to, detailed mathematical models all focus on FH2 domain[5-7]. However, Pollard et al have found the number of profilin binding sites on FH1 is directly linked with the acceleration effect of the formin motor. Due to the importance of FH1 domain, any model trying to explain the formin working mechanism without considering this essential FH1 domain is not complete.

Since FH1 domain is intrinsically disordered, it is not cost effective to investigate the ensemble of its conformations using an experimental approach. Molecular simulations of the disordered FH1 domain can generate large ensembles of structures in reasonable time. It evades the traditional structure-
solving problem and at the same time incorporates experimental constraints to make sure the sampling is consistent with experimental results. TraDES is a molecular simulation tool that has been used earlier for understanding intrinsic disordered proteins and their functions [9, 10]. The simulation of formin FH1 domain would help to elucidate the conformational changes in formin when it exercises the motor function in actin filament assembly. It would also shed new light on the mechanism of other motor proteins with intrinsically disordered regions. This chapter provides an overview of intrinsically disordered proteins, the molecular simulation methods that have been employed to study the intrinsically disordered proteins. Then we discuss the importance and our current understanding of formins.

1.1 Intrinsically Disordered Proteins (IDPs)

1.1.1 New revelation of protein structure-function relation

Ever since Emil Fischer proposed the lock and key model for enzyme and substrate recognition in 1894, the notion that protein functions rely on their unique folded structures was one of the most accepted hypothesis about protein structure and function relationships. Denatured proteins lose their biological activity and are not capable of crystallization while native proteins form characteristic crystals [12]. From above observation, Pauling concluded that the uniquely defined configurations of native proteins are responsible for their distinctive properties [12, 13]. Increasing number of protein denaturation studies later supported this view [14, 15] and subsequently tens of thousands of atomic resolution protein structures were solved. Thus, the theory of protein function depending on its specific 3D structure was well accepted. Serum
albumin, a protein that binds multiple small anionic molecules with different shapes, changed the concept of one protein corresponding to one structural configuration [16]. This finding broadens the fixed structure-function view by adding multiple structures of one protein into consideration [16, 17]. Inspired by the multiple conformations observed for serum albumin, Daniel Koshland suggested the induced fit model [18] which was experimentally verified [19-21]. In contrast to the lock and key model that treats binding as resulting from complementary geometrically rigid structures, the induced fit model proposes that conformational changes of the active sites of proteins could be induced for better binding. Conformational changes of protein structures are proposed to be important for protein function [17, 18]. It indicates the possibility of one conformer shifting to another under the effect of its binding partner. Originally, different conformations are in equilibrium in solution and the binding event selects a certain appropriate population with the required conformation. This selection pushes the equilibrium of all conformers to a new balance, and favors the one that facilitates binding. Association of glucose with hexokinase is a good example of conformational changes that occur upon binding [19]. In contrast to the above canonical views that protein function depends on its well defined structure or associated conformational changes, numerous disordered proteins have been reported to be functional from X-ray crystallography and NMR experiments [22-26]. Without well-defined secondary and tertiary structures while actively functioning in vivo on a wide range, intrinsically disordered proteins (IDPs) challenge the structure-function paradigm, which states that ordered 3D structure is essential for protein function.
This emerging trend in the study of disordered proteins have given rise to terms such as natively denatured, unfolded, intrinsically unstructured [1, 27]. According to Uversky, denatured and disordered means “any set of non-rigid conformations of polypeptide chains including different compact partially folded conformations” while unstructured and unfolded are interchangeable terms referring to the subset of disordered proteins which lack any ordered structure. Disorder is a continuum level of various degrees of compactness and amount of folded structure [28, 29]. Among the four conformations of proteins (ordered, molten globules, pre-molten globules and random coils), disordered state covers three of them: from collapsed molten globules to extended form, as pre-molten globule or random coils [27]. Due to the various conformations that IDPs adopt, no consensus has been reached for the definition of disordered states [30]. For clarity, the IDPs discussed here refer to proteins without well-defined ordered structure under physiological conditions. Unstructured segments either in isolation or within full-length proteins will be referred to as intrinsically disordered regions (IDRs) [31].

1.1.2 Structural features of IDPs

Although there are no precise measurements to pinpoint IDPs now, a few structural features can help in identifying them. Compactness is the most unambiguous feature to differentiate IDPs from globular proteins [27]. The hydrodynamic radius is a good indicator of compactness of proteins. The hydrodynamic radius, which is derived from Stokes-Einstein viscosity relation, is defined as the radius of solid spherical particles that exhibits the same frictional coefficient as the protein. The hydrodynamic radius is larger for unfolded proteins compared to globular proteins of the same molecular
weight [32, 33]. Size exclusion chromatography (SEC) and dynamic light scattering are commonly used methods to determine protein hydrodynamic radius [33, 34]. The degree of globularization is another trait that differs in ordered and disordered proteins. Ordered proteins possess a tightly packed core, which IDPs lack. IDPs are enriched with charged residues and lack hydrophobic residues. It is hard to form a compact hydrophobic core with the small amount of hydrophobic residues found in IDPs, compared to folded proteins. A maximum can be found from a scattering curve in Kratky plot via SAXS analysis for globular proteins but not for IDPs [35, 36]. Several IDPs are found with low content of secondary structure from far-UV CD analysis [37-40] and serve as one more feature of IDPs. NMR studies also report increased intra-molecular flexibility for disordered proteins [1, 41, 42]. Although IDPs do not follow the canonical structure-function paradigm, these loosely packed proteins have their unique structural character and therefore are responsible for certain roles in cells that correspond to their special structural traits.

1.1.3 Diverse roles of IDPs

In general, more than 30% of eukaryotic proteins contain disordered segments with at least 30 consecutive amino acids across genomes [43]. For cancer related proteins, almost 80% of sequences are predicted to have disordered regions [44]. This is mainly due to the connection between cell-signaling and cancer related proteins [45], a great many cell-signaling proteins are found to play a part in cancer development and are structurally disordered [46-48]. Disorder gives a protein the conformational freedom to bind multiple partners [49] which might enable a broad range of cell-signaling pathways.
Disordered proteins are key to signaling networks, which in turn contribute to the regulation of cancer cells. Disordered proteins are found to be more common in eukaryotes than prokaryotes in a sequence analysis study [50]. This may have been a result of the emergence of more complex signaling and regulatory networks in eukaryotes [44]. From the analysis of sequence functional annotation, it was found that sequences playing a part in regulation, transcription and development are likely to have disordered regions, while sequences involved in transportation and catalytic activity are inclined to have ordered structures [51]. A reasonable explanation for this phenomenon is the fact that disordered proteins work in a way complementary to ordered proteins. Some examples for the diverse roles of IDPs are listed in Table 1.1.

Molecular recognition is the most common role listed here for disordered proteins. Calsequestrin is a calcium storage protein with the capacity of binding up to 40–50 calcium ions. Such binding is with high specificity but low affinity, which enable quick uptake and release of calcium from sarcoplasmic reticulum for muscle relaxation and stimulation. From the crystal structure of rabbit skeletal muscle calsequestrin, the electron densities of a seven-residue disordered loop and a twenty-residue carboxy terminus are missing. These regions contain multiple negatively charged residues and are supposed to bind calcium [52]. Lac repressor is a protein that specifically recognizes certain DNA sequences and inhibits transcription of the corresponding operon [53]. Both X-ray diffraction and NMR experiments show the difference between the disordered free DNA binding domain and its ordered structure upon DNA binding [54, 55]. Cyclin-dependent kinase inhibitor p21Waf1/Cip1/Sdi1 is completely disordered as indicated by various
experiments [56]. A disorder to order transition is also observed for this protein upon binding to CDK2 [56]. The caspase-activated DNA fragmentation factor DFF45 has an intrinsically disordered N terminal domain. Binding of DFF40 induces it to form a folded structure [57]. NMR experiments show that the DNA binding domain of human vitamin D receptor undergoes disorder to order transition upon metal and DNA binding [58]. 4E-binding protein is an intrinsically disordered protein that executes its inhibitory function upon binding eIF4E [59]. A 50 amino acid long peptide of DNA polymerase I was identified to be disordered from NMR and CD experiment. It is reported to bind dNTP and duplex DNA [60]. The steroidogenic acute regulatory protein (StAR) enables cholesterol to move from the outer mitochondrial membrane to inside. Its N-terminal protease-resistant domain has a molten-globule structure and this facilitates the interaction with membranes [61]. As indicated from missing electron density of its crystal structure, residue 1 to 15 of Trypsinogen is recognized as disordered. The proteolytic digestion of inactive trypsinogen results in an active trypsin. The cleavage site is within this disordered region and the cleavage induces it to go through a disorder to order transition [62]. The C-terminal domain of Hirudin is completely disordered in its free state. To inhibit blood coagulation, the C-terminal binds to thrombin and inhibits the conversion of fibrinogen to fibrin [63, 64].

Disordered protein could facilitate molecular assembly with their advantage of being versatile in conformations. The disordered region of tobacco mosaic virus (TMV) capsid is a positively charged 25-residue loop implied by missing electron density and NMR experiments [65, 66]. It exhibits
a disorder to order transition upon RNA binding. This is shown to be essential for the coat protein assembly [67]. The GP120 protein from HIV-1 has a 24-residue V3 loop that impacts its infectivity. The disorder to order transition of this region facilitates diverse entry mechanisms into the cell [68-70] FlgM is a whole length disordered protein in its native state. It regulates flagellum assembly by its transportation into the basal body assembly channel. Once FlgM starts to accumulate inside the channel, the synthesis of the mRNA of the flagellum subunit will terminate. The small diameter of this channel requires the disordered nature of FlgM for transportation purposes [67, 71].

Neuromodulin is a protein that stores calmodulin for maintaining the calcium balance at the growing tips of nerve cells. The disordered characteristic of this protein permits high quantity, low affinity binding with calmodulin [72-74].

Disordered structure possesses excess flexibility that permits conformational regulation for different needs. The fd pIII protein contains three folded domains with two flexible linkers. Disordered linkers are needed for the movement between the domains so that they can adopt various orientations and spacing, which is a common function for disordered linkers [75]. The linker inside kinesin enables a stepping motion as a result of its disorder to order transition. Ordered and disordered crystal structures are both observed for this linker region [76, 77]. Titin is a large protein that stretches across a broad space in muscle cells. It is proposed to function as an entropic spring that assists overstretched muscle cells to recover to their natural length [78]. A flexible chain regulates the K+ channels in nerve axon. It possesses the ability to block the channel. The flexibility of the chain contributes as an entropic clock that controls the open/closed status of the channel [79-81].
Neurofilament H is a motor neuron axon with a 679 amino acid long disordered end that keeps the filaments well spaced by its thermally driven motion. This is an example of an entropic bristle that enables small molecules’ movement while retaining the shape of the axon [82]. Bone sialoprotein (BSP) and osteopontin (OPN) both belong to the SIBLING family and are shown to be disordered by NMR [83]. The flexibility of these proteins makes binding multiple partners possible and often function in bridging two folded proteins into a complex [83]. Dehydrin-related Dsp 16 protein is a desiccation stress protein that accumulates when the resurrection plant Craterostigma plantagineum is under dehydration stress. No well-defined structure was observed for this protein in solution. NMR experiments showed a small chemical shift dispersion pattern, which is characteristic of disordered proteins. The extended structure makes the Dsp 16 protein highly sensitive to proteolysis while its high content of serine and threonine makes an easy target for phosphorylation. It is a good example of disordered protein that binds to water and can be phosphorylated at the same time [84].

The reason for the prevalence of disordered proteins could be inferred from their special structural characteristics. Disordered proteins are smaller compared to a folded protein with equivalent size of intermolecular interfaces, thus, lower the energy cost of protein-protein interactions [85]. They facilitate multiple interactions with sufficient interfaces while keeping protein and cell to a minimal size [86]. Being disordered is essential for the diverse roles that IDPs play. The flexible linkers that connect ordered domains in multidomain proteins facilitate a high degree of freedom for the orientations they can adopt. The disorder to order transition upon binding decreases the conformational
entropy in the system. It also results in specific reversible interactions [71, 87-89] and enables binding with high specificity and low affinity. The open structure of IDPs enables them to bind multiple partners in various binding events [90-92]. Their disordered nature permits IDPs to adopt different conformations upon binding distinct partners [48, 83, 93]. This contributes to the versatile roles of IDPs and they are recognized as interacting hubs in signaling [1, 41]. Conversely, a binding partner may also bind to many IDPs resulting in many-to-one signaling [30]. The half-lives of IDPs are generally shorter than those of folded proteins and are extended by modifications or by binding. This promotes the sensitivity of cells to different environmental conditions [94]. In summary, functions of IDPs mainly fall into four categories: molecular assembly, protein modification, molecular recognition and entropic chain activities [95, 96].
<table>
<thead>
<tr>
<th>Protein</th>
<th>Disorder</th>
<th>Function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calsequestrin</td>
<td>21 amino acid disordered tail</td>
<td>Ca(^{2+}) binding</td>
<td>[97-99]</td>
</tr>
<tr>
<td>Trypsinogen</td>
<td>Residue 1-15</td>
<td>Folding inhibitor</td>
<td>[62, 100]</td>
</tr>
<tr>
<td>TMV capsid</td>
<td>25 positively charged amino acid loop</td>
<td>RNA binding</td>
<td>[65-67, 101]</td>
</tr>
<tr>
<td>Lac repressor</td>
<td>61 amino acid tail</td>
<td>DNA binding</td>
<td>[53-55]</td>
</tr>
<tr>
<td>p21(^{^\text{Warf/Cip1/Sdi1}})</td>
<td>Whole length (164 residues)</td>
<td>Cyclin-dependent kinases inhibitor</td>
<td></td>
</tr>
<tr>
<td>V3 loop of HIV-1 gp120</td>
<td>24 amino acid long segment</td>
<td>Cell surface attachment</td>
<td>[66-68]</td>
</tr>
<tr>
<td>FlgM</td>
<td>Whole length (97 aa)</td>
<td>Transcription promoter that regulates flagella assembly</td>
<td></td>
</tr>
<tr>
<td>4E-binding protein</td>
<td>Whole length (118 aa)</td>
<td>Translation inhibitor</td>
<td>[57, 100]</td>
</tr>
<tr>
<td>Neuromodulin</td>
<td>Whole length (239 aa)</td>
<td>Calmodulin storage protein</td>
<td>[70-72]</td>
</tr>
<tr>
<td>fd pIII</td>
<td>A 21-residue and a 40-residue linker</td>
<td>Flexible linker</td>
<td></td>
</tr>
<tr>
<td>Kinesin</td>
<td>15-residue</td>
<td>Stepping motor</td>
<td>[74, 75, 102]</td>
</tr>
<tr>
<td>Protein</td>
<td>Description</td>
<td>Function/Property</td>
<td>References</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Titin</td>
<td>2174 amino acid</td>
<td>Entropic spring</td>
<td></td>
</tr>
<tr>
<td>K⁺ Channel</td>
<td>64 residue</td>
<td>Entropic clock determine the timing of open/closure status of the channel</td>
<td>[77-79]</td>
</tr>
<tr>
<td>Neurofilament H</td>
<td>679 amino acid disordered end</td>
<td>Entropic bristles separating neighboring filaments</td>
<td></td>
</tr>
<tr>
<td>Bone sialoprotein (BSP)</td>
<td>Whole length (327aa)</td>
<td>Hydroxyapatite binding</td>
<td>[81]</td>
</tr>
<tr>
<td>DNA fragmentation factor 45</td>
<td>N-terminal domain residue 1-116</td>
<td>Nuclease inhibitor</td>
<td></td>
</tr>
<tr>
<td>Vitamin D receptor</td>
<td>Residue 21-96</td>
<td>DNA binding</td>
<td></td>
</tr>
<tr>
<td>P19⁰⁰⁰ tumor suppressor</td>
<td>37 N-terminal amino acid</td>
<td>Inhibit p53 degradation</td>
<td></td>
</tr>
<tr>
<td>DNA polymerase I</td>
<td>50 residue</td>
<td>dNTP and DNA binding</td>
<td></td>
</tr>
<tr>
<td>Steroidogenic acute regulatory protein (StAR)</td>
<td>N-terminal domain residue 63-193</td>
<td>Membrane binding</td>
<td></td>
</tr>
<tr>
<td>Osteopontin (OPN)</td>
<td>Whole length</td>
<td>Factor H binding</td>
<td>[81]</td>
</tr>
<tr>
<td>Prion protein</td>
<td>N-terminal residue 23-126</td>
<td>Cu²⁺ binding, Convert PrPC to PrPSc</td>
<td>[105-108]</td>
</tr>
<tr>
<td>Protein</td>
<td>Length/Region</td>
<td>Function</td>
<td>References</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Dehydrin-like dessication stress protein</td>
<td>Whole length</td>
<td>Water binding and phosphorylation</td>
<td></td>
</tr>
<tr>
<td>Prothymosin α</td>
<td>109 aa</td>
<td>Interact with histone H1</td>
<td>[32, 109]</td>
</tr>
<tr>
<td>Fibronectin-binding domains</td>
<td>130 aa</td>
<td>Insert to cell wall</td>
<td></td>
</tr>
<tr>
<td>α-Fetoprotein</td>
<td>Whole length 590 aa</td>
<td>Regulate cell division and metabolism</td>
<td></td>
</tr>
<tr>
<td>Breast cancer type 1 susceptibility protein (BRCA1)</td>
<td>1480 aa long central region</td>
<td>Mediate interactions with DNA and p53</td>
<td>[112-114]</td>
</tr>
<tr>
<td>α-Synuclein</td>
<td>Whole length 140 aa</td>
<td>Link synucleinopathies, high propensity to aggregate</td>
<td>[115-118]</td>
</tr>
<tr>
<td>Amyloid β protein</td>
<td>40–42 residue</td>
<td>Transcription factor, kinase activator</td>
<td>[119, 120]</td>
</tr>
<tr>
<td>Hirudin</td>
<td>Amino acid 50–65 C terminal region</td>
<td>Recognize α-helix, thrombin binding</td>
<td>[121-124]</td>
</tr>
<tr>
<td>Lymphoid enhancer-enhancer-</td>
<td>Whole length (399 aa)</td>
<td>DNA binding</td>
<td>[125]</td>
</tr>
<tr>
<td>Protein Name</td>
<td>Domain Details</td>
<td>Function</td>
<td>Reference</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>-----------------------------------------</td>
<td>------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>binding factor (LEF-1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human T-cell leukemia virus type I tax peptide</td>
<td>Residue 89-110</td>
<td>DNA binding</td>
<td>[126]</td>
</tr>
<tr>
<td>Diphtheria toxin repressor protein (DtxR)</td>
<td>C terminal residues 149-226</td>
<td>Divalent metal cations binding</td>
<td>[127]</td>
</tr>
<tr>
<td>Cyclic natriuretic peptide (CNP)</td>
<td>22 amino acids</td>
<td>Cell surface binding</td>
<td>[128]</td>
</tr>
<tr>
<td>WASP</td>
<td>GBD domain (residues 230-310)</td>
<td>Activation switch: Cdc42/Rac binding</td>
<td>[129, 130]</td>
</tr>
<tr>
<td>E-cadherin</td>
<td>C terminus cytoplasmic domain</td>
<td>β-catenin binding</td>
<td>[131, 132]</td>
</tr>
</tbody>
</table>
IDPs are implicated in various diseases like cancer, cardiovascular disease, diabetes and neurodegenerative diseases, thus making them attractive drug targets [28]. P53 is an extensively studied transcription factor that regulates cell cycle and apoptosis with more than 150 genes as its target [133, 134]. Loss of function of p53 is the cause for a major class of cancers [135]. Around 70% of interactions with p53 are mediated by IDRs [92]. The conformation of Tau protein in Alzheimer’s disease is random coiled before aggregation while the disordered N-terminal of prion protein is found to induce prion protein into aggregation [136-138]. A mouse Arf protein containing 37 N-terminal amino acids is found to be unstructured in solution. This region is known for its regulatory role on p53 and is an important research target for human cancers [103]. Studies on full-length hamster prion protein show a highly flexible N terminus and is predicted to participate in the conversion of PrPC to PrPSc [105]. The natively unfolded prothymosin alpha is a highly conserved protein that is widely distributed in a variety of tissues. Different experimental characterizations have confirmed its structure to be a random coil [32]. Another long fragment of disordered region is found in fibronectin-binding protein and is shown to play a part in protein-protein interactions [110]. BRCA1 is a tumor suppressor gene product that is active in many essential biological pathways including cell-cycle checkpoint control, DNA damage response and tumor evolution. The central region of BRCA1 is indicated to be a long flexible scaffold for multiple intermolecular interactions, from combined NMR and CD studies [112, 113]. Free wild-type alpha-synuclein is disordered in solution and its conformational plasticity is important for the development of Parkinson disease (PD) [115, 118]. Deposits
of amyloid beta-protein are hypothesized to give rise to Alzheimer’s disease (AD). The transformation of the disordered structure of beta-amyloid to an ordered fibrillar aggregates is a topic of intense study for AD researchers [120, 139, 140]. Early study shows 79% proteins in cancer and 55% proteins in cell-signaling are predicted to contain 30 residues or longer disordered segments [42]. Abundance of IDPs in various diseases reiterates the need to identify and characterize disorder. The sequence composition of IDPs is demonstrated to be remarkably different from that of folded proteins and this fact can be used to distinguish IDPs from ordered proteins.

