5-2016

Complex Non-equilibrium Structural Dynamics in Globular Proteins

Xiaohu Hu
University of Tennessee - Knoxville, xhu12@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Xiaohu Hu entitled "Complex Non-equilibrium Structural Dynamics in Globular Proteins." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Jeremy C. Smith, Major Professor

We have read this dissertation and recommend its acceptance:

Jerome Baudry, Hong Guo, Tongye Shen, Xiaolin Cheng

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Complex Non-equilibrium Structural Dynamics in Globular Proteins

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Xiaohu Hu

May 2016

To my parents, Weiping Hu and Ping Jing.

Acknowlegement

First of all, I would like to thank my PhD advisor Dr. Jeremy C. Smith for giving me the opportunity to carry out my PhD research at the UT/ORNL Center for Molecular Biophysics (CMB) and his continuous support of my work throughout the years, who provided me guidance, resources and most importantly, academic freedom for me to pursuing my own scientific interest, which made this fruitful PhD project possible at the first place. Next, I would like to thank Dr. Xiaolin Cheng from Oak Ridge National Laboratory and the former post-doc researcher Dr. Liang Hong at CMB, who helped me in many of the difficulties I have encountered and provided me endless hours of discussion during the years of my PhD, and as well as Dr. Thomas Neusius whose expertise on non-ergodic subdiffusion provided important suggestions and inspiration for my own research. Finally, I would thank my loving parents Weiping Hu and Ping Jing who have been always given me encouragement and support in all aspects of my life and my pursuit in science.

Abstract

Internal structural motions in proteins are essential to their functions. In this present dissertation, we present the results from an extensive set of molecular dynamics simulations of three very different globular proteins and demonstrate that the structural fluctuations observed are highly complex, manifesting in non-ergodic and self-similar subdiffusive dynamics with non-exponential relaxation behavior. The characteristic time of the motion observed at a given timescale is dependent on the length of the observation time, indicating an aging effect. By comparing the simulation results to the existing single-molecule fluorescence spectroscopic data on other globular proteins, we found the characteristic relaxation time for a distance fluctuation within proteins, such as inter-domain motion, increases with the length of the observation time in a simple power-law that appears to be universal and independent of protein species, spanning over enormous 13 decades in time ranging from picoseconds up to hundreds of seconds. We argue that the observed self-similar dynamics arises from the fractal nature of the topology and geometry of the underlying energy landscape. Diffusion of a fictive walker over the complex hierarchical energy landscape leads to structural dynamics that are best described by a noisy, subdiffusive continuous time random walk, consistent with the aging and observed broken ergodicity. In comparison with data from single-molecule experiments in the existing literature, the present results suggest that the structural dynamics of single protein molecules is likely to remain non-ergodic and out of equilibrium on most timescales over which protein functions occur, eventually persists up to typical lifespan of proteins *in vivo*.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Biological systems, such as proteins, cells or an entire organism, are, among all physical systems, probably the most interesting and challenging ones. Biological systems are characterized by a high degree of complexity, self-organization, adaptivity, and most of all, the interplay of different components within these systems give the rise of a uniquely complex macroscopic behavior, known as *life*.

Proteins are a crucial part the elementary building blocks in any biological systems known so far (anno 2016). The word protein is derived from the ancient Greek word *proteios*, meaning "I take the first place", emphasizing its central role in all living organisms. This denomination was introduced by the Swedish chemist Jöns Jakob Berzelius in 1838 [102]. Inside a cell, proteins serve virtually all important functions, ranging from providing structural integrity, *i.e.* cytoskeleton, the structural frames that give eukaryotic cells their physical shapes, to biochemical catalyst that enabling important chemical reactions to occur in a relatively "timely" fashion. In the chemical sense, proteins are polymers that consists of amino acid monomers which "fold" to unique and mostly well-defined[1] three-dimensional structures determined by the primary sequence of amino acids. This folded state is also referred to as the "native state". The structure of the protein is directly related to its function as any misfolding of the primary peptide sequence leads to loss of the function of the protein and, in many cases, can result in many degenerative conditions and diseases,such as Alzheimer's, Parkinson's, or diabetes [117], just to name a few.

However, despite the importance of the tree-dimensional structure of the protein for its function,

[1]There are also proteins that fold to a less well-defined or partially disordered structure. These are referred to as "intrinsically disordered proteins".

there is another crucial aspect of protein function, namely the *dynamics*. A protein needs to move to perform its function. The ability of the protein being able to move arises from the fact that even in the native states, the folded protein is able to adopt a large number of similar conformations, the so-called conformational substates, and these substates are likely to have functionally different properties [60]. At sufficiently high temperatures, *e.g.* room temperature, the protein can quickly switch between these different conformational substates resulting in structural fluctuations that can be considered as diffusion on a rugged potential energy landscape [44] where a local minimum is associated with a conformational substate. It has been demonstrated experimentally that a protein can only be functional at sufficiently high temperature [114], i.e. above the so-called dynamical transition temperature [38, 106, 114, 115], for the protein to perform certain anharmonic motions, i.e. transition between different conformational substates, to enable its function [38, 106, 115]. Although the exact value for the dynamical transition temperature is still a subject under debate [84].

It is of great importance to understand the physics behind the internal protein dynamics. Driven by thermal energy, proteins are the molecular machineries that carry out virtually all essential functions in a cell. However, details of these motions are still elusive. One hand, the thermal structural fluctuation of protein is a random, stochastic process, but on the other hand these seemingly random fluctuations do somehow result in, at least averaged over an ensemble and a finite period of time, a highly specific and precise function. As once pictorially described by Professor Igor M. Sokolov from Humboldt University, Berlin, Germany, "A protein is not a Carnot engine, and not a BMW motor"[2]. Besides the understanding of the basic physics behind this phenomenon, protein dynamics is especially important for the understanding of protein's biological functions, such as ion channel gating, allosteric, cell signaling, enzymatic activity, etc., *i.e.* how do these microscopic giggling or conversion between slightly different conformations around the native structures will manifest in simple or complex macroscopic observables, such as reaction rates or regulation of cellular activities.

Early models assumed that the internal dynamics of protein is an overdamped Brownian motion in an external confining potential that can be modeled by a classic Langevin equation [88]. However, in the subsequent decades, a large body of experimental and numerical simulation works demonstrated that this simple, Brownian picture of the internal protein dynamics under physiological conditions has become increasingly questionable [52, 53, 70, 81, 95, 110, 143]. The present dissertation is intended to address this question and introduces a novel model for protein dynamics and interpretation of the internal dynamics of single, globular proteins, which has significant implications on its ensemble averaged behavior in the cellular environment.

One very important protein function is the catalysis of essential biochemical reactions in a cell, that, otherwise, will not occur (at least not on any timescales relevant to the lifespan of any living beings) due to the extremely high natural activation energy barrier separating the substrate and

[2]Personal communication

product. Here, proteins, or more precisely, enzymes, can facilitate these reactions by lowering the barrier so they can occur on much shorter timescales meaningful to the cell. The detailed knowledge of how proteins perform these functions on the molecular level, and also how to enhance or interfere with these functions, are crucial for research areas such as protein engineering or medical drug design and discovery.

One important quantity that is used to characterize the level of activity of a chemical reaction is the rate of the reactions, *i.e.* catalysis events occurring per time unit. Simple chemical reactions can be modeled as a classic Kramer's barrier crossing event [68]. In this picture, the reaction is modeled by the diffusion of a fictive random walker along a certain reaction coordinate in a potential of mean force. Starting in an initial local minimum representing the reactant state, the walker can cross over the confining barrier, or the activation energy, and reach another minimum representing the product state. The rate of the reaction, k, is related to the activation energy barrier $\Delta E^{\ddagger}$ and the temperature T via the well-known Arrhenius' relation