1.1.4 Sequence preference of IDPs

For ordered proteins, it is believed that their amino acid sequences determine their secondary and tertiary structures [142]. Once folded proteins denature, they lose native functions [10]. IDPs are naturally disordered and functional, so the presence of distinct proportions of amino acids in their sequence, that determines their conformational states, has been extensively studied. Early studies have shown that the amino acid composition of some protein segments deviate from the average amino acid composition with statistical significance [143, 144]. This revealed a new perspective to relate protein sequence composition to the disorder-ness in their structure. Dunker and his colleagues compared ordered proteins sequences from a non-redundant PDB dataset with less than 25% identity [145] to disordered sequences taken from missing residues in X-ray, NMR and CD experiments. Order promoting amino acids such W, C, I, Y and V were found depleted in disordered datasets while disorder promoting amino-acids such as R, S, P, E and K were favored in disordered proteins [146]. Bulky amino acids L, V, I and aromatic amino
acid like W, Y and F are found to assemble the hydrophobic core of ordered globular proteins. This results in hydrophobic collapse and is one of the principle reasons for the protein to fold into compact native structures. Disordered proteins lack the ability to pack tight, partly due to a shortage of these hydrophobic residues. On the contrary, IDPs are enriched in polar residues like R, G, Q, S, P, E and K. Structure breaking amino acids G and P are even more prevalent in IDPs. Based on these amino acid compositional differences between ordered and disordered proteins, disorder prediction tool such as PONDERs (Predictors of Natural Protein Disordered Regions) have been developed using artificial neural networks, which brings a new aspect to protein sequence – functional study of disordered proteins [146]. GlobPlot is another disorder-order differentiation tool developed by Linding et. al. using a non-redundant set of proteins from the SCOP database [147] as the ordered protein data set and proteins with missing electron density from PDB files as the disordered set. Prediction is based on the propensities of amino acids in disordered or globular proteins but it can not be used for quantitative analysis across the whole proteome [148]. Later improvements in artificial neural networks generated the second version of the predictor DisEMBL. This tool enabled a broader definition of disorder and enhanced the accuracy in coil prediction [149]. DISOPRED [150], TOP-IDP [151], and FoldIndex [152] were developed using similar methods with higher prediction accuracies. Another set of disorder predictors that do not depend on a training dataset of IDPs, but instead forecast the potential of a protein to fold from its amino acid composition. These kinds of predictors include IUPred [153], FoldUnfold [140], and UCon [154]. Relying on residual interaction energy to differentiate
the folded and disordered states, IUPred yielded 76.33% true positives indicating that disorder occurs in proteins with too few stabilizing interactions within a protein [155]. More than 50 disorder predictors have been developed so far. Predictors based on compositional bias alone can reach an accuracy of 87% [156]. This indicates that composition is a determinant of disorder. More specifically, disordered proteins with a high content of prolines and glycines represent a crucial class [141]. Different from ordered proteins it is the composition rather than exact conservation of sequence that is shown to be important for disordered proteins.

Because of the compositional bias, disordered proteins tend to possess higher net charge and lower hydrophobicity compared to globular proteins [157]. Due to fewer hydrophobic groups in IDPs and IDRs, the driving force to form a compact hydrophobic core is reduced [157], and this explains the lower degree of compactness of IDPs. However, it is not feasible to confirm a disordered protein solely by its net charge and mean hydrophobicity. Low sequence complexity is another commonly recognized feature of disorder proteins [146]. But high sequence complexity alone does not distinguish IDPs from ordered proteins. Combination of multiple prediction tools could improve the accuracy in identifying IDPs. META-Disorder predictors incorporating multiple unrelated predictors can achieve higher accuracy than each individual component [158]. This is because different training datasets enable each predictor to detect disordered proteins from different aspects. Though results may vary depending on individual predictor, the abundance of IDPs in nature is confirmed through various predictors. More than 30% eukaryotic proteins are predicted to possess at least 30 amino acid long
disordered regions [159, 160]. For mammals, 75% signaling proteins are predicted to have long disordered regions[48]. With such significant amounts of IDPs, the experimental characterization and simulation of their structural ensembles are in urgent need.

1.1.5 Experimental characterization of IDPs

X-ray crystallography defines disordered proteins with missing electron density, which is caused by positional variation of flexible atoms that fail to scatter coherent X-rays [52, 60, 145, 161, 162]. But technical difficulties could also lead to missing electron density, which leaves this method with uncertainties in determining IDPs. Even though a crystal of a protein is solved, it only represents one conformation rather than the ensemble in solution. NMR spectroscopy is an important technique to obtain dynamic 3D structure information of proteins in solution. Hydrogen exchange labeling enables high-resolution structural characterization with NMR spectroscopy. Many complete backbone resonance assignments of IDPs and IDR [54, 82, 163-167] prove the capability of NMR to investigate conformational changes for disordered proteins. However, difficulties arise when proteins of high molecular mass (>40kDa) are studied in NMR. Poor signal dispersion could also give false positive results in identifying IDPs.

Circular dichroism (CD) spectroscopy provides representative structural signature of proteins. Far-UV CD spectrum determines protein secondary structure and distinguishes random coil from ordered conformations. Near-UV CD spectrum provides information on tertiary structure by indicating the surrounding environment of aromatic groups [168-170]. The limitation of these methods is that no information on the residue
location can be obtained besides the secondary structure implication. So these are incapable to give clear descriptions of proteins composed of both ordered and disordered segments. CD has been shown to falsely report on the disorder of structural loops [155]. Small-angle X-ray scattering (SAXS) is a structural method especially suitable for the study of flexible proteins in solution. The average large sizes of unstructured proteins can be detected in SAXS by looking at the radius of gyration. A newly developed ensemble optimization method (EOM) generates a pool of random conformers that covers the protein conformational space and then selects those conformations that fit the experimental scattering pattern. The characterization of denatured lysozyme is a good example of the application of this method [171]. As a low-resolution method, SAXS is not able to measure atomic positions inside the protein and so is an ideal complement to X-ray crystallography and NMR. A recent finding on the disordered p53 TAD (transcription activation domain) is from a combination of studies using NMR and SAXS [167].

FRET could determine the distribution of end-to-end distances of IDPs with fluorescent dyes labeled at termini. Rapid and slow conformational fluctuations can be detected by time-resolved FRET and single molecule FRET [172, 173]. Accumulating FRET based reports of conformational behavior of IDPs reveal great potential [174-179]. However, no single experimental technique could discern residues to be ordered or disordered. Combination of multiple approaches would enhance the credibility of detection, but sometimes they disagree since they characterize disordered proteins using different parameters. Due to the highly flexible nature of IDPs, it is more reasonable to describe them as an ensemble of interconverting
structures based on the probability of adopting certain conformations rather than solve a fixed set of atomic coordinates. Descriptions of IDPs from diverse experimental approaches describe the averages of structural parameters and are inadequate at specifying the intrinsic conformational distributions. Even the most exhaustive experimental restraints from NMR cannot effectively describe disordered structure ensembles but they could serve in simulation as constraints to generate conformational ensembles to be consistent with experimental data [27].

1.2 Molecular Simulations

The first molecular dynamics based programs were used to simulate liquids with hard sphere models [180]. Subsequently, models simulating proteins were developed. With the growth in computing power, simulations of proteins have been developed from in-vacuum systems to implicit or explicit water solvation systems. The first simulation of a biological macromolecule was carried out on the protein bovine pancreatic trypsin inhibitor in vacuum for 9.2 ps by McCammon et al. in 1977 [181]. This work significantly changed our view of proteins from relatively rigid structures to dynamic systems with internal motions. Since then, more studies have been done to investigate the internal motions of proteins and nucleic acids. All-atom models represent every atom in the system thus they require substantial computer resources when modeling large systems. Coarse-grained models are more suitable for the investigations of longer time and proteins of longer lengths by reducing amount of information. Conformational sampling in molecular simulation includes molecular dynamics (MD) and Monte Carlo (MC) simulations. MD is a deterministic approach that determines future positions
of atoms with Newton’s equation of motion. It provides the actual trajectory of the system and enables characterization of the thermodynamic state of the system. For large molecular systems, it is a computationally intensive task to carry out MD simulations. MC is a stochastic approach that accepts or rejects a configurational change according to an energetic criterion. Thermodynamic properties of the system are calculated from an accumulated number of configurations in MC. Molecular simulations of IDPs/IDRs incorporate their empirical potentials to characterize ensembles. Experimental restraints can guide simulations to produce ensembles that are consistent with experimental characterization [182]. According to the relatively flat energy landscapes of IDPs/IDRs (Figure 1.1), conformations of proteins generated by MC simulation can “jump” around the energy landscape without being stuck in local minima [183, 184] while in MD simulation the conformational space of a protein may be trapped in a small number of low energy states which lead to poor conformational sampling.
Folded proteins possess a well-defined global energy minimum while IDPs correspond to a relatively flat energy landscape with numerous local energy minima. This figure was modified from [185].
The Trajectory Directed Ensemble Sampling (TraDES) program generates unfolded random protein conformers in all-atom detail using a trajectory distribution as input to direct the sampling. This method is not bound by explicit potential scoring functions. Information from a protein sequence alone could produce sterically plausible conformers. The input trajectory file is a collective set of the frequency distribution of each amino acid, which is based on a trajectory distribution dictionary made from a non-redundant set of 834 PDB structures [186]. TraDES has been used to generate ensembles of structures of unfolded proteins successfully [187-190]. Recent hybridization of experimental restraints and conformational sampling is a popular method generally used to produce structural ensembles of disordered proteins. ENSEMBLE is such an example that incorporates chemical shifts, NOEs, PREs, solvent-accessible surface area, hydrogen exchange protein factors and hydrodynamic radius to generate conformers that are consistent with experimental restraints to minimize the chance of over-fitting [7]. In ENSEMBLE, TraDES provides the initial conformational pool of unfolded proteins for experimental data fitting. Another similar approach is the flexible-mecccano algorithm, which samples the conformational space of each amino acid in a disordered protein from a coil conformational library [191]. Experimental restraints can be used to match ensemble data from conformers with experimental data but experimental methods in obtaining these restraints may cause perturbations in ensemble conformations [27]. Since simulations actually do not require experimental restraints as input, it is advisable to use experimental data for validation rather than input to minimize the chance of introducing artifacts [192].
1.3 Formin

Woychik first introduced the name “formins” for proteins he discovered to be encoded by the mouse *limb deformity (ld)* gene. Disruption of this gene would cause limb defects [193]. Later Castrillon and Wasserman found out the protein encoded by *Drosophila* gene *diaphanous* is homologous to the formin protein and is required for cytokinesis. Studies on protein Bni1p, which is a product of a *Saccharomyces cerevisiae* gene, showed it also has the same domains of above mentioned proteins, so the term formin homology domains came into use, with formin homology domain 1 (FH1) and formin homology domain 2 (FH2) first defined [194]. Till now more formins are identified in a wide range of organisms among plants, animals and fungi [195-199]. The cellular processes in which formins participate vary from cytokinesis, cell polarity, filopodium formation, cell adhesion and migration, morphogenesis, endocytosis and microtubule-actin cross talk [200, 201]. As the key regulators of cytoskeletal organization, formins are involved in actin nucleation, polymerization, membrane integration and spindle association [198, 202]. The most studied type-I plant formins are shown targeting to the plasma membrane [203-205]. Cdc12 is found forming cytokinesis ring while Bni1p and Bnr1p stay at the bud tip [206-208]. dDia2 and mDia2 are detected on the filopodia tip [209, 210].