$$k = A(T) \exp\left(-\frac{\Delta E^{\ddagger}}{k_B T}\right), \tag{1.1}$$

where k_B the Boltzmann constant and $A(T)$ is a temperature-dependent prefactor. In the conventional sense, if one speaks of a reaction rate *constant*, one usually implicitly implies the system one refers to has already reached the thermodynamical equilibrium, therefore, a time-independent average rate exists (thus the word *constant*) and provides a good representative measure of the reactive activities going on in the system. The existence of a *time independent* rate constant requires the satisfaction of several criteria [60]; Besides well-defined reaction coordinate and activation barrier that is at least a few $k_B T$ high separating the reactant and product states, most importantly, the relaxation times of all degrees of freedom, other than the reaction coordinate, involved must be faster compared to the motion along the reaction coordinate [60]. Furthermore, a constant rate k implies that the kinetic of the reaction follows a single exponential behavior. For example, a reaction of the interconversion between two substances

$$\mathrm{C}_1 \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} \mathrm{C}_2$$

can be expressed with a first order differential equation

$$\frac{dC_1(t)}{dt} = k_{12} C_2(t) - k_{21} C_1(t) \tag{1.2}$$

$$\frac{dC_2(t)}{dt} = k_{21} C_1(t) - k_{12} C_2(t). \tag{1.3}$$

Assuming the systems is fully equilibrated and the rates k_{12} and k_{21} are time-independent, the

relaxation towards the equilibrium concentration up on any perturbation δC can be expressed as

$$\delta C_1(t) \equiv C_1(t) - C_1^{\text{eq}} = \delta C_1(t=0)\exp(-kt) \tag{1.4}$$

$$\delta C_2(t) \equiv C_2(t) - C_2^{\text{eq}} = \delta C_2(t=0)\exp(-kt) \tag{1.5}$$

with $k = k_{12}+k_{21}$ and equilibrium concentrations C_1^{eq} and C_2^{eq} of the substances 1 and 2, respectively. These requirements described above are fulfilled for simple chemical reactions, *e.g.*, in the gas phase where the kinetic energies of the colliding molecules precisely follow the Boltzmann distribution and will occasionally reach a high enough value to overcome the activation energy. However, reactions catalyzed by proteins or involving proteins, *e.g.* opening and closing of an gated ion channel or protein folding, the situation is more complicated and the prerequisites for the a well-defined reaction rate are often violated.

The pioneering experiments of myoglobin rebinding of CO and CO_2 after photodissociation at cryogenic temperatures carried out by researchers around Hans Frauenfelder in the 1970s [5] revealed the existence of different conformational substates of the same protein in solution, and more importantly, the ensemble averaged reaction rate shows a broad power-law distribution and deviates from Arrhenius behavior at low temperatures ($T < 180$ K) [5]. At this temperature range, proteins in the sample are frozen in different conformational substates and unable to relax. One amazing realization made was that these different conformers can pose very different effective activation energy barriers, leading to the observed board distribution. At higher temperatures, the onset of relaxation processes allows proteins to rapidly change from one conformation to another, and the observed ensemble averaged reaction rate starts to narrow and becomes more Arrhenius-like.

However, quite often, experimental data have shown that many protein reactions still exhibit anomalous, nonexponential kinetics even at room temperatures [70, 110, 121]. Different models have been proposed to explain these observations of the "rate constants" that are apparently no longer constant. Most of them are based on two major hypothesis; The first one is the so-called *static disorder* [25, 142], proposing that the same species of proteins can folded into different conformational substates within the native state, each has a fixed but different activation energy barrier. The second one is the so-called *dynamic disorder* [145], which states that the thermal fluctuation of the protein structure causes the protein to interconvert between different conformational substates each with a different reaction barrier. As result, the effective reaction barrier is a stochastic function of time which leads to the fluctuation in the reaction rate and broad distribution. However, it was impossible to distinguish exactly which one of the two mechanisms is responsible for the observed nonexponential kinetic in protein reactions in the usual ensemble averaged experiments, until the advancements achieved in the single-molecule fluorescence spectroscopy enabled the direct measurement of the catalytic activity of single protein molecules over a long period of time [80, 96, 141]. Modern single-molecule experiments are capable of observing individual reaction