1.3.1 Formin domain organization

Formins are defined by a highly conserved FH2 domain and a proline-rich FH1 domain (Figure 1.2). The FH2 domain initiates *de novo* actin filament nucleation [211], FH1 resides before the N-terminal of the FH2
domain and the two domains are flanked by conserved coiled-coil regions [194]. FH2 is a well-folded domain that forms a tethered dimer by a flexible linker in between [212]. It binds to the barbed ends of actin filaments [211, 213], where the actin filament assembly is promoted [214]. Each half of the FH2 dimer is competent to bind the barbed ends of actin filaments [212]. With both half dimers binding to the filament ends, FH2 acts as a capping protein to protect the filament ends from other cappers [215, 216], nonetheless allows processive barbed end polymerization when half of the dimer detaches from the end and binds to the newly inserted actin monomer [217]. Thus came the stair stepping mechanism of formins explaining its processive motion [212]. It was also shown that FH2 stays persistently on the barbed ends of actin filaments and new subunits are inserted between formin and the filament barbed ends. This proved the stair stepping hypothesis, but a lack of supercoil formation on microscope slides through immobilized formin indicated the necessity of formins to reorient on actin filament barbed ends [217]. A screw mode later incorporated FH2 dimer rotation by including torsion elastic stresses [218]. The rotational motion of formin along the actin filaments was validated by single-molecule fluorescence polarization recently. [219]. FH2 domain nucleates unbranched actin filament but inhibits polymerization [220]. FH1 domain changes this situation by recruiting actin bound profilin and thus promotes barbed end actin assembly. FH1 is a compositionally biased domain with a high content of prolines [6, 195]. Interaction of FH1 domain and profilin vastly promotes actin monomers addition to formin-associated filament ends [6, 221].
Besides the most recognized FH1 and FH2 domain, there are studies updating information on other common domains as well. Petersen et al. first described FH3 domain as a localization domain for fission yeast formin Fus1 and later it was also found to have the same function for mDia proteins [222-224]. The N-terminal GTPase-binding domain (GBD) of formins is found to be a target of Rho GTPases [197] participating in cytoskeleton reorganization [225, 226]. The location of GBD often overlaps with FH3 domain [222]. The Diaphanous-autoregulatory domain (DAD) is localized at the C-terminal of the FH2 domain. It binds to GBD when formin remains inactive. When activated Rho GTPases bind to GBD, DAD is released from GBD turning formin into its active state [227-229]. The DAD is reported to have a relatively more conserved N terminus than its C terminus [199]. Through comparative

Figure 1.2 Schematic illustration of formin domain organization. Crystal structure is adapted from PDB with accession number 3EG5 for GBD/FH3 domain, 1UX4 for FH2 domain and 2F31 for DAD domain.
sequence analysis the GBD/FH3-FH1-FH2-DAD architecture is revealed to be a common representation of Dictyostelium, fungal and metazoan formins [230].

1.3.2 Disordered FH1 domain

FH1 domain is always studied together with the FH2 domain for they are the defining feature of the formin protein. Only after GBD/FH3 and DAD these regulatory domains were mapped to their corresponding functions, detailed functional studies of FH1 and FH2 became the focus in the research of formins. Different from the well-folded FH2 domain, FH1 domain is a low complexity sequence enriched with prolines. No conserved segments can be found in FH1 via multiple sequence alignment. The number of polyproline tracks inside each formin FH1 domain differs dramatically; however, the presence of polyproline tracks seems a conserved feature for FH1 domain. After querying formin homology 1 in UniProtKB, there were 37 entries and those including functional studies are listed in Table 1.2. At least one polyproline track was found in these formins, which indicates the essential role of polyproline track for FH1 domain function.
<table>
<thead>
<tr>
<th>Protein</th>
<th>Origin</th>
<th>FH1 domain length (amino acids)</th>
<th>Number of polyproline tracks</th>
<th>Function</th>
<th>Accession number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formin-2</td>
<td>Mus musculus</td>
<td>390</td>
<td>14</td>
<td>Response to DNA damage, actin-binding, actin nucleation factor, spindle positioning</td>
<td>Q9JL04</td>
<td>[231-233]</td>
</tr>
<tr>
<td>Formin-1</td>
<td>Mus musculus</td>
<td>101</td>
<td>5</td>
<td>Polymerize linear actin cables, form adherent junction</td>
<td>Q05860</td>
<td>[234, 235]</td>
</tr>
<tr>
<td>Formin-2</td>
<td>Homo sapiens</td>
<td>511</td>
<td>28</td>
<td>response to DNA damage, protect cell against apoptosis</td>
<td>Q9NZ56</td>
<td>[236, 237]</td>
</tr>
<tr>
<td>Formin-1</td>
<td>Homo sapiens</td>
<td>88</td>
<td>4</td>
<td>Form adherent junction, polymerize linear actin cables</td>
<td>Q68DA7</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>Drosophila</td>
<td></td>
<td></td>
<td>Actin nucleation factor, involve in cytoskeleton dynamics and transport</td>
<td>Q24120</td>
<td>[238, 239]</td>
</tr>
<tr>
<td>Protein</td>
<td>melanogaster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formin-I</td>
<td>Dictyostelium</td>
<td>62</td>
<td>2</td>
<td>Nucleation of actin filament</td>
<td>Q54PI9</td>
<td>[230, 240]</td>
</tr>
<tr>
<td>FHOS</td>
<td>Homo sapiens</td>
<td>129</td>
<td>3</td>
<td>Assemble actin filament, involve in cell elongation and</td>
<td>Q9Y613</td>
<td>[241-243]</td>
</tr>
<tr>
<td>Gene</td>
<td>Species</td>
<td>Accession</td>
<td>Molecular Function</td>
<td>Protein Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>-----------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRF1</td>
<td>Homo sapiens</td>
<td>O60610</td>
<td>Assemble F-actin, nucleate actin filament, regulate actin dynamics, control cell shape</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mDia1</td>
<td>Mus musculus</td>
<td>O08808</td>
<td>Assemble F-actin, stabilize microtubules, regulate cell morphology and cytoskeleton</td>
<td>[220, 245-248]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRF3</td>
<td>Homo sapiens</td>
<td>Q9NSV4</td>
<td>Promote actin polymerization, regulate actin dynamics</td>
<td>[249, 250]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formactin-2</td>
<td>Homo sapiens</td>
<td>Q2V2M9</td>
<td>Actin polymerization</td>
<td>[251, 252]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bni1p</td>
<td>Saccharomyces cerevisiae</td>
<td>P41832</td>
<td>Control mitotic spindle position, regulate F-actin, organize microtubules</td>
<td>[253-255]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formin-A</td>
<td>Dictyostelium discoideum</td>
<td>Q54WH2</td>
<td>Actin nucleation</td>
<td>[230, 256]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formin-B</td>
<td>Dictyostelium discoideum</td>
<td>Q54SP2</td>
<td>Actin nucleation</td>
<td>[230, 256]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inverted formin-2</td>
<td>Mus musculus</td>
<td>144</td>
<td>7</td>
<td>Sever actin filament, promote actin dynamics</td>
<td>Q0GNC1</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>--------------</td>
<td>-----</td>
<td>---</td>
<td>---------------------------------------------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>Cdc12p</td>
<td>Schizosaccharomyces pombe</td>
<td>233</td>
<td>2</td>
<td>Form cell division ring, cap and nucleate actin filament</td>
<td>Q10059 [258, 259]</td>
<td></td>
</tr>
<tr>
<td>DAAM1</td>
<td>Homo sapiens</td>
<td>72</td>
<td>3</td>
<td>Direct actin filament nucleation and elongation</td>
<td>Q9Y4D1 [260, 261]</td>
<td></td>
</tr>
<tr>
<td>Formin-G</td>
<td>Dictyostelium discoideum</td>
<td>27</td>
<td>1</td>
<td>Nucleate linear actin filament</td>
<td>Q1ZKK2 [230, 256]</td>
<td></td>
</tr>
<tr>
<td>Formin-J</td>
<td>Dictyostelium discoideum</td>
<td>27</td>
<td>2</td>
<td>Nucleate linear actin filament</td>
<td>Q54ER5 [230, 256]</td>
<td></td>
</tr>
<tr>
<td>Formin-F</td>
<td>Dictyostelium discoideum</td>
<td>124</td>
<td>5</td>
<td>Nucleate linear actin filament</td>
<td>Q5TJ56 [230, 256]</td>
<td></td>
</tr>
<tr>
<td>Formin-E</td>
<td>Dictyostelium discoideum</td>
<td>63</td>
<td>3</td>
<td>Nucleate linear actin filament</td>
<td>Q5TJ57 [230, 256]</td>
<td></td>
</tr>
<tr>
<td>Formin-H</td>
<td>Dictyostelium</td>
<td>27</td>
<td>2</td>
<td>Control filopodial dynamics</td>
<td>Q54N00 [209, 240,</td>
<td></td>
</tr>
<tr>
<td>dDia2</td>
<td>discoideum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRF2</td>
<td>Homo sapiens 75 4 Regulate endosome dynamics</td>
<td>O60879</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fhos1</td>
<td>Mus musculus 157 4 Assemble F-actin, involve in cell elongation</td>
<td>Q6P9Q4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fhos3</td>
<td>Mus musculus 32 2 Form stress fiber, involve in cell elongation</td>
<td>Q76LL6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mDia2</td>
<td>Mus musculus 71 2 Promote actin polymerization, regulate actin dynamics</td>
<td>Q9Z207 [226, 248, 264]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mDia3</td>
<td>Mus musculus 77 5 Regulate microtubule dynamics</td>
<td>O70566</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1.3.3 Interaction of profilin and formin

Src homology 3 (SH3) and WWP/WW domains are the first two binding motifs discovered to bind the proline-rich FH1 domain among formin binding proteins (FBPs) [266]. Watanabe et al. found mouse formin p140mDia (mDia1) binds to profilin in regulation of actin filament formation [197]. This finding put formin and profilin as the focus in studies concerning actin filament assembly. Profilin is a small (14-17 kDa) actin-binding protein universally present in eukaryotes [267]. Since it was first discovered in 1977 as actin filament dynamics regulator, more binding partners characterizing profilin as a multifunctional protein have been found such as actin related protein [269], polyphosphoinositides [270, 271] and poly-L-proline (PLP) [272, 273]. Later, structural and biochemical studies proved that profilin binds to G-actin and PLP independently [274, 275]. Furthermore, mutation assay shows binding to poly-L-proline is an essential function for profilin. Although the sequences of different profilins vary, they reserve similar three-dimensional structures. The sequence variety of profilin results in diverse affinities for their ligands [276]. The role of profilins in actin polymerization was first proposed to keep monomeric actin [268]. New experimental observations gave the detail of profilin-actin filament interaction that profilin inhibits actin filament elongation more at the pointed end than the barbed end, and the inhibition effect is stronger on filament nucleation than elongation [277]. Profilin is known to bind poly-L-proline containing proteins while formins are the only class of poly-L-proline containing proteins that binds profilin in plants. This indicates the importance of profilin-formin interaction in actin dynamics regulation [203]. Binding with formin changes the role of
profilin in actin filament formation. Besides promoting formin mediated barbed end filament assembly, profilin also enhances formin mediated nucleation [278].

1.3.4 Formins in actin filament assembly

Major discovery of the role of formin in actin filament assembly comes from Sagot et al. and Evangelista et al. Both groups found that yeast formin nucleates actin filaments in an Arp2/3 independent manner [208, 279]. Studies later verified that FH1 and FH2 domains are both essential for nucleation of profilin bound actin monomers [213, 278]. Pring et al. confirmed the contribution of proline-rich FH1 domain in filament nucleation by deleting it, which leads to decreases in employing profilin-actin for nucleation [280]. Apart from being actin filament nucleator, yeast formin cdc 12p is shown as a barbed end capping protein as well. Actin polymerization experiments with yeast formin Bni1 FH1-FH2 and FH2-only peptide indicated the capping function is FH2 dependent [280]. Kovar et al. discovered that the polymerization rate of formin bound filament grows at the same rate on barbed end as free filament. Considering formins’ capping attribute, they are postulated to nucleate filament exclusively from pointed end [259]. The study using Bni1p also confirmed the capping role of formins during actin elongation, but the formin protein was found to serve as a leaky cap that stays at the barbed end of filament slowing down both polymerization and depolymerization according to the observation of fluorescent filament growth. As a leaky cap, the formin protein still allows nucleation from the barbed end which is contrary to the pointed end growth model that Kovar et al. suggested [214]. Thus two opinions have formed towards how formins facilitate actin
filament elongation while binding to their barbed ends. Later Higashida et al. obtained live cell images using different mDia1 mutants and discovered that formin is persistently associated with the barbed end and moved processively along it. Furthermore, they discovered that formins move independent of myosin, which leads to the rationale that formins may be a motor [246]. Kovar et al. at the same time also unveiled this phenomenon through measuring the force generated by formin polymerization with single filament anchored on glass. This method is noteworthy as it enables direct observation of the addition of new actin monomers between the FH2 domain and the barbed end of actin filament [217].

1.3.5 Molecular motors

Molecular motors are biological machineries inside cells that powers cytoplasmic transportation. The most well studied molecular motors so far are actin-based myosins, microtubules-based kinesins and dyneins. At first there was little in common found in these different motors. Crystal structures later revealed that the ATP-binding sites in myosin and kinesin are identical in their structures. Comparative study of these three motor families indicated a common evolutionary origin shared with G proteins [281]. Myosin detaches from its track with every step it moves; while kinesin and dynein move along the microtubules for many steps before dissociation, which is referred to as processivity. One common feature of these motors is that they turn chemical energy into mechanical work. The source of chemical energy is usually the hydrolysis of ATP-related compounds or the polymerization of proteins or nucleic acids. The basic structural component of a molecular motor is the motor domain that binds to its track, such as actin filaments and microtubules.
It is the place where the enzymatic activity occurs. The motor domains are linked by coiled-coil stems to tail domains that associate with cellular cargoes like vesicles and organelles [282]. Myosins, kinesins and dyneins will be hereafter mentioned as canonical molecular motors.

Formins are known to be molecular motors that are involved in actin-based motility. The available crystal structures of formins indicated FH2 domain works in a similar manner to the motor domain of canonical molecular motors. The track of formins is the actin filament they are bound to and the FH2 domain walks along the barbed end processively as well. The FH1 domain that binds to profilin-actin works like the tail domain of canonical molecular motors in binding cellular cargos. About the energy source of formin motors, it is still in dispute whether ATP hydrolysis plays a part in this system. The main difference between canonical molecular motors and formins is that canonical molecular motors primarily assist in transportation not assembly. In contrast, formins work in an architect style, which builds sprouting motile motif or constructs essential parts of cytoskeleton that serves as a track for myosin, e.g., the FH1 domain recruits actin filament building blocks: free profilin-actin monomers for filament assembly, rather than merely translocate the protein to a different location.
Figure 1.3 Schematic illustration of the direct transfer polymerization model. (A): Profilin-actin-ATP (P•A) binds to the polyproline tracks (X) to form a X•P•A complex and the affinity-modulated clamp locks to the end of a filament. (B): P•A is directly transferred to the end of filament. (C): The clamped subunit is subjected to ATP hydrolysis after new monomer is added and the clamp moves to the new subunit. The bound profilin is dissociated and starts a new cycle of monomer addition. This figure was modified from .

Abundant studies of actin-based motility demonstrated the importance of profilin-actin concentration in the polymerization zone. The locomotion was halted with free polyproline segments injected into cells of Listeria [284], Shigella [285], and vaccinia [286]. It suggested that free profilin-actin recruiting motif failed to contribute to actin-based motility. Dickinson et al. considered the importance of profilin-actins in actin filaments assembly and proposed a direct transfer polymerization model (Figure 1.3), in which the profilin-actins are recruited directly by motifs belong to “actoclampin motor” that is attached to the motile surface [283]. The distance between the filament and the motile surface is very short according to the high concentration of profilin-actin needed for the polymerization rate measured in experiments
Among the models hypothesized for molecular motors, the Elastic Brownian Ratchet model is well known which specifies that unattached filaments driven by thermal motions produce a directed force [288]. However, this Elastic Brownian Ratchet model cannot explain various tethered filament mediated protrusion and the staccato stepwise motions found in *Listeria* [289]. In contrast, the direct transfer polymerization model attaching to the motile object explains these stepwise motions in *Listeria* [290]. Formins work similar to the “actoclampin motors” proposed by Dickinson [283]. The FH1 domain uses its multiple polyproline tracks to accumulate a reservoir of profilin-actins, while the FH2 domain functions as the “affinity-modulated clamp” [283].

Studies on FH2 domain showed the “clamp” structure as a donut shaped FH2 dimer that remains persistently on the barbed end. Studies about the Rho-related pathway also revealed the attachment of formins to the cytoplasmic membrane, which leaves the behavior of FH1 domain in the mist.

### 1.3.6 Formins as motors

The FH1 domain binds to profilins with low affinity [196, 291] and it delivers profilin-actin rapidly to the barbed end where the FH2 domain stays. Carlier *et al.* tested the function of FH1 domain in actin filaments polymerization by conducting filament assembly assay using two different groups of formins: mDia1 FH1-FH2 and mDia1 FH2 domain alone. The polymerization rate of formin-mediated filaments at the barbed end was found to be 15 fold faster than the free barbed ends. In contrast, mDia1 FH2 bound filaments grew 2 fold slower than free barbed ends [292]. Therefore, the FH1 domain was proved to promote actin filament polymerization together with the FH2 domain, which alone inhibits polymerization even in the presence of
profilin. Later Kovar and colleagues also proved that formin increases the elongation rate upon binding profilin, although in different magnitudes, the FH1 domain is always required [293]. A study using formin Bni1p constructs composed of one to eight profilin-binding sites showed that the elongation rate of actin filament barbed end increases with the number of profilin binding sites in the FH1 domain [6]. Therefore, formin FH1 domain is demonstrated to be essential for the acceleration effect it exerts on actin filament polymerization and the number of profilin binding sites inside FH1 domain is positively linked to this effect. Since the FH2 domain would form a dimer when it binds to the barbed end of actin filament, the FH1 domain always appears in pairs this way. To elucidate whether the FH1 domain is able to work independently while the FH2 domain dimerizes, the essay of hybridized formin Bni1(FH1FH2+FH2)p was used to test the acceleration effect against the wild type Bni1(FH1FH2)p. The result showed that two FH1 domains of a formin dimer do promote elongation with twice the speed of only one FH1 domain, yet one FH1 domain still promotes assembly [294]. From the above we can see that the number of FH1 domain as well as the number of profilin binding site inside FH1 domain relate to the formin motor behavior closely. Many models have been proposed to fit formin-driven actin polymerization so far, but they all fail to take FH1 domain into consideration [221, 295, 296]. In consequence, the challenge here is to include the essential FH1 domain for a more complete description of the formin motor behavior.

About the energy source of formin motor, Blanchoin et al. found the hydrolysis rate of ATP is 2 orders less than the addition rate of profilin-actin to the barbed end of filament [297]. So it is deduced that ATP hydrolysis is
not required for profilin-actin based assembly. However, Romero et al. hold the opinion that the free energy for formin motor derives from ATP hydrolysis, which is accelerated by FH1-FH2 during profilin-actin polymerization [292]. Since profilin needs to dissociate from the barbed end to make it available for addition of new profilin-actin for filament elongation and ATP hydrolysis decreases profilin actin affinity, it is a reasonable deduction that actin bound ATP hydrolysis is accelerated by formins as well. On the other hand, the observation of single filament polymerization indicates a 30-fold higher elongation rate of formin promoted assembly compared to ATP hydrolysis rate. This new evidence separated the role of ATP hydrolysis from the filament elongation mediated by formins again [217]. More discoveries from Kovar’s group showed that, ATP hydrolysis is unrelated with the formin accelerated elongation activity. It was reported that ADP-actin-profilin-formin assembles faster than ADP-actin-profilin-free end in which ATP does not participate at all [293]. In addition, Romero’s group did a similar experiment using ADP-actin while their result indicated that ADP-actin abolishes the assembly activity because of the inhibition of profilin [298].

1.4 Aims and Objectives

From the previously published reports we could see that most research work concerning formins is FH2 domain focused. Studies on the FH1 domain that confirmed its importance for the formin motor function did not address the structural information of FH1 domain during the profilin-formin binding event. Thereafter, the conformational changes that the FH1 domain goes through to realize the formin motor function is still elusive and the arguments on the energy source of formin motor have not reached a consensus yet.
There has no available high-resolution structural information of formin FH1 domain by now from experimental approaches. The simulation software such as TraDES is therefore a great tool to the study of the conformational ensemble of the FH1 domain in bound and free states. The capability of the FH1 domain in contributing energy to formin motor function is unclear as well. As a result, it is intriguing to detect the conformational changes of formin FH1 domain and to explore the thermodynamic profile from its free state to profilin-bound state. The objectives of this thesis are to:

- investigate the conformational changes of formin FH1 domain through analyzing its structural ensembles generated by TraDES package
- examine the thermodynamic data of formin-profilin binding by Isothermal Titration Calorimetry
- evaluate the conformational changes of formin FH1 domain using unbound mDia1 FH1 domain and profilin-bound mDia1 FH1 domain by luminescent resonance transfer and luminescence decay experiments

The results of this thesis would provide insights into the conformational behavior of the disordered FH1 domain of formin motor. Additionally, the study here helps achieve a better understanding on how the disordered FH1 domain functions inside formin motor while incorporating profilin into the system. The examined thermodynamic profile may tell us the mechanism in which the disordered FH1 domain actively turns the formin protein into a motor with the help of profilin. The analysis in the conformational behavior and energetic profile of the disordered FH1 domain gives an alternative explanation about the energy source of the formin motor.
The approaches employed in this thesis also apply to the studies of other intrinsically disordered proteins in mechano transduction pathways.

This thesis focuses on the behavior of formin FH1 domain when the binding of profilin happens. Although the FH2 domain is also essential for formin normal function, the motor function of formins is mainly dependent on the FH1 domain. So the FH2 domain is beyond the scope of this thesis and is excluded in all the experiments designed for this thesis. To search the energetic source within the self-motivated model, external energy source like ATP is not considered in this thesis.

The next chapter gives more details about the TraDES software package. Since the study is on disordered domain of a protein, TraDES has the advantage in generating structural ensembles of disordered proteins in reasonable time. The simulation results from TraDES will be discussed as well.
Chapter 2
Monte Carlo Simulation of the Disordered Formin FH1 Domain

2.1 Background

Actin based motility is essential for various cell activities such as cell migration, adhesion, cytokinesis as well as morphogenesis and endocytosis [299, 300]. Because the actin filament nucleation by actin monomers alone is not energy favorable, nucleating factors are needed to initiate a nucleus to start [301]. There are plenty of factors in cytoplasm that depolymerize actin filaments or cap their ends. Auxiliary complex is needed to overcome the energy barrier of initiating actin filaments polymerization and protect the growing barbed end from capping proteins. Formins, the unbranched actin filament nucleators, are capable to do both tasks.

Formins, widely found in fungi, plants, and animals regulate cytoskeleton structure as a downstream target of the small GTPase Rho [197, 225, 302]. They actively participate in cytokinesis, cell polarization and stress fiber formation [195, 302]. The role of formins in modulating nucleation and polarization of unbranched actin filaments is in an Arp2/3 complex independent manner [208, 211]. Functioning as “leaky caps”, formins protect the barbed end of actin filament from tight capping proteins and increase the elongation time frame of the filament [214]. Formins stay persistently along the barbed-end of actin filament, partially inhibiting both polymerization and depolymerization by 50% [280]. Profilin, an actin assembly regulator that binds to both actin and poly-L-proline, speeds up the polymerization of monomeric actin [259]. With evidence from diverse formins, profilin turns the
formin protein from a capper into a processive motor accelerating filament elongation [293].

As highly conserved multidomain proteins, formins are characterized by their common domains, Formin-homology-1 (FH1) and formin-homology-2 (FH2) [303]. FH1 is a coiled-coil domain of variable lengths. Due to its high proline content and shortage of conserved amino acids, FH1 domain is believed to be unstructured [304]. The multiple poly-L-proline stretches in the FH1 domain are targets of WW domain containing proteins and profilins. FH2 is a dimeric donut shaped domain that persistently associates with the barbed end of actin filaments with *de novo* filament nucleation function [211]. A 20–30 amino acid long sequence located at the C-terminal of FH2 domain is the conserved Dia autoregulatory domain (DAD) [227]. Together with the N-terminal autoinhibitory region, which is named as diaphanous inhibitory domain (DID), DAD and DID form an auto-inhibitory formin structure [305]. The GTPase-binding domain (GBD) partially overlaps with the DID domain and it relieves the auto-inhibition of formins upon Rho GTPases binding [228, 230].

The FH2 domain alone is found to inhibit actin polymerization [220] and this effect is not reversed by adding profilin [213]. On the contrary, FH1-FH2 domain accelerates profilin-actin polymerization by up to 15-fold [292]. It is obvious that the FH1 domain is essential for the motor effect on the FH2 domain associated barbed end elongation via profilin-FH1 interaction. A study using formin Bni1p constructs composed of one to eight profilin-binding polyproline tracks showed that the barbed end elongation rate increases with the number of profilin binding polyproline tracks in the FH1 domain [6]. It
was elucidated that the FH1 domain works independently while the FH2 domain dimerizes and two FH1 domains promote actin elongation at twice the speed of one FH1 domain [294]. This suggests that the FH1 domain works in a cooperative manner and the number of polyproline tracks in the FH1 domain directly affects formin motor function.

Mouse Diaphanous-related formin 1, mDia1, gained researchers’ attention originally because it is the effector of Rho GTPase and required for stress fiber formation [306]. After that, the effect of mDia1 on actin polymerization is revealed [307]. Till now mDia1 is the most investigated formin protein with great details in the functions of its FH1 and FH2 domains [213, 220, 305, 308, 309]. Therefore, mDia1 could serve as a representative for the study of formins. The FH1 domain of mDia1 contains 13 tandem polyproline tracks and its capability of binding profilin is verified by peptide blotting essay [310]. The crystal complex structure of profilin 2a and mDia1 is solved at atomic resolution by Kursula et al in 2008. The complex is composed of two profilin 2a molecules and two consecutive polyproline tracks that are separated by a disordered region of five residues whose electron density is weak. The solved crystal structure renders mDia1 to be an ideal target to study the conformational behavior of the long disordered formin FH1 domain containing multiple polyproline tracks.

Amino acid sequence is supposed to contain enough information for small monomeric proteins to fold into its three-dimensional structure based on Anfinsen’s findings that totally denatured globular proteins could recover their functional structure in seconds [142]. Discovering protein sequence-structure relationship turned out to be an intriguing task. Molecular simulations are
frequently used to predict protein structures and study molecular interactions. Molecular dynamics (MD) and Monte Carlo (MC) sampling are the two major approaches to sample protein conformations. Intrinsically disordered proteins are not confined to adopt one static conformation and they should be represented by a structural ensemble. MD simulations could calculate disordered ensemble using time-averaged restraints but the high computational requirement limits its application to short averaging time or short peptides [312]. MC sampling is used to study large disordered proteins with experimental restraints on ensembles of their conformers [188]. TraDES (Trajectory Directed Ensemble Sampling) is a probabilistic sampling method that generates off-lattice all-atom protein conformers based on one-dimensional 3-state secondary structure information [313]. It is an effective method that narrows down plausible conformational space but still keeps the native states of disordered proteins. TraDES has been used to do unbiased conformational sampling to study several disordered proteins.

Plenty of biochemical studies have addressed the importance of the FH1 domain for its acceleration effect on profilin-actin nucleation and polymerization [6, 292, 293]. However, no exploration has been done on the conformational behavior of this domain because it is non-conserved and highly unstructured. Several working models of formin motor have been proposed but none of them incorporates the function of the essential FH1 domain. Therefore, the motion mechanism of formin motor is still unclear due to a lacking of information about the FH1 domain. Herein, we report the conformational change upon profilin binding to the structure ensemble of mDia1 FH1 domain, which contained four million conformations generated
using TraDES package. To see the simulation result in a reasonable time frame we use a mDia1 FH1 segment that contains six consecutive polyproline tracks for this study. We examined the availability of profilin binding site and the volume exclusion effect of multiple bound profilins bound in the FH1 domain.

2.2 Results

2.2.1 Disordered FH1 domain

The formin FH1 domain is accepted as disordered due to its biased amino acid composition and the absence of its crystal structure in current structure databases. Results from multiple disorder prediction tools indicated the FH1 domain is disordered. Figure 2.1 shows the results from three disorder prediction software RONN [316], IUpred [153] and Globplot [148]. Results from RONN shows the “distance” of formin-FH1 from well-characterized prototype sequences. IUpred predicts disorderness of formin-FH1 based on the number of interresidue interactions. Globular proteins possess plenty of interresidue interactions to stabilize folding while IDPs don’t. Globplot presents the sum of the propensity of amino acids in formin-FH1 to be disordered. They all predicted that FH1 domain has a greater probability to be disordered than being ordered. The structural ensemble of the disordered domain FH1 can be generated by TraDES.
Figure 2.1 Prediction results of the mDia1 FH1 domain [UniProt O08808] using disorder prediction software. Graphical prediction results from RONN A), IUpred B) and Globplot C) are displayed.
2.2.2 Polyproline tracks in the FH1 domain and the pattern for profilin binding sites

Polyproline tracks are mostly viewed as the profilin binding sites in the formin-profilin interaction. The number of proline residues in polyproline tracks varies even within the same FH1 domain. Several research groups have performed formin-profilin binding essays using polyproline tracks of distinct lengths as profilin binding sites. It was demonstrated that long polyproline tracks can bind more than one profilin but the required minimum number of consecutive prolines to bind profilin ranges from six to ten [272, 317]. Additionally, the pattern for single profilin binding sites in formins has not been reported so far. Kursula and colleagues did experiments to bind profilin to the 13 polyproline tracks in mDia1 FH1 domain, and they found that tracks with less than five consecutive prolines lack the ability to bind profilin [310]. Based on these findings, we only consider sequences with no less than five consecutive prolines as a polyproline track in this study and rule out shorter segments.

To search a pattern for profilin binding site in the formin FH1 domain, we queried UniProtKB with “FH1 domain” and obtained 37 hits. Among the 37 FH1 sequences acquired, 182 polyproline tracks were found. Figure 2.2 summarizes the lengths of all polyproline tracks. Short tracks are dominant among these annotated FH1 domains. Pentaproline tracks amounted to 43% and hexaproline tracks took up 21% of the distribution respectively. Long tracks with sixteen and seventeen consecutive prolines are scarce.

The frequencies of amino acids that appear in the upstream and downstream positions surrounding the polyproline tracks is presented in
Figure 2.3. Leu is found to be the primary residue that immediately follows a polyproline track; in contrast the residue that immediately precedes a polyproline track has 31.6% chance to be Ile, 18.9% to be Gly and 16.1% to be Leu. One of the characteristics of IDPs/IDRs is the depletion of bulky hydrophobic amino acids Ile, Leu and Val in their amino acid sequences. Therefore, the conservation of isoleucine and leucine around polyproline tracks is a feature special for the FH1 domain. In the solved crystal structure of mouse profilin structure with mDia1 segment, it revealed that Ile and Leu residues serve as anchors for pentaproleine segment at both ends, each packing against a tyrosine residue in profilin. Based on the above observations, we took IPPPPL as the pattern of profilin binding site.

![Distribution of the length of a polyproline track.](image)

Figure 2.2 Distribution of the length of a polyproline track.
Figure 2.3 Amino acid composition in the upstream and downstream positions surrounding the polyproline tracks.
2.2.3 Classification of free and bound profilin binding sites

All mDia1-FH1 conformers are superimposed with profilins to test the availability of each profilin binding site. The solved crystal structure [PDB:2V8F] of two mouse profilins (Chain A and B) and one segment of mDia1-FH1 peptide containing two profilin binding sites (Chain C) provides the conformational relationship of profilin and its binding site in mDia1 FH1 domain. In this study, the first profilin binding site in Chain C of [PDB:2V8F] was aligned to each profilin binding site in mDia1-FH1 conformers. The Chain C bounded profilin structures were then automatically relocated to the corresponding binding sites in mDia1-FH1. The superimposition of profilins and mDia1-FH1 may result in clashes between a mDia1-FH1 conformer and profilins that “binding” to it or clashes between profilins “binding” to different sites in one mDia1-FH1 conformer. Therefore, a clash check is performed after the superimposition to distinguish the status of a profilin binding site whether it is profilin bound without clashes or it is a free site that does not accommodate previously superimposed profilin. Clash check between atoms are based on van der Waals radii. When the distance between two atoms is less than the sum of these two atoms’ van der Waals radii, a random number is employed to decide whether the collision is allowed.

A successful binding event only happens when TraDES generated mDia1-FH1 sequence aligned well with the Chain C in [PDB:2V8F]; at the same time the number of steric clashes inside the merged structure is within tolerance while docking. The root mean square deviation (RMSD) of all mDia1-FH1 conformers ranges from 0.04Å to 3.56Å. The number of steric clashes for each profilin-FH1 docking complex ranges from 0 to 6949. To
determine a RMSD threshold for a good alignment and a clash tolerance for sound docking, the last 500 conformers of mDia1-FH1 were picked for visualization in MacPyMOL. It was found that 324 structures followed the trace of superimposed [PDB:2V8F] Chain C. The RMSD distribution of these conformers was within 0.5Å. Meanwhile, the rest 176 conformers hugely deviated from the Chain C of [PDB:2V8F] and their RMSD distribution centered around 2.5Å (Figure 2.4). From the clear distinction of the RMSD values of the well-aligned conformers and misaligned ones, 0.5Å was selected as the proper threshold to separate these two groups of conformers. Among the profilin-merged mDia1-FH1 conformers, 110 structures were free of clashes, 129 structures had side chain contacts, which could be released through rotation, and 260 structures had major clashes with two chains clashed together at multiple locations. Figure 2.5 displays the clash number distribution for these 500 profilin merged mDia1-FH1 conformers. Clashes below 65 are not distinguishable. Using the determined RMSD and clash number threshold, we evaluated every profilin binding site in the mDia1-FH1 conformers. According to the availability of each binding site, we classified all conformers into seven groups. Conformers with RMSD less than 0.5Å and clash number less than 65 at all six sites are bound with six profilins. The same rules applied to conformers bound with five to one profilins, and the rest of the conformers excluded from the above six groups are categorized as the seventh group which were free from profilins. The number of structures in each group decreases as the number of bound sites increases (Table 2.1).
Figure 2.4 Determination of the root mean square deviation (RMSD) threshold for a good alignment. RMSD distribution of conformers overlapping with the superimposed [PDB:2V8F] chain C (black), and conformers deviating from the trace of [PDB:2V8F] chain C (green) are shown in this figure. The small bump on the green line is caused by conformers with one to three misaligned prolines but the rest part still fitted the trace of [PDB:2V8F] chain C. Structures generated from TraDES are colored in green and [PDB:2V8F] chain C in pink. Classification was made via structural visualization in MacPYMOL. Alignment samples with RMSD of 0.178, 0.468, 0.88 and 2.56 are displayed.