catalyzed by a single enzyme over a long period of time (up to $\sim 10^2$ seconds) [36, 80] and can directly measure the dynamics of the internal structural fluctuation of a single protein molecule on the same timescale [95, 143]. These data have clearly shown that not only does the catalytic rate of a single enzyme fluctuates with the time, more importantly, these fluctuations rather occurs on a long timescales that is comparable to those of the catalytic event itself [80, 141]. Further more, the structural fluctuation of the protein also highly correlates with the fluctuation of the catalytic rate, exhibiting very long correlation time up to seconds [80, 95, 96, 143]. Elaborated statistical analysis of these measured single molecule time series data provided clear evidence that the dynamic disorder is responsible for the temporal fluctuation of the catalytic rate [80, 96, 141].

Early single molecule experiments showed that the time averaged, cumulative catalytic activity of individual enzyme molecules are vastly different from each other [25, 142]. Xue et al. have shown that even after a long observation time of ~ 2 hours, clear discrepancies remain in the total amount of product catalyzed between individual enzyme molecules, although the activity for individual enzymes seems to have reached a steady state on the timescale of the experiment [142]. These results have been interpreted as supporting evidence for static disorder, suggesting the different protein molecules may have intrinsically different catalytic activities due to different conformers they adopted. However, these data do not directly contradict the dynamic disorder because it is not clear, what is the timescale of the slowest mode of the temporal fluctuations in the catalytic rate or the protein structure, or even such an upper limit does indeed exist. The picture of the static disorder would be valid if the longest timescale of the fluctuation in the catalytic rate associated with the protein thermal motion is significantly shorter than the length of the observation time. However, it is well known that extremely high barriers between conformational substates can exist that can require very long time to cross. In this scenario, even on comparably long timescale such as hours, the fluctuation of the catalytic rate due to dynamic disorder can appear static [141].

This dissertation introduces a novel model for conformational dynamics in globular proteins [52] that favors the picture of dynamic disorder, but yet, still consistent with the seemingly static but vastly different catalytic rates among individual enzyme molecules as seen in the time averaged single molecule experiments such as ref. [142]. We propose an alternative interpretation of the protein dynamics, namely, we argue that the protein structural fluctuation is a nonergodic stochastic process and out of the dynamical equilibrium on timescales for most protein functions. As a consequence of the broken ergodicity, the structural dynamics and the associated functional properties of individual protein molecules are non-stationary and will exhibit aging [90] and such processes can result in population splitting [90, 120]. In other words, during the same period of observation, some proteins may seem inactive whereas others may appear to be highly active or exhibiting an varying activity somewhere between the extremes. This prediction is consistent with the observation of vastly different, but yet static appearing catalytic activities between individual enzyme molecules

over a long period of observatin time ($\sim$ 1-2 hours) in single molecule experiments [142]. Such non-stationary and non-equilibrium dynamical behavior of proteins are highly unintuitive and can have significant impact on the current understanding of protein functions in general, and as well as the biological processes on larger scales such as on cellular or even organism level, which are essentially based on the collective, concerted functional behavior of individual proteins molecules.

Chapter 2

Background and Theoretical Concepts

In this chapter, we introduce the relevant concepts and theories needed for the discussions on the internal structural dynamics of proteins, and also provide readers, who may not be familiar with the subject, background information in order to better understand the content of this dissertation. We first present a brief review on protein biophysics with the focus on the relationships between protein structure, dynamics and function together with a brief historical review on this topic. We then introduce some relevant basic concepts of statistical mechanics, and followed by an introduction of anomalous diffusion which is the key approach of modeling the protein structural dynamics in the present dissertation.

2.1 Protein biophysics - a brief perspective on protein structure, dynamics and function

All living things can be viewed in a hierarchy of

(DNA) $\leftrightarrow$ Proteins$\leftrightarrow$ Organelles$\leftrightarrow$ Cells $\leftrightarrow$ Tissues $\leftrightarrow$ Organs $\leftrightarrow$ Organisums [45].