Figure 2.5 Determination of the clash tolerance for sound profilin docking. Clash number distribution for profilin merged mDia1-FH1 conformers with no visible clashes (black), simple side chain contacts (green) and more than one clashes (blue) are shown in this figure. Visualization is made via MacPYMOL. The FH1 domain is colored in green and profilin in cyan. Samples of the merged conformers with a clash number of 25, 54, 120, and 251 are demonstrated above.
2.2.4 Formin FH1 domain opens up and elongates upon profilin binding

Radius of gyration ($R_{\text{gyr}}$) is an indicator of a protein structure’s compactness. The smaller the radius of gyration is, the tighter packing the protein has. Figure 2.6 A) displays $R_{\text{gyr}}$ distribution of the seven structure ensembles that were classified by RMSD threshold and clash tolerance. The curve representing the free mDia1-FH1 conformers lies on the left end with the smallest average $R_{\text{gyr}}$ among all groups. This indicates free conformers unavailable for profilin binding pack tighter than conformers bound with profilins. Also, it means conformers bound with profilins tend to open up compared to the free ones. Curves standing for conformers bound with N+1 profilins shift to the right of those bound with N profilins ($0 \leq N \leq 5$). Table 2.1 lists out average $R_{\text{gyr}}$ values of all seven groups. There is a consistent increase of no less than 2Å in the $R_{\text{gyr}}$ average value of the structural ensemble when one more profilin binding site becomes available each time. However, this is not observed for the structural ensemble without bound profilins. The increment of $R_{\text{gyr}}$ from free conformers to one profilin bound conformers is less. But such increment is still statistically significant (Table 2.2). These findings infer that mDia1-FH1 gradually extends its conformation as more sites bind to profilin.
Figure 2.6 Radius of gyration ($R_{\text{gyr}}$) A) and End-to-end distance (NCdist) B) distribution of conformers with 0 – 6 profilin binding sites. The curve corresponding to mDia1-FH1 structural ensembles with all sites free (black), one of the six sites bound with profilin (red), two of the six sites bound with profilin (green), three of the six sites bound with profilin (blue), four of the six sites bound with profilin (cyan), five of the six sites bound with profilin (purple) and all six sites are bound with profilin (yellow).
2.2.5 End-to-end distance

The end-to-end distance distributions of the seven structural ensembles showed the same right shifting trend (Figure 2.6 B). When a free binding site gets bound by profilin, its end-to-end distance is found to increase 5Å ~ 7Å on average in the structural ensembles bound with profilin (Table 2.1). Free conformers without bound profilins grow comparatively less in the end-to-end distance after binding to one profilin. The increase in the average values of end-to-end distances of the structural ensembles indicates mDia1-FH1 goes through a conformational elongation when free sites become bound by profilins. Although both $R_{\text{gyr}}$ and end-to-end distance increase on average when more sites in the FH1 domain are bound by profilins, these increasing trends of two properties are different. The average $R_{\text{gyr}}$ grows consistently for structural ensembles with one to six sites bound by profilins (Table 2.1). On the contrary, only a drastic increment happens in the average end-to-end distance of the structural ensemble when two free sites are bound by profilin, but with more sites are bound, such increment becomes decreased. In other words, every extra site bound by profilin contributes less to the end-to-end distance elongation.
<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>All sites</th>
<th>One site</th>
<th>Two sites</th>
<th>Three sites</th>
<th>Four sites</th>
<th>Five sites</th>
<th>Six sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Structures</td>
<td>4000000</td>
<td>1088661</td>
<td>1268351</td>
<td>1063272</td>
<td>459039</td>
<td>107178</td>
<td>12876</td>
<td>623</td>
</tr>
<tr>
<td>Mean $R_{gyr}$</td>
<td>26.79</td>
<td>25.15</td>
<td>25.58</td>
<td>27.8</td>
<td>30.09</td>
<td>32.44</td>
<td>34.76</td>
<td>36.83</td>
</tr>
<tr>
<td>Min $R_{gyr}$</td>
<td>12.39</td>
<td>12.39</td>
<td>12.79</td>
<td>13.38</td>
<td>14.43</td>
<td>15.04</td>
<td>18.23</td>
<td>24.14</td>
</tr>
<tr>
<td>Max $R_{gyr}$</td>
<td>55.61</td>
<td>55.13</td>
<td>51.74</td>
<td>52.45</td>
<td>54.99</td>
<td>55.61</td>
<td>55.15</td>
<td>54.56</td>
</tr>
<tr>
<td>$R_{gyr}$ increment</td>
<td>0.43</td>
<td>2.22</td>
<td>2.29</td>
<td>2.35</td>
<td>2.32</td>
<td>2.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean end-to-end dist</td>
<td>65.376</td>
<td>59.396</td>
<td>61.34</td>
<td>69.174</td>
<td>76.755</td>
<td>84.149</td>
<td>91.394</td>
<td>96.6</td>
</tr>
<tr>
<td>Max end-to-end dist</td>
<td>181.345</td>
<td>172.731</td>
<td>154.626</td>
<td>166.094</td>
<td>171.697</td>
<td>172.501</td>
<td>181.345</td>
<td>176.37</td>
</tr>
<tr>
<td>End-to-end dist increment</td>
<td>1.944</td>
<td>7.834</td>
<td>7.581</td>
<td>7.394</td>
<td>7.245</td>
<td>5.206</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: All: initial structural ensemble; All sites free: structural ensemble with no site available for docking; One site bound: structural ensemble with one site available for profilin docking; Two sites bound: structural ensemble with two sites available for profilin docking; Three sites bound: structural ensemble with three sites available for profilin docking; Four sites bound: structural ensemble with four sites available for profilin docking; Five sites bound: structural ensemble with five sites available for profilin docking; Six sites bound: structural ensemble with six sites available for profilin docking.
Table 2.2 Mann-Whitney-Wilcoxon Test (unpaired) for $R_{gyr}$.

<table>
<thead>
<tr>
<th></th>
<th>Six sites bound</th>
<th>Five sites bound</th>
<th>Four sites bound</th>
<th>Three sites bound</th>
<th>Two sites bound</th>
<th>One site bound</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>X</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Y</strong></td>
<td>Five sites bound</td>
<td>Four sites bound</td>
<td>Three sites bound</td>
<td>Two sites bound</td>
<td>One site</td>
<td>All sites free</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>$&lt; 2.2e-16$</td>
<td>$&lt; 2.2e-16$</td>
<td>$&lt; 2.2e-16$</td>
<td>$&lt; 2.2e-16$</td>
<td>$&lt; 2.2e-16$</td>
<td></td>
</tr>
</tbody>
</table>

Null hypothesis: mean (X) = mean (Y); Alternative hypothesis: mean (X) > mean (Y)

NOTE: X and Y are the $R_{gyr}$ distributions of corresponding structural ensembles listed in the first and second rows of the table. E.g. X = Six sites bound, Y = Five sites bound, this is to check whether the mean $R_{gyr}$ of structural ensemble with all six sites available for proline docking equals to the mean $R_{gyr}$ of structural ensemble with five sites available for proline docking. The p value is less than 2.2e-16 means the chance of accepting the null hypothesis is less than 2.2e-16.
Table 2.3 Mann-Whitney-Wilcoxon Test (unpaired) for single site distance distributions of conformers bound with only one profilin.

<table>
<thead>
<tr>
<th>Site B bound</th>
<th>Site C bound</th>
<th>Site D bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Dist C</td>
<td>Dist B</td>
</tr>
<tr>
<td>Y</td>
<td>Dist D/E/F/G</td>
<td>Dist D</td>
</tr>
<tr>
<td>P</td>
<td>&lt; 2.2e-16</td>
<td>&lt; 2.2e-16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site G bound</th>
<th>Site E bound</th>
<th>Site F bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Dist F</td>
<td>Dist E</td>
</tr>
<tr>
<td>Y</td>
<td>Dist B/C/D/E</td>
<td>Dist D</td>
</tr>
<tr>
<td>P</td>
<td>&lt; 2.2e-16</td>
<td>&lt; 2.2e-16</td>
</tr>
</tbody>
</table>

Null hypothesis: mean (X) = mean (Y); Alternative hypothesis: mean (X) > mean (Y)

NOTE: X and Y are the single site distance distributions of conformers bound with one proline. The header specifies the condition for X and Y. e.g., in the second column with header “Site B bound”, X is “Dist C” which represents the distribution of site C distance when its neighbor site B is bound with proline; Y is “Dist D/E/F/G”, which represent the distribution of site D distance, the distribution of site E distance, the distribution of site F distance or the distribution of site G distance when site B is bound with proline; p value for this group of X and Y are all less than 2.2e-16, so they are listed in one cell for simplicity. This means a less than 2.2e-16 chance of accepting the null hypothesis: the average distance of site C equals to the average distance of site D, E, F or G when B is proline bound.
2.2.6 The cooperativity of the neighbor sites closest to the bound site

Single site distances were measured for conformers with one site bound by profilin. Despite the obvious open conformation of the bound site compared to its free state, there are also changing trends among the other free sites affected by their neighbor site’s conformational behaviors. When a free site is next to a profilin bound site, it tends to extend a bit longer than other free sites. The free sites that reside next to the bound site could be viewed as the neighbors to the bound site. The cooperativity of the neighbor site is: when a site is bound with profilin, its neighbors elongate as well. Figure 2.7 exhibits the positive cooperativity of neighbor sites in the structural ensemble with one site bound by profilin. The six sites are named alphabetically from B to G. In the conformers with one site bound by profilin, the distance of site B is on average longer when its neighbor site C is bound with profilin. Similarly, distance of site C is generally longer when either its neighbor site D or site B is bound with profilin. The same observation can be found for the other four sites in Figure 2.7. Although it seems that the free site at C terminal next to the profilin bound sites elongates longer than the free site at N terminal, Wilcoxon signed-rank test listed in Table 2.3 does not show any statistical significance for this seemingly imbalanced neighbor effect except for site E. For sites C, D and F, either of their neighbors on both ends have the same effect: they extend to the same length generally. Thereafter, a profilin bound site would induce both of its neighbor sites to stretch to a similar length. As to site E, it elongates less when its downstream neighbor site F is bound by profilin than when its upstream neighbor site D is bound by profilin. In the cellular context, the C terminal of FH1 domain is linked to the FH2 domain and its N terminal
attaches to the plasma membrane via the diaphanous inhibitory domain. This imbalanced neighbor site effect observed on site E might imply that it is easier to extend its conformation through the free sites near membrane side while the site near FH2 domain is bound rather than the other way around.
Figure 2.7 Comparison of the distance distributions of single sites. Site B and site G are the sites on the two terminals. All four sites in the middle tend to elongate when either of their neighbor sites is bound by profilin. E.g. Site C Distance distribution: The distance of site C when its upstream neighbor site B is profilin bound (red) and the distance of site C when its downstream neighbor site D is profilin bound (blue) are larger than (shift to the right of) the distances of site C when site E (cyan), site F (magenta) or site G (yellow) is profilin bound.


2.3 Discussion

2.3.1 Radius of gyration

A disordered protein lacks a stable tertiary structure and has higher conformational flexibility. Due to highly biased amino acid composition, disordered proteins generally cannot form a tightly packed hydrophobic core. As a consequence, the radius of gyration, \( R_{\text{gyr}} \), of a disordered protein is relatively larger than that of an ordered protein with the same amino acid length. As a global structural property of proteins, \( R_{\text{gyr}} \) indicates the degree of structural compactness and serves as a characteristic measurement for classifying proteins. A statistical analysis on four major protein domains revealed that each domain has its own particular value of \( R_{\text{gyr}} \) [318]. For disordered polypeptide chains, \( R_{\text{gyr}} \) is currently the main property describing their dimensions. Small-angle x-ray scattering (SAXS) could directly measure \( R_{\text{gyr}} \) and is widely used in protein characterization. Globular and non-globular conformations are found with large differences in their SAXS patterns [34]. SAXS is especially suitable to study IDPs that cannot crystalize. Recent developments on the Ensemble Optimization Method enable SAXS to quantitatively characterize IDPs with more details by fitting multiple scattering patterns from conformational ensembles [171]. The protein sample required for SAXS experiment is typically 1-3 mg and should be highly pure. For low yielded proteins, it is usually painful to prepare enough samples with high quality. The protein samples need to remain soluble at high concentration in case that the aggregation will interfere with the scattering data. Therefore, insoluble proteins or proteins with low solubility are beyond the study scope of SAXS. Molecular simulation can save us from preparing these protein...
samples. Seventeen proteins were successfully reconstructed using TraDES when it was released [313]. The minimum $R_{gyr}$ value of the TraDES generated conformers was at the same order of a native structure. The study on disordered drkN SH3 domain reported the $R_{gyr}$ value of experimentally refined structures, and it matches the calculation from TraDES-sampled conformers [187]. TraDES, therefore, can help us measure a reliable $R_{gyr}$ of protein structures when it is difficult to perform SAXS experiments. It evades the aggregation problem we normally encounter for protein samples and delivers a chance to study IDPs that hardly express. FH1 is relatively shorter compared to the FH2 domain and contains many prolines that increase its insolubility. Expression of the FH1 domain alone is impossible so past studies on FH1 always expressed FH1 and FH2 together. TraDES could do conformational sampling on the FH1 domain without the interference of the FH2 domain.

In the mDia1-FH1 domain simulation study, the $R_{gyr}$ of FH1 structural ensemble is observed to grow larger on average with every one more site bound by profilin (Figure2.6A). It indicates that the bound profilins tend to induce the FH1 domain to adopt a more open conformation. Figure 2.8 shows the comparison between two structural ensembles with and without profilin binding capability. All conformers were aligned at the same profilin binding site. The ensemble that is capable of binding profilin (A) vacates the space for profilin. In contrast, the structural ensemble not available for binding profilin does not accommodate profilin (B) and spans randomly in all directions taking space near the binding site. From the comparison, we could observe that the bound profilin serves as a spatial constraint imposed on the disordered profilin binding site in the FH1 domain. This spatial constraint induces the
conformation of the FH1 domain to elongate and expels the rest of the peptide to scatter in other directions. The same effect is also detected in the intracellular domain of LRP6 protein in the initiation of Wnt canonical signaling pathway [190]. The profilin binding sites in the FH1 domain are short and arranged in tandem. One bound profilin may affect the conformational space for its closest neighbor sites on both sides in the FH1 domain. This could be due to the cooperativity of neighbor sites we observed in Figure 2.7. While profilin-actins bind to these sites in FH1 in vivo, the excluded volume effect would be more obvious for larger size of the actins.

Figure 2.8 Comparison of mDia1-FH1 conformers with (A) and without (B) profilin binding capability. All conformers were aligned at the same profilin binding site (resi 26-30). Conformers capable of binding profilin vacate the space for profilin (A). In contrast, the structural ensemble not available for binding profilin does not accommodate profilin (B) and spans randomly in all directions, and the binding site was crowded with its probabilistic residue distribution.
2.3.2 End-to-end distance

The end-to-end distance is a measure of the distance between the two terminal $\alpha$-carbon atoms of two amino acids in a protein structure. For tightly packed peptide chains, the end-to-end distance would be small, but for ensembles of IDPs, the end-to-end distance have wide distributions. The statistical model of random coils is often employed to describe disordered proteins. According to Flory the average end-to-end distance of a random coil chain has a power law depending on the number of composing units. It is proportional to $N^v$, where $N$ is the number of composing units and $v$ is a scaling factor. A fully disordered chain is predicted to have $v$ equal to 0.6 [319, 320]. However, this theory is not true for long denatured polypeptides [321, 322]. The end-to-end distance is another commonly used measurement for characterizing disordered proteins. FRET is frequently used to determine the end-to-end distance of IDPs for its high sensitivity and ideal time resolution. Both single-molecule FRET and ensemble FRET have been used to study IDPs. The end-to-end distance determination of the N terminal domain of tumor suppressor p53 is an example of using these two methods to study the conformational dynamics [174]. As a common concern for experimentalists using protein samples in solution, the low solubility of sample peptides will limit the application of FRET. TraDES has been used to reevaluate the end-to-end distance of bovine pancreatic trypsin and the results fitted well with the FRET result from Haas’ group [186]. It is proved to be an alternative to obtain the end-to-end distance distribution of proteins. The mDia1 FH1 domain is only 162-amino-acid long and is predicted to be totally
insoluble [323]. In silico conformational sampling is extremely useful to study the disordered mDia1 FH1 domain.

Figure 2.9 Illustration of excluded volume effect on the FH1 domain. PDB structure of mouse profilin 2a (Green) with a profilin binding site (Orange) from the mDia1 FH1 domain (pdb id: 2V8F) is positioned along the Z-axis. Formin FH2 domain (Cyan, pdb id: 1UX4) is under the x-y plane.

From the end-to-end distribution of seven structure ensembles of mDia1-FH1 (Figure 2.6B) we could deduce that the entire peptide grows longer with every site bound by profilin. This could attribute to the excluded volume effect. Profilin 2a is a 15 kDa protein composed of 140 amino acids, more than ten times larger in size compared to one profilin binding site in formins that is only 12-amino-acid long. Hence, profilin serves as a spatial constraint confining the available conformational space of FH1 along the Z-axis (Figure 2.9). Upon binding, profilin blocks the originally available space of binding site in the FH1 domain, forcing it to stretch in the unblocked direction. In vivo, one end of FH1 domain is linked to the structured FH2
domain that associates with the barbed end of growing actin filaments and the other end of FH1 is linked to the plasma membrane via other regulatory domains of formins. Because FH2 is larger than FH1 in size, there is little chance that FH1 could penetrate the x-y plane imposed by the FH2 domain and extend to the minus side of Z-axis. Formin FH1 domain usually contains multiple binding sites to be bound by profilins, which can exclude FH1’s volume along the z-axis significantly. With one end of FH1 fixed by FH2-associated actin filament, the excluded volume effect could only drive FH1 to extend towards the plus side of Z-axis, which represents the membrane \textit{in vivo}.

Contrary to Flory’s “isolated pair hypothesis” that treats each \(\varphi, \psi\) pair of the peptide backbone as sterically insensitive to their neighbors, Pappu \textit{et al.} found that local steric effects could restrict the accessible conformational space significantly. The bound profilin opens up the profilin binding site in the FH1 domain and each residue inside the binding site loses a great part of conformational freedom so that these steric effects further restrict the accessible conformational space of their closest neighbor sites.

2.3.3 Addition of profilin bound site reduces the contribution of individual binding site in end-to-end distance expansion

It has been reported that each additional profilin bound site contributes less to the actin filament assembly rate as the total number of binding sites in the FH1 domain increases [6]. In our simulation study, we also found that there is a declining trend in the elongation degree of FH1’s conformation when every additional profilin bound site is docked according to the average end-to-end distance of all seven FH1 structural ensembles. This suggests a
correlation between the conformational elongation of FH1 and the activity of formin motor in actin polymerization. An additional profilin bound site introduces a smaller end-to-end distance increment to the FH1 domain, which adopts an extended conformation at a slower speed. This would reduce the chance for profilin’s quick access to the binding site. Hence, such situation may slow down profilin’s accumulation and in turn reduce the velocity of the free actin monomers to be added on the actin filament.

2.3.4 Relationship between conformational elongation and formin motor function

Formin FH1 domain adopts a random coiled form in nature since it is intrinsically disordered. This can also be inferred from the $R_{\text{gyr}}$ and end-to-end distance distributions in which FH1 samples a large conformational space. Every bound profilin pushes the loosely packed coil to a more extended form according to the $R_{\text{gyr}}$ distributions (Figure 2.6A). The end-to-end distance distributions also show that the extension stretches in a longitudinal way rather than curling up into a ball (Figure 2.6B). The scatter plot of $R_{\text{gyr}}$ versus end-to-end distance shows this linear relationship between the over all shape of FH1 and its end-to-end distances of individual binding sites (Figure 2.10).
Figure 2.10 Scatter plot of $R_{gyr}$ vs end-to-end distance of FH1 structural ensembles. Colored dot represent structural ensembles with all six sites free from profilin (pink), one site bound by profilin (orange), two sites bound by profilin (green), three sites bound by profilin (cyan), four sites bound by profilin (blue), five sites bound by profilin (purple) and all six sites bound by profilin (magenta).

The increment in the average end-to-end distance with one additional site bound by profilin indicates that the disordered FH1 domain transits from a random coil to an elongated form upon binding with profilins, and the elongation extent depends on the number of bound profilins in the FH1 domain. Prior experiments with formin Bni1p constructs composed of one to eight profilin-binding polyproline tracks showed similar dependence on the number of profilin binding polyproline tracks for actin filaments polymerization [6]. It has been reported that each additional profilin bound site contributed less to the actin filament assembly rate as the total number of
binding sites in the FH1 domain increases [6]. Interestingly, we found that there is a decline in the elongation degree of FH1’s conformation when every additional profilin bound site is docked, according to the average end-to-end distance of all seven FH1 structural ensembles. This suggests a coupling of the conformational elongation of FH1 and the acceleration activity of formin motor in actin polymerization. Hence, we propose a cooperative “jack” model of random coil-to-elongation transition of the FH1 domain to explain the formin motor behavior in the presence of profilins (Figure 2.11). In the cellular context, Rho protein activates and recruits formins to the plasma membrane leaving diaphanous inhibitory domain (DID) associated with the plasma membrane such that FH1 spans between actin filament barbed end and the plasma membrane. Free profilin-actins associate with the profilin binding sites in FH1 creating the excluded volume effect, which turns FH1 coiled conformation into a regular “stair case” stacked with the bound profilins. Kursula’s group reported profilin dimerisation upon binding to FH1 peptide [310]. The dimerization could further stiffen the staircase shaped FH1. Due to the conformational elongation of FH1, the plasma membrane is being pushed forward. After unloading profilin-actins to profilin binding sites, the FH1 domain delivers these bound profilin-actins to the barbed end of actin filaments and then returns to its random coiled state. The FH2 domain elevates the working base of the FH1 domain for a new session of elongation after inserting actins delivered by the FH1 domain.

In the formin dimer elasticity model, a pulling force on the FH2 domain is proposed to lower the critical actin concentration and thus increases the barbed end polymerization. FH1 in the cooperative “jack” model we
propose here would elongate and push forward the membrane due to the volume exclusion of profilin-actins and the dimerization of profilins that bind to adjacent profilin binding sites, so that there would be a pulling force applied on the FH2 while the profilin-actins stacked FH1 pushing the membrane. Our simulation results indicate that the coil-to-elongation transition of the FH1 domain provides the structural basis for the formin motor function and adds a new perspective to previous processive model of FH2 and the end-tracking model for formin general mechanism from the FH1 side. The energy source for the “jack” model could derive from the binding energy released when the FH1 domain associates with profilins. Experiments using optical tweezers could measure the elongation of the formin-FH1 while binding to multiple profilins and magnetic tweezers could measure the forces generated during binding.
Figure 2.11 The proposed cooperative “jack model” for formin FH1 in the actin filament elongation. Random coil status: the FH1 domain spreads loosely between the membrane and the FH2 domain. Jagged “stair case” status: a group of profilin-actins stack in by binding to the FH1 domain and push the membrane outward. One cycle of formin motor movement starts from the random coiled state represented by the most left complex, then profilin-actins induce the formin motor to a stacked “stair case” represented by the 2nd, 3rd and 4th complex and the membrane is progressively pushed forward at a distance of $\Delta R_i (\Delta R_1 > \Delta R_2 > \Delta R_3)$. After releasing the bound profilin-actins to the FH2 associated barbed end, FH1 domain restores to its random coiled state again with a higher “stair case” base presented by the most right complex. For simplicity purpose, only two profilins binding to FH1 are presented.
2.4 Methods

2.4.1 Generation of structure ensembles of mDia1-FH1 domain

The FH1 domain of mDia1 is composed of thirteen polyproline tracks, six out of which are linked in tandem and were used in this study for simulation. The sequence corresponding to the six profilin binding sites, a 73-residue long FH1 segment of mDia1 was retrieved from UniProtKB shown as following:

>sp|O08808|DIAP1_MOUSE Protein diaphanous homolog 1 OS=Mus musculus GN=Diaph1 PE=1 SV=1|635-707
IPPPPPLPGVASIPPPPPLPGATAIPPPPPLPGATAIPPPPPLPGATAIPPPPPLPGGTPGIPPPP
PLPGSVDPPPPLPGGPG

This segment sequence was used as the input to generate mDia1-FH1 conformers using VISTRAJ and TRADES from TraDES package [186]. VISTRAJ firstly generated a trajectory distribution file containing the probability distribution of $\phi/\psi$ angles in the conformational space for each residue. The five consecutive prolines in each profilin binding site were confined to the $\phi/\psi$ angles observed from the crystal structure [PDB: 2V8F] as follows:

<table>
<thead>
<tr>
<th>Residue</th>
<th>Phi $\phi$ (°)</th>
<th>Psi $\psi$ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro14</td>
<td>-74.8</td>
<td>150.5</td>
</tr>
<tr>
<td>Pro15</td>
<td>-60.4</td>
<td>140.0</td>
</tr>
<tr>
<td>Pro16</td>
<td>-76.0</td>
<td>164.1</td>
</tr>
<tr>
<td>Pro17</td>
<td>-75.9</td>
<td>160.1</td>
</tr>
<tr>
<td>Pro18</td>
<td>-65.0</td>
<td>148.4</td>
</tr>
</tbody>
</table>
This crystal structure contains mouse profilin 2a (Chain A and B) and two profilin binding sites in the mDia1 FH1 domain (Chain C), which has exactly the same sequence as the profilin binding site we used for this study. Since FH1 is regarded as a disordered region, its trajectory distribution output from VISTRAJ was generated under the “coil” mode. This method samples the probability distribution of $\psi/\phi$ angles for each residue according to a non-redundant PDB data set composed of coils as secondary structure. TRADES then used the trajectory distribution output file to generate off-lattice unbound all-atom protein structures. It samples the Ramachandran space with probability distribution of each residue contained in the trajectory distribution file. The workflow is illustrated in Figure 2.12.
2.4.2 Alignment of [PDB:2V8F] to mDia1-FH1 conformers

The residues 9-13 in [PDB:2V8F] Chain C are five consecutive proline residues in mDia1-FH1 that bind to profilin and they were aligned with each of the six binding sites in mDia1-FH1. SALIGN is a program in TraDES package that takes two protein structures with the same length and aligns the second structure to the first one by superimposing them using the Singular Value Decomposition (SVD) method. The output of this program is a transformed structure of the second structure. Here SALIGN aligned the backbone atoms of [PDB:2V8F] Chain C onto those of each profilin binding site in mDia1-FH1 conformers. This program also allows all-atom and α-carbon alignment method. The root mean square deviation (RMSD) between the superimposed structures measures their divergence from each other. It is used as an indicator of the quality of alignment and is calculated as:

\[
RMSD = \sqrt{\frac{1}{N} \sum_{k=1}^{N} |R_k^i - R_k^j|^2}
\]  \hspace{1cm} (2.1)

where \(R_k^i\) represents the position of the \(k\)-th α-carbon in the main chain of structure \(i\).

2.4.3 Merging and checking clashes

Chain B of [PDB: 2V8F] is profilin 2a that binds to prolines in mDia1-FH1. After the structural alignment step run by SALIGN, STRMERGE, another program in TraDES package, was used to merge Chain B of [PDB:2V8F] with mDia1-FH1 conformers to a single .val file. It is necessary to check 1) the steric clashes between profilin 2a and mDia1-FH1 and 2) the
clashes between profilins that bind to different sites. CRASHCHECK from the TraDES package counts all atom-atom clashes inside a structure file. A profilin binding site was considered as bound with profilin when no clashes were reported from the merged profilin-FH1 complex. All atoms in each chain of the merged structures went through the clash-checking step. The atom-atom pairs with distance less than the Van der Waal distances of two atoms were recorded. The latest TraDES package can be found at ftp://ftp.blueprint.org/pub/TraDES/.

2.4.4 Measurement of $R_{\text{gyr}}$ and end-to-end distances

Radius of gyration is the root mean square distance of each atom from the center of the structure. It is used in polymer physics for describing the dimensions of a polymer chain. It is defined as:

$$R_{\text{gyr}}^2 = \frac{1}{N} \sum_{k=1}^{N} (r_k - r_{\text{mean}})^2$$  \hspace{1cm} (2.2)

Where $r_k$ represents the position of each atom in the structure, $r_{\text{mean}}$ stands for the average position of all atoms, which is the gravity center of the structure. Radius of gyration indicates the openness of a structure. Conformers generated via TraDES contain the coordinates of all atoms including hydrogen, which are used to calculate the radius of gyration. $R_{\text{gyr}}$ distributions of the structural ensembles without any profilin binding and with one to six sites bound by profilins were calculated to determine the overall openness of the structure.
The end-to-end distance, $R$, measures the distance between the $\alpha$-carbons of the N-terminal residue and the C-terminal residue. It is the sum of all covalent bonds, each of which is taken as a vector, $b_i$. Thus,

$$R = \sum_{i=1}^{n} b_i$$

(2.3)

The distance $R$ varies with different residues in between and can be either positive or negative. To avoid the effect of the sign, $\sqrt{R^2}$ is used to represent the end-to-end distance here. The single binding site distance is calculated the same way with two terminal residues substituted by desired binding site boarder residues.

The end-to-end distances include the distances between I1 and G73. The distance of site B, site C, site D, site E, site F and site G were calculated respectively from I1 to S12, from I13 to A24, from I25 to A36, from I37 to G48, from I49 to V61 and from V62 to G73.
Chapter 3
Thermodynamic Study of the FH1 Domain and Profilin Binding

3.1 Background

The first description of titration calorimetry is more than 40 years ago recording the simultaneous determination of $K$ and $\Delta H$ [325]. This method was originally used for determination of equilibrium constants for metal-ligand complexes [326]. The calorimetric instrument available at that time was only able to measure the equilibrium no more than $10^{-4}-10^{-5} \text{ M}^{-1}$ [327]. For association constants beyond this limit, more dilute solutions are required but the instrument at that time was not sensitive enough to make accurate measurement. The introduction of titration calorimetry into biological systems was first published in 1978 followed by a more detailed discussion on each of the steps in fulfilling a typical calorimetry experiment [329]. After ten years development, this technique was finally commercialized for biological equilibrium studies. MicroCal, the first isothermal titration calorimetry (ITC) apparatus commercially available, was advertised as a device to determine binding constants in minutes. ITC nowadays is used widely in characterizing the thermodynamics of reactions in biological systems. With the development of the technology, modern ITC apparatus is capable of measuring heat as low as $0.1 \mu\text{cal}$ and determining $K$ as large as $10^{-8}-10^{-9} \text{ M}^{-1}$.

Most chemical reactions are accompanied by a change in heat. The endothermic process is indicated by the absorption of heat and the exothermic process is accompanied by the release of heat. A measure of the heat absorbed from the environment for an endothermic reaction or heat unleashed to the environment for an exothermic reaction equals to the amount of heat exchange
occurred during the reaction. The rate of heat exchange with the environment simply equals to the rate of the reaction. ITC is an ideal method to measure either how much heat exchange occurred during a reaction or the rate of the reaction. Comparing to traditional optical methods ITC has the advantage to detect spectroscopically silent reactants. It does not require a fluorescent tag for the reactant or translucent solution for measurement. A wide range of biologically pertinent conditions could suffice for a good measure. ITC is a highly useful technique that measures the binding energetics of protein-ligand binding. This technique can precisely determine a group of thermodynamic parameters, which include enthalpy, entropy, heat capacity changes and the Gibbs energy, associated with binding. Also, it is usually used in obtaining binding constants for protein-ligand binding.

In the case of determining binding constants, heat is treated as simply a phenomenological signal of the saturation extent of the ligand-binding site. Measurement of K depends on the intrinsic factors of the system such as instrumental stability, systematic artifacts and the availability of obtainable device as well as the system-specific factors such as the magnitude of K, kinetic thermal and pH stability. These various factors finally decide the appropriateness of methods for a given experiment. Due to the generality of heat changes during chemical reactions, microcalorimetry has gained the popularity in measuring the binding constants of protein-ligand complexes. The value of ΔH in records for a protein single site binding with a ligand molecule spans from +4 kcal/mol to -25 kcal/mol. Changing the nature of the buffer may escalate the enthalpic signal by about 12 kcal/mol for ligand-binding reactions that involve proton ionization process. Calorimetric studies
with the sensitivity available now are able to measure \( K \) in micromolar range no matter it is a negative or positive signal. Another parameter frequently measured for ligand-binding reactions during calorimetry is \( \Delta H \). It can be achieved with two orders of magnitude in precision for titration calorimetry.

The binding of profilin and the FH1 domain concentrates free profilin-actins for actin filament barbed end assembly and contributes to the formin motor function. It is still unclear whether the interaction between these two polymers makes energetic contributions to the motor function. As the FH1-profilin binding is governed by thermodynamics, quantitative estimates of the free energy of the binding are essential to verify its energetic role for formin motors. To elucidate the structural features of the FH1-profilin interactions, investigation on the thermodynamics of binding of mDia1-FH1 and profilin 2a is performed. ITC is used in this chapter to study FH1-profilin interactions.

3.2 Calorimetry Theory

Calorimeter is the instrument to measure the heat involved in a reaction. The first ice-calorimeter Antoine Lavoisier used to determine the heat produced by the guinea pig marked the beginning of thermochemistry. There are generally three methods in measuring the heat change for a calorimeter [327]. For a temperature change instrument, the heat change caused by the reaction will be represented by a change of the cell temperature measured in the calorimeter. The output is the calorimeter cell temperature as a function of time. For a power compensation instrument, the temperature of the measured cell is kept constant by a control heater. For example, in an exothermic reaction, the power supplied to the control heater would decease when the heat released from the reaction is sensed. This heat feedback
mechanism makes the combined contribution of the control heater and the heat from the reaction constant in the measuring cell. The output is the power applied to the control heater to keep the measurement cell temperature constant as a function of time. For a heat conduction calorimeter, the measurement cell temperature is kept constant by heat flow sensors that are linked to a heat sink secured to a constant temperature. The output is the voltage as a function of the small temperature change developed by the heat flow sensors. ITC instruments employ a power compensation method in measuring the heat change in reactions. The temperature difference between the control cell and the sample cell due to the reaction is measured and detected by changing the power applied to the control heater. Each injection of the ligand solution from the syringe forms a certain amount of protein-ligand complexes and causes the release (exothermic process) or absorption (endothermic process) of heat.