Among these entities, proteins are a class of fascinating biomacromolecules. They are the molecular work horses that virtually serve all functions in cells. From chemical point of view, proteins are linear, unbranched polymers composed of monomeric units from a space of twenty different amino acids. When translated into a lengthy polymer by the ribosomes, this seemingly limited set of twenty available amino acids can incorporate an can enormous space of possible sequences. For the average

protein primary sequence length of roughly 300 amino acids [76], the possible number of proteins can be made is on the order of $20^{300} \sim 10^{390}$. This number is in fact so huge that even only a tiny fraction of a trillionth of a trillion of these sequences are biologically meaningful in some way, the potential number of proteins encoded by this sequence length would be still way beyond the mind-blowing order of $\sim 10^{300}$.

The term *Biophysics* first appeared in a paper by J. R. Loofbourow published in the journal *Reviews of Modern Physics* in 1940 entitled "*Borderland problems in biology and physics*" [45, 77]. Generally speaking, biophysics includes theoretical and experimental approaches to study problems faced in biological systems. With the advancement in the development in computer hardwares and numerical algorithms, the computational approach has become a substantial part of biophysics complementing different experiments. As fundamental as the central dogma of molecular biology, *i.e.*

$$\text{DNA} \xrightarrow{transcription} \text{mRNA} \xrightarrow{translation} \text{Protein},$$

describing the relationships and dependencies between DNA, mRNA and protein, there is a similar central dogma of proteins, namely the relationships between protein structure, dynamics and function. Both experimental and theoretical approaches from physics are central in studies and understandings of this trinity.

2.1.1 Protein structure and x-ray crystallography

The primary source for protein structure determination is the x-ray crystallography. The capability of resolving crystal structures may have been the most important influence in molecular biology coming from physics [45]. Ever since German physicist Wilhelm Conrad Röntgen discovered the mysterious radiation he referred to as the "x-ray" in 1895 (for which he received the first Nobel price in 1901), physicists, such as Max von Laue, who introduced x-ray diffraction, and W. L. Bragg, who derived the famous diffraction law of waves on lattice (Bragg's law), have made the determination of the crystal structure possible and paved the way for one of the most crucial scientific discoveries in the 20th century by James Watson and Francis Crick in the 1953. The duo concluded the double helix structure of the DNA, the molecule that encodes the genetic blue prints of all known living beings, based on the x-ray diffraction data. The first protein crystal structure was resolved in 1958 – the sperm whale myoglobin, by John Kendrew, and shortly afterwards, together Max Perutz, they resolved solved the structure of the much larger hemoglobin. In 1971, the today's well-known protein structure data bank PDB was introduced and H. C. Watson and J. Kendrew deposited the first myoglobin structure under the four-letter PDB identification code "1MBN".

All these early pioneering crystallography effort were indeed heroic [45] (for which Perutz and Kendrew received Novel price in Chemistry in 1962), given the experimental difficulties and the

minimal level of existing knowledge at the time[1], resolving these structures often took years of hard work. Nevertheless, these structures provided first crucial atomistic picture and insight into DNA and proteins, thus established the important connection between the structures and functions of biomacromolecules.

Thanks to the advancements in many technological areas in the past decades, such as synchrotron x-ray source, free electron laser, better detectors and as well as the ever increasing computing capabilities, the number of protein structures resolved and deposited into the PDB has been increasing exponentially. As of 2016, the PDB already contains over 110,000 crystal structures. Although the x-ray structures of protein gives an impression of the uniqueness, this structure is not a rigid construct but rather the reflection of an ensemble average of protein structure under the crystalline environment. Dynamical activities around this average is possible, and in solution, large conformational changes can occur and as a part of functional motions in proteins. In the subsequent subsections, we will further detail the complexity around protein structure, dynamics and function.