To initiate a reaction in the calorimetry measurement cell, three methods can be used: batch, titration and flow method [327]. In a batch microcalorimeter, the solutions of protein and ligand are separately incubated to reach thermal equilibrium. After that the two identical-solvent-based solutions are mechanically mixed. Due to the long equilibration time length caused by the slow response of the instrument, it usually takes several months to obtain a complete binding curve. In a flow microcalorimeter the solutions are mixed in a thermostatic sensing chamber. Compared to the 2.5hr long equilibrium time required by a batch microcalorimeter, flow microcalorimeter reduced this process to around 15 minutes. This improvement shortened the working time to get a complete binding curve to one week. However, a change
in the operating temperature needed as long as three days to reach equilibration. The major drawback of this method is the requirement of large quantities of reactants to keep a stable baseline due to the constant flow nature of the system. Modern ITC instruments use the titration mode with a series of injections made at the pre-set time intervals to complete the reaction. Titration calorimeter works in a similar way as batch calorimeter when mixing two solutions. Both introduce a small portion of one reactant into the other and mix the two solutions in a stepwise manner. The heat fluctuation from the mixture is measured and translated into an integrated enthalpy at each step and would finally produce a complete enthalpic titration curve. Initially this method was held back by its low sensitivity, which requires a relative high concentration of the reactant. Hence, any attempt to determine the dissociation constant less than 500 \( \mu \text{M} \) is impractical. An internally located syringe later rescued the degree of sensitivity, but it still required a long initial incubation period. Improvements made till now could guarantee a faster and more sensitive measurement for a complete binding curve. The research development in this area is booming, and titration calorimeter almost has substituted microcalorimeter in a nonspecific context.

### 3.3 Thermodynamics of Binding

A typical binding interaction involves a ligand and a receptor with vacant binding sites. The binding equilibrium of ligand L binding non-covalently to protein receptor P can be represented by:

\[
P + L \leftrightarrow P \cdot L
\]  

(3.1)
Where $P\cdot L$ is the non-covalent protein-ligand complex. The binding constant $K_a$ for a single binding site case is given by equation:

$$K_a = \frac{[P\cdot L]}{[P][L]} \quad (3.2)$$

Where $[P]$ indicates concentration of protein receptor $P$. The free energy of binding $\Delta G$ can be written as follows:

$$\Delta G = -RT \ln K_a \quad (3.3)$$

In equation (3.3) $\Delta G$ is the change of free energy when one mole of ligand binds to one mole of protein under standard states (Figure 3.1), $R$ is the universal gas constant, $T$ is the absolute temperature in Kelvin. $\Delta G$ is an indication of the strength of binding between the ligand and the protein. It ranges from around -50 kJ mole$^{-1}$ in the tightest binding interactions to around -17 kJ mole$^{-1}$ in the weakest ones. Compared to binding constant $K_a$, the dissociation constant $K_d$, which is the reciprocal of $K_a$, is more commonly used to imply the binding affinity. It equals the concentration of free ligand at which half of the total protein molecules are associated with ligands at equilibrium. For a natural binding interaction, the dissociation constant is lower than the physiological concentration of the ligand by a factor of 10 to 100. A picomolar to nanomolar range of $K_d$ reveals tight binding; while, a value in millimolar suggests a weak binding. When expressed with the enthalpy change $\Delta H$ and entropy change $\Delta S$ of binding, the free energy of binding can be described by:
\[ \Delta G = \Delta H - T\Delta S \]  

As indicated from equation (3.4) the binding affinity is determined by enthalpy and entropy. The binding enthalpy change represents the sum of energy change after breaking old bonds and forming new ones. The entropy change represents the sum of changes in solvation entropy and conformational entropy.

Figure 3.1 Schematic diagram of the Gibbs energy change during the interaction of ligand and protein.
3.4 Materials and Methods

3.4.1 Peptide and protein

The mDia1-FH1 peptide sequence used for isothermal titration is IPPPPLPGVASIPPPPPLPG synthesized from GL Biochem (Shanghai). Mouse 2a profilin was expressed in E. coli BL21 cells from plasmid pET28 (profilin sequence cloned from pMK-T vector, purchased from geneArt). Transformed cells were incubated overnight in LB broth with 100 mg/ml Ampicilin at 37 °C and shaken at 250 rpm. The cells grew until OD600 reached around 0.6 and 0.4 mM IPTG was added for induction. The cell pellet was resuspended in Buffer I (5 mM EGTA, 0.1 mM EDTA, 50 mM Tris-HCl (PH = 7.3), 50 mM KCl, 10 mM DTT, 8 M urea) followed by sonication. Impurities were cleared by centrifugation at 15,000g for 1h. The supernatant was extensively dialyzed against buffer without urea (Buffer II) after overnight dialyzation against buffer with 4 M urea. The supernatant after another centrifugation at 15,000g was loaded onto a poly-L-proline column (Poly-L-proline and Cyanogen Bromide-Activated Matrices, Aldrich) constructed in a standard procedure [330]. The profilin was eluted from the column with Buffer I after exhaustive washing with 1M urea. The eluted profilin was dialyzed again by following the above-mentioned steps.

3.4.2 Instrumentation

The most widely used ITC instruments are Nano ITC from TA and VP-ITC from Microcal. Based on isothermal power compensation they operate in similar principles. There are two identical cells, sample and reference, made by highly efficient thermal conducting materials in an
adiabatic shield (Figure 3.2). The shield is cooled by a circulating water bath. The temperature difference between the reference cell and the shield is continuously monitored by sensitive thermocouple circuits to keep a constant temperature. A feedback control system tries to keep the temperature difference as close to zero as possible and outputs the measured signal. One of the reactants is contained in the sample cell; while buffer or water without reactants is placed in the reference cell. The other reactant is in the injection syringe as the titrant reactant to be added to the sample cell. Before the titration process, the reference cell is given a constant power to form the baseline signal. During each injection step, the heat is released or taken up by forming new macromolecular complex. The power supplied to the sample cell correspondingly decreases or increase to maintain the temperature as the same as the reference cell. Exothermic reactions produce negative signals as shown in Figure 3.2 and endothermic reactions result in positive voltage signals. The heat change in a reaction is proportional to the percentage of bound ligand, so that the initial concentrations of both reactants are of extreme importance. For the initial injections, most of the reactants form macromolecular complexes, resulting in a large signal. With more injections of the titrant, the reaction gradually saturates and a less heat change is involved.
3.4.3 Isothermal titration calorimetry

The calorimetric titration of mouse 2a profilin with mDia1-FH1 was carried out at 25 °C in a VP-ITC isothermal titration calorimeter (Microcal). Before titration, all samples were prepared in 10 mM HEPES, pH = 7.5 and degassed for 10 min using vacuum pump. The titration consisted of 27 injections of 10 µl each of a 2 mM FH1 solution following an initial injection of 2 µl. The syringe was loaded with 500 µl FH1 and was injected into 0.2 mM profilin that is contained in the sample cell with a volume of 2 ml. The duration of each injection is 20 second (sec) separating by a 220 sec space. The stirring speed in the sample cell was 500 rpm.
3.5 Results and Discussion

Thermodynamic characterization of the interaction of mouse 2a profilin with mDia1-FH1 was performed using isothermal titration calorimetry. Figure 3.3 shows the raw data of heat for each injection and the nonlinear fit to a two-site binding model. This was obtained after adjusting the raw data with control titrations and removing the first injection data point. The red line fitting to the data points represents an ideal heat profile produced by a 1:2 complex. The best-fit value for each parameter is listed in the box at the bottom right corner of the graph. The two-site reaction model was fitted using Origin ITC software (MicroCal).
Figure 3.3 Titration of mouse 2a profilin bound with mDia1-FH1. Top panel represents the raw ITC from 27 equal injections of mDia1 FH1 into profilin. Bottom panel represents the nonlinear regression fit of the raw data after the subtraction of blank experiments and the deletion of the first data point of 2 µl initial injection. The experiment was performed in 10 mM HEPES, pH = 7.5, at 25 °C. The concentrations of reactants were 0.2 mM profilin (in cell) and 2 mM FH1 (in syringe).
The titration analysis proved that profilin-FH1 binding is an
exothermic reaction. No additional heat is needed when titrating profilins to
saturation. The FH1 peptide used in this study contains two profilin binding
sites. The number of binding site n = 0.47 matches the molar ratio of FH1
versus profilin which correspond to a 1:2 stoichiometry. The dissociation
constant $K_d = 13.3 \ \mu\text{M}$, which is in the range detected for profilin and proline-
rich peptides [331]. The binding of FH1 and profilin is enthalpically favored
($\Delta H = -10.5 \ \text{kcal/mol}$) with unfavorable entropy ($\Delta S = -12.96 \ \text{cal mol}^{-1}\text{deg}^{-1}$)
but to a lesser extent comparing to L-Pro$_8$ (eight consecutive L-prolines) and
L-P$_5$AP$_5$ (two segments of five consecutive L-prolines with an Alanine). The
entropy change for profilin binding to L-Pro$_8$ and L-P$_5$AP$_5$ are -142 cal mol$^{-1}$
$\text{deg}^{-1}$ and -62.24 cal mol$^{-1}$deg$^{-1}$ [331]. This indicates the entropy cost of
binding L-Pro$_8$ is larger than that of IP$_5$L, which is the profilin binding site
pattern seen in chapter 2. The enthalpy change of profilin binding formin-FH1
sums up the energy change in breaking solvent hydrogen bonds around both
formin-FH1 peptide and nonpolar groups in the profilin binding site and
energy changes in forming intermolecular interactions upon profilin-FH1
binding. The favorable enthalpy change of profilin-FH1 binding is likely to be
derived from hydrogen bonding and hydrophobic interactions. The negative
entropy change renders the binding of profilin-FH1 unfavorable. The
contribution of overall entropy change is the sum of configurational entropy
change of the peptide and the entropy change of the solvent [332]. Recall the
results we obtained in Chapter 2 that formin-FH1 undergoes a coil-to-
elongation transition upon profilin binding. Profilin-FH1 binding lowers the
disorderliness of formin-FH1, which is likely to play a major contributory role in the negative entropy change.

Poly-L-prolines composed of ten or more residues are found to bind profilin with similar affinities and shorter L-proline homopolymers with dramatically lower affinities according to FRET data acquired by Petrella in 1996 [331]. The profilin binding site in formin-FH1 possesses only five consecutive prolines but shows a higher binding affinity than L-Pro_{11} for profilin, as estimated by Metzler et al [317]. This comparison reveals the important roles of the isoleucine and leucine anchors at the two ends of polyproline in binding profilin. The favorable enthalpy change in FH1-profilin binding denotes a more stable complex formation. Our experiment on mDia1-FH1 and profilin shows the exothermal nature of their binding. The energy released could participate in the pushing process of random-coil shaped FH1 into an extended form when profilin binds to it. The IP_{3}L pattern is shown to be more stable in binding profilin after comparing the dissociation constant and entropy cost with the previously determined data of poly-L-proline peptides with at least five consecutive prolines. Non-proline residues may have a more significant role in facilitating profilin binding than we used to think.
Chapter 4
Terbium Chelate as a Luminescent Probe for Studying Formin FH1 Conformational Change

4.1 Background

Fluorescence resonance energy transfer (FRET) is frequently applied to measure molecular distances and is common in detecting conformational changes. It involves two fluorophores, one as donor and the other as acceptor. Energy may transfer from excited state donor to acceptor and the efficiency is used to determine the distance between the donor and acceptor. This method enables sensitive detection of small distance within 100 Å. Though it is a feasible method to determine the end-to-end distance of IDPs, it has limitations. The short donor lifetime (usually in 10^{-9} scale) is difficult to measure with accuracy. The sensitized emission signal to background ratio is low because of the direct excitation of the acceptor [333]. The orientation between the donor and acceptor is hard to determine, which directly affects the calculation of energy transfer.

Ever since the lanthanide-binding tags (LBT) have been developed to sensitize the metal ion luminescence, lanthanide has obtained a popularity serving as the donor in FRET. Inspired by a variety of natural terbium binding proteins that enhance terbium luminescence through quenching an aromatic residue near the binding site, LBT is gradually developed into successful donors with long lifetime (millisecond), sharp emission spectra and unpolarized luminescence [334, 335]. For the nature of energy transfer from lanthanide is different from fluorescence that arises from a singlet to singlet transition, lanthanide emission is more properly called luminescence
resonance energy transfer (LRET). The millisecond luminescence lifetime donor enables collecting the emission data after an initial delay, which eliminates background luminescence derived from the excitation of acceptor.

In this chapter, we used a LBT fused mDia1-FH1 peptide segment with two profilin binding sites and labeled with AF555 as acceptor to measure the conformational changes of FH1 under the effect of profilin binding. From previous simulation results of mDia1-FH1 domain, we observed an overall elongation for the end-to-end distance of profilin bound FH1 and neighbor cooperativity caused by profilin binding. So this donor-acceptor system was constructed to test the simulation result. Supposing the LBT fused mDia1-FH1 is an ideal chain that follows a completely random coil model, the root mean squared end-to-end distance will be 37.79 Å calculating by Paul Flory’s model

\[ \sqrt{R^2} = l \cdot N^{0.6} , \]

where \( R \) represents the end-to-end distance of the polymer, \( l \) is the unit length here representing the distance between the adjacent \( \alpha \)-carbon atoms which is 3.8 Å. This value taken from the work of Zagrovic and Pande [336]. They verified the predictions of the ideal random chain with unit length of 3.8 Å by analyzing internal distances of several unfolded ensembles of proteins. \( N \) represents the number of units refereeing to the numbers of amino acid inside LBT fused FH1 which is 46. Herein the maximum end-to-end length of our peptide is within the distance range of LRET. FH1 end-to-end distance changes indicated by the difference of energy transfer efficiency before and after FH1 binds profilin support the simulation results of end-to-end elongation after binding profilin. The decay study on profilin binding of donor-only FH1 is performed to test neighbor cooperativity observed from the simulation results in Chapter 2.
4.2 Results and Discussion

4.2.1 Overlap of donor and acceptor spectra

The donor consisted of terbium chelate that was fused to the N-terminal of mDia1 FH1 domain. The donor-integrated peptide was synthesized from GL Biochem (Shanghai) with a cysteine introduced at the C-terminal of the peptide for thiol reactive acceptor labeling. The terbium chelate used to enhance terbium luminescence was a LBT reported by Imperiali et al [337]. It has been disclosed to have a lifetime around 2.24 milliseconds (ms) that is long enough to collect decay data on a standard fluorometer. Figure 4.1 shows the structure of a terbium chelate available in PDB with the same coordinating residues and labeled in yellow. The six coordinating residues forms a hexadentate ligand structure shielding Tb(III) from water O-H groups at the first coordination sphere [338]. The trivalent terbium ion quenches tryptophan fluorescence and enhances subsequent phosphorescence. The mDia1-FH1 sequence used in this study contains two profilin binding sites with the typical IPPPPPL pattern, which adopts PPII helix conformation in most of the time, and one flexible linker in between which is supposed to elongate upon profilin binding based on our previous simulation result.
Figure 4.1 Stick representation of the X-ray structure of Tb(III) bound LBT. The metal-ligating residues are shown in yellow. The terbium ion is shown as a red sphere in [PDB: 1TJB] using MacPyMOL.

To choose an appropriate acceptor to label the FH1 peptide, the absorbance spectrum of the acceptor fluorophore should overlap with the emission bands of Tb(III). As indicated from Figure 4.2, Tb(III) has three emission bands peaked at 490 nm, 545 nm and 590 nm as its spectral characteristics. For such a wide range of emission spectrum, many organic fluorophores will have spectra overlap with trivalent terbium. Alexa Fluor dyes stand out for their high extinction coefficients, high photostability, high quantum yields and better brightness. There are commercially available amine, thiol, aldehyde and carboxylic acid reactive forms of Alexa Fluor dyes to label both peptides and proteins. For this study, Alexa Fluor 555 C₂-maleimide (AF555, cat. A20346; Invitrogen) was chosen as the acceptor. It has spectral
overlaps with two of the three Tb(III) emission bands at 490 nm and 545 nm (Figure 4.2).

Figure 4.2 Spectra of mDia1 FH1-labeled with either terbium or AF555. Solid dot and triangle are the emission and absorption spectra of FH1, respectively. Dotted line is the emission spectrum of LBT-FH1 with terbium and was obtained with 100 µs delay after flash.

4.2.2 LRET study

All emission scans were conducted in a Cary Eclipse Fluorescence Spectrophotomer equipped with a phosphorimeter that enables pulsed excitation and a delay after flash. Direct fluorescence of the acceptor fluorophore was eliminated by this temporal discrimination. The long lifetime, usually in milliseconds, of lanthanides could facilitate energy transfer after the delay period. LBT labeled mDia1-FH1 was excited at 280 nm that induces tryptophan sensitized luminescence of Tb(III). The scan data were collected after a 100 µs delay, so that direct fluorescence of the AF555 had decayed...
away (The lifetime of AF555 is 0.3 ns http://www.lifetechnologies.com/us/en/home/references/molecular-probes-the-handbook/tables/fluorescence-quantum-yields-and-lifetimes-for-alexa-fluor-dyes.html). Signals arising around 565 nm (emission maxima of AF555 is 565 nm) after the delay was purely derived from sensitized emission.

Energy transfer experiment began by adding terbium ions to AF555 labeled mDia1-FH1 and collecting the luminescence data. Figure 4.3 shows that the donor alone labeled FH1 (Tb-formin) is the only group that had emission signal after the delay. There was no detectable luminescence once the acceptor was added. A possible explanation for this phenomenon is that AF555 completely quenched the LBT-terbium donor, and acceptor sensitized emission decayed within the initial delay of 100 µs. Another explanation is the donor-acceptor labeled mDia1-FH1 preferred a loop conformation with acceptor very close to the tryptophan residue in LBT, so that the terbium ion failed to obtain luminescence enhancement and energy directly transferred from the excited tryptophan to acceptor. To verify these two explanations, the lifetime of the sensitized emission at 565 nm was measured. If AF555 completely quenched the donor, there would be lifetime enhancement at the emission maxima at 565 nm. Figure 4.4 shows the luminescence decay of acceptor-only (formin-AF555) and acceptor-donor (Tb-formin-AF555) groups within 155 µs at 565 nm after a delay of 10 µs. During 10ms, only the first few points had luminescence intensity above average, and data after the first 15 points were thus omitted in this figure. A mono-exponential fit was tested for these two groups resulting in two lifetimes at 10⁻⁶ scale. With too few points showing sensitized emission, little about acceptor lifetime can be concluded
from this data set. Unpaired t test reported these two groups were from the same population with $p = 0.18$. To summarize, acceptor-only and acceptor-donor group have the same decay rate for acceptor fluorophore. This indicates the trivalent terbium did not quench tryptophan in LBT, so the long lifetime feature of lanthanide dyes is absent here. The lifetime measurement of the quenched donor at 545 nm produced a flat line for luminescence intensity. This also suggests the absence of sensitized terbium as well as a loop conformation adopted by LBT-fused mDia1-FH1.

Figure 4.3 Emission spectra of donor and acceptor labeled peptides. Free Tb(III) as control (o), LBT labeled mDia1-FH1 added with Tb(III) as donor only (solid line with dots), mDia1 FH1 labeled with AF555 as acceptor only (+), and the donor-acceptor labeled FH1 with both LBT-Tb(III) and AF555 (×) are shown in this figure. Samples were excited at 280 nm. The gated emission was recorded from 400 to 600 nm. All data were normalized at 545 nm.
4.2.3 Measurement of luminescence decay

Once the acceptor was labeled, there was no detectable luminescence from decay experiment. So we used only donor-labeled FH1 to see if profilin binding had any effect on the lifetime of trivalent terbium. Figure 4.5 shows the lifetime data of LBT-Tb-FH1 (donor-only) and LBT-Tb-FH1-profilin (donor with profilin). All measurements were conducted with excitation at 280 nm that sensitized tryptophan in LBT. The collection of the luminescence intensity was at 545 nm after an initial delay of 50 µs. As shown in Figure 4.5, the donor with profilin group decayed faster than the donor-only group. The curves were fit using R package. The luminescence signal of donor-only group was a single exponential decay fit $I(t) = I(0) \cdot e^{-t/\tau}$, and the emission decay of donor with profilin group is a biexponential fit $I(t) = I(0)_1 \cdot e^{-t/\tau_1} + I(0)_2 \cdot e^{-t/\tau_2}$, where $I(t)$ is the luminescence value at time $t$, $I(0)$ is the initial luminescence value at time 0 and $\tau$ is the lifetime.
The lifetime of donor-only group was 2.27 ms. While, the donor with profilin group resulted in two lifetime values from the bi-exponential fit 
\[ I(t) = 0.64 e^{-t/2.11} + 0.39 e^{-t/0.72} \]. The long lifetime component corresponded to the donor-only group and the short lifetime component corresponds to the donor-profilin group. The percentages of each component stand for their amplitude. This means the donor-profilin group is composed of 64% donor-only FH1 peptide and 39% FH1 peptides bind with profilins, leading to a shorter lifetime of 0.72 ms. \( R^2 \) residuals of both curves were > 0.99.
There were no acceptors for these two groups, the only changes that occurred for the donor-profilin group is binding of profilin to the formin-FH1 which lead to a shorter lifetime of donor-only labeled FH1 at 545 nm. The characteristic emission spectrum of trivalent terbium recorded for the decay experiment indicates that the decay signal is indeed from sensitized terbium chelate and the shortened lifetime of terbium after profilin binding is probably due to conformational changes of the terbium chelate. As discovered by Horrocks et al. in 1978, the observed luminescence decay constant, $k$ (the reciprocal of the lifetime $\tau$), is extremely sensitive to the number of water molecules occupying the first coordination sphere of Tb(III); and there is a linear relationship between $k$ and the mole fraction of H$_2$O ($\chi_{H_2O}$). In other words, the lifetime $\tau$ is inversely proportional to the mole fraction of H$_2$O. The longer the lifetime is, the less water molecules occupy the coordination sites of Tb(III). The nonradiative energy transfer from Tb(III) to the water O-H oscillators greatly reduced terbium ion luminescence, because the excited-state energy transforms to O-H bond vibrational energy much faster than the luminescence emission [339]. Compared with donor-only labeled formin-FH1, the group with profilin added tends to have a long lifetime population (donor-alone labeled) and a relatively short lifetime population (profilin bound). Profilin binding here could decrease the lifetime of the sensitized Tb(III) by introducing more water molecules in the first coordination sphere of trivalent terbium. A conformational change of the terbium chelate is likely to be the cause. The chelate originally is in a loop conformation with trivalent terbium in the center of the chelate. Only with a more open conformation could water molecules incorporate into the first coordination sphere. Recall the results
from Chapter 2 that profilin binding not only induces a direct coil-to-
elongation transition of the formin-FH1 domain but also leads to
conformational changes in a neighboring site, which is known as neighbor site
cooperativity. The binding of profilin in one binding site affects the amino
acid residues near the binding site so that unbound sites also extend their
conformation upon occupation of their neighboring sites. In this case, profilin
binding to the formin-FH1 binding site induces an elongation of the LBT
which is right next to the profilin binding site in formin-FH1. In addition, the
terbium chelate becomes accessible for water molecules to enter the inner
coordination sphere of Tb(III).

4.3 Conclusion

In this study, mDia1-FH1 was genetically fused with an LBT to study
the conformational changes of FH1 under the effect of profilin binding in the
approach of LRET. The other end of FH1 was labeled with AF555 as the
LRET acceptor. When FH1 was labeled both with Tb(III) and AF555, neither
sensitized emission of donor nor acceptor was detected. Through
luminescence decay experiment, the donor was found to abort its quenching of
tryptophan residue of the LBT, which was supposed to enhance Tb(III)
luminescence. With the acceptor labeled, the tryptophan residue directly
transferred energy to the acceptor, so the conformation of both donor and
acceptor labeled FH1 was most likely a loop with two ends in close proximity.
The luminescence decay of donor-only and donor-profalin group indicates an
increased number of water molecules coordinated Tb(III). This in turn
suggests a conformational elongation of FH1 upon profilin binding. The
disordered FH1 peptide segment used in this study was possibly too short, thus
it mostly looped around in solution when both ends are labeled. However, the decay measurement enabled the detection of the neighbor effect as mentioned in chapter 2 because of a long lifetime donor.

4.4 Materials and Methods

4.4.1 Peptide and profilin

The synthesized donor labeled FH1 peptide sequence is MFIDTNNDGWIEGD ELLLEEGIPPPPPLPGVASIPPPPPLPGACAS. Mouse 2a profilin was expressed and purified as described in Chapter 3.

4.4.2 Terbium titration

Terbium(III) chloride anhydrous powder from Aldrich (99.999%) was dissolved in deionized water with a few drops of concentrated HCl to reach 100 mM. The concentration of terbium was determined by titration against standardized 0.1M EDTA solution using xylenol orange (0.1% w/v) as an indicator. The stock solution was later diluted as needed.

4.4.3 LRET

Emission measurements were made on a Cary Eclipse Fluorescence Spectrophotometer. The luminescence scans were recorded from 400 nm to 600 nm with 1 nm increments after a lamp pulse at 280 nm and a delay of 50 μs after the flash. The excitation slit width was 5 nm and the emission slit width was 5 nm. Samples were 20 μM FH1-AF555, 20 μM FH1-AF555 with 100 μM Tb(III), 20 μM FH1 with 100 μM Tb(III), in 10 mM HEPES buffer (pH = 7.5) in a 500 μL cuvette.
4.4.4 Measurements of decay

Luminescence lifetimes were measured in a Cary Eclipse Fluorescence Spectrophotometer. Samples were excited at 280 nm and there was a lamp pulse from the xenon flash lamp. After an initial delay of 50 µs, the intensity at 545 or 565 nm was monitored with 60 µs increments for 10 ms. The excitation slit width was 5 nm and the emission slit width was 10 nm. Samples were 20 µM FH1-AF555, 20 µM FH1-AF555 with 100 µM Tb(III), 20 µM FH1 with 100 µM Tb(III), 20 µM FH1 with 100 µM Tb(III) and 50 µM profilin, in 10 mM HEPES buffer (pH = 7.5) in a 500 µL cuvette.
Chapter 5

Conclusions and Future Research

This thesis focuses on the conformational changes of the disordered formin FH1 domain under the effect of profilin binding. We propose that the structural ensembles of formin FH1 domain will adopt an elongated form as the dominant conformation when profilin associates with the binding site. Multiple profilin binding events taking place simultaneously will induce the random coil shaped FH1 to transit into a jagged jack that pushes forward the membrane in front of the polymerizing actin filaments. A structural simulation method was developed to examine the effect of profilin binding, which may elucidate the biological relevance of formin FH1’s conformational change and its motor function in promoting actin filament barbed end polymerization. Most formins possess multiple profilin binding sites and the number of profilin binding site directly relates to the motor function. Thus, it is crucial to identify the conformational behavior of these profilin binding sites. We used a luminescent tag to inspect the conformational changes of FH1 upon profilin binding. We checked the thermodynamic profile of FH1-profilin binding and found that the reaction was exothermal, which agreed with the earlier report of this association. This project incorporates a variety of research fields that includes structural simulation, biostatistics and biophysics.

We studied the conformation of disordered mDia1-FH1 structural ensembles using a probabilistic sampling method. Profilin-bound FH1 was found to elongate comparing to the unbound FH1 structures and the end-to-end distance of the whole structure averagely increases with the number of bound profilins. This is realized by imposing profilin to the binding site.
“IP$_3$L” in FH1 structural ensembles to mimic the docking of profilin, which serves as an innate spatial constraint confining the available conformational space of FH1. The conformational changes of FH1 structural ensembles are identified from the comparison of the distributions of $R_{gyr}$ and end-to-end distance of structural groups bound with different number of profilins. As the bound profilins increased, the accessible conformational space of FH1 domain decreased dramatically and the FH1 structural ensembles generally shift to have a larger $R_{gyr}$ and a longer end-to-end distance. Such novel observation reveals that FH1 structural ensembles can rearrange under the effect only from profilin binding and the number of bound profilins is directly associated with the openness of the conformations. We argue that the binding of profilin alone is responsible for the elongation of FH1 ensembles and this conformational extension affects the formin motor function. The elongation of FH1 ensembles with an increased number of profilin-bound sites is statistically demonstrated to correspond to the boost efficiency of the formin motor. Although this proof is not sufficient for a causation conclusion, we derive a hypothesis that FH1’s conformational elongation is the missing link that connects the number of bound profilins and the efficiency of the formin motor, so that a profilin jagged jack model is proposed. In this model, the disordered FH1 domain gets stiffened upon profilin binding. Experimental data has shown that profilins dimerize when they bind to two concatenated binding sites, resulting in a pushing force to the N-terminal associated plasma membrane. This secures a high concentration of profilin-actin monomers to enhance the barbed end polymerization of growing actin filaments. It has not been explained why the number of bound profilins will promote the formin motor function. Hence, it
is intriguing to find out the role of the missing link “conformational
eelongation”, which provides a plausible mechanic explanation for profilin-
directed formin motor. The elongated conformation could possibly be required
for a disordered region normally to adopt to serve as a mechanical jack in
other similar mechano biological systems. An obvious extension of one
binding site is shown in the end-to-end distance distributions of all the
structural ensembles while its nearest neighbor sites are occupied with
profilins. Such neighbor cooperativity indicates that the excluded volume
effect of profilins not only applies to the site it associates with, but also the
nearest regions. Since one bound profilin could lead to the extension of more
than one binding sites in the FH1 ensembles, those profilins associated later on
the neighbor binding sites would induce lesser extension based on the
maximum conformational space one binding site can approach. For the last
binding site bound with profilin, the increment of its end-to-end distance
compared to its free state would be the smallest among all other binding sites
because there is no more neighbor site to display extra excluded volume effect
in the downstream. The neighbor cooperativity explains the decreased
contribution of single profilin-bound site to the end-to-end distance average
value observed from the distributions of structural ensembles bound with zero
to six profilins. Earlier experiments demonstrated that more profilins bound in
the FH1 domain tend to provide fewer stimuli in the actin polymerization.
This again connects our FH1 conformational elongation with formin motor
efficiency. Although more profilins are bound, the conformational elongation
is slowing down due to the reduced end-to-end distance of single binding site.
This further decreases the conformational space between FH2 and the plasma
membrane, so that the concentration of profilin-actins gets lower leading to a diminished acceleration effect of formin motor.

Binding of profilin and formin FH1 induces statistical elongation of the FH1 ensembles and stiffens the random coil shaped FH1. If there is energy release during binding, it could probably work through the elongated distance. Isothermal titration calorimetry study of profilin-FH1 binding was performed, and we found that the association of profilin and FH1 is an exothermal reaction, which releases the energy for our proposed “conformational elongation”-driven formin motor. The negative change of enthalpy demonstrated a stable binding event where energy is released. Additionally, the dissociation constant denoted a higher affinity of our chosen IP₃L pattern of profilin binding site than the peptide Poly-L-Proline with more number of proline residues. This finding reveals that the non-proline residues in FH1 play crucial part in profilin binding as well.

For a direct observation of the elongation of FH1 upon profilin binding, we genetically fused a lanthanide binding tag to the N terminal of FH1 peptide with two consecutive IP₃L binding sites. With milliseconds lifetime of sensitized trivalent terbium donor introduced into system, it is a surprise that no emission signal was detected from either donor or acceptor in the donor-acceptor labeled FH1. In the decay experiment, ten microseconds of delay time rescued some data point of acceptor-only and acceptor-donor group to show a rapid decay of acceptor fluorescence, which demonstrated no difference with or without terbium ion. Too close a distance between the tryptophan in lanthanide binding tag and the acceptor is, thus it was deduced that the terbium failed to quench the tryptophan and showed no sensitized
emission in the luminescence scan experiment. This suggests a loop conformation of the donor-acceptor labeled FH1. In the decay experiment of donor-only and donor with profilin group, a fraction of the donor labeled FH1 peptide was displayed to have a shorter lifetime after adding profilin. This decrease in lifetime in the absence of the acceptor could not possibly be derived from acceptor quenching; the only explanation is a change of the donor itself. Earlier experimental data reported the negative linear relationship of mole fraction of water molecule that associates with terbium in its first coordination sphere and the lifetime of its sensitized state [339]. This indicates an elongation of the lanthanide binding tag for the neighbor cooperativity introduced by the bound profilin.

In summary, we detected a conformational elongation of FH1 ensembles upon profilin binding and the extension value is positively linked to the number of bound profilin. The findings lead us to a hypothesis of “conformational elongation”-driven formin motor model that poses FH1 as a cooperative jack. The FH1 jack pushes forward the membrane before the growing actin filaments while binding profilins. The thermodynamic profile of profilin-FH1 binding showed an energy release upon the binding and provides the potential energy for the formin jack motor. The luminescence decay experiment showed a decrease of lifetime from donor-only peptide to donor-profilin peptide. This indicates more water molecules coordinated the terbium ion inside the lanthanide binding tag, which could only result from the elongation of the LBT as the neighbor cooperativity effect of the FH1 domain upon profilin binding.
Bibliography


117. Syme, C.D., et al., A Raman optical activity study of rheomorphism in caseins, synucleins and tau - New insight into the structure and


229. Wallar, B.J., et al., The basic region of the diaphanous-autoregulatory domain (DAD) is required for autoregulatory interactions with the


308. Torda, A.E., R.M. Scheek, and W.F. Van Gunsteren, TIME-DEPENDENT DISTANCE RESTRAINTS IN MOLECULAR-