2.1.2 Protein folding and energy landscape

Proteins are complex systems [44, 45, 60] not only because of its high degrees of freedom but also due to its complex behavior. Unlike most other organic compounds which have a well defined structure and properties upon the synthesis, proteins come first as a random linear heteropolymer of amino acids after the translation by ribosome. It needs to undergo a subsequent folding process in order to acquire its final native and functional structure. The folding process itself is anything but trivial and still an very active area of research. First of all, in order to fold into the right native structure, the protein needs to have the correct primary sequence, often a single point mutation can cause the folding to fail completely and resulting in serious consequences for the cell and even the entire organism. On the other hand, despite the sequence specificity required for the proper folding, many completely different sequences can fold into essentially the same structure and carry out the same function, such as members of the lysozyme family [7].

Shortly after the pioneering protein unfolding/refolding experiments on ribonuclease A by Christian Anfinsen [4] in the early 1960s, American molecular biologist Cyrus Levinthal already noted the complexity of such a process with his famous Levinthal paradox: Due to the large number of the possible conformation the a protein can adopt, the folding time required would be astronomical [20]. He presented a thought experiment in the following way: Assuming each amino acid backbone can adopt two different conformations, *i.e.* ϕ, ψ angles, for a protein with 100 amino acid, there would be $2^{100} \sim 10^{30}$ possible protein conformations. If the time required for the sampling of each conformation is 1 ps, the total time required to sample all these conformations

[1] In fact, the father of x-ray diffraction Max von Lauer believed that structures of biological molecules cannot be resolved via the technique he has developed, until he was proven wrong by Watson and Crick in 1953 [45].

would be $\sim 10^{10}$ years [20]. However, in reality, the protein folds on much faster timescales of micro- to milliseconds, therefore appears to be paradoxical in Levinthal's argument. As we will show shortly later, this paradox can be explained by the concept of energy landscape and folding funnel, which offers a unified description of the protein folding and the relationship between structure and dynamics.

The experiments of CO and CO_2 rebinding of myogolbin after photodissociation with frozen protein samples at cryogenic temperatures by Austin et al. [5] have shown that for an ensemble of already folded proteins identical in their primary sequence, individual proteins can adopt slightly different conformations around the native structure in solution and were frozen in these different conformers. This new view involving a broad distribution of protein properties due to different conformations led to the birth of the energy landscape concept introduced by Hans Frauenfelder et al. [44].

A natural description of the full system, *i.e.* the large number of related, but yet different protein conformational substates associated with the native folded state at a constant temperature and pressure, is the Gibbs free energy [20]. The effective free energy, implicitly averaged over all solvent conformations, can be expressed as a function of the full atomic coordinates of the protein [20]. Such a function can be pictured as a high-dimensional "landscape" of free energy. Each point of the landscape is associated with a conformation, regardless whether it is a folded protein or just a unfolded random peptide chain. In this way, all protein conformations are characterized in a statistical and probabilistic manner, since the free energy of associated with a set of atomic coordinates is reciprocal to the probability of finding the protein adopting that particular conformation. The false assumption made in the Levinthal's paradox is that all conformations a peptide chain can adopt are somewhat equally likely. In that sense, the energy landscape is comparable to a relatively flat golf course with a single hole at a unknown location somewhere on the course representing the native state. In this analogy, folding a protein is equivalent to hitting the golf ball randomly into different directions until the ball finally rolls into the hole, a process that obviously can take a very long time to complete. In reality, different conformations are associated with vastly different probabilities, where as partially folded molten globular conformations in solution have much lower free energies comparing to random chains and with the folded native structures at the very bottom [32, 146]. Therefore, if the entire golf course is shaped like a giant funnel with the hole at the very bottom, even a blind-folded golfer would be able to bring the ball into the hole quickly, simply because after each strike, the ball will always roll closer the hole. This funnel-shaped global energy landscape for peptide chain is referred to as the folding funnel, reflecting the thermodynamical driving force for the protein folding.

As we will discuss in the later chapters, even at the bottom of the funnel, the energy landscape around the global minimum associated with the folded native state is not a plain smooth surface

but rather of a sufficiently large area with complex fractal, self-similar substructures with many local minima. These features give the rise of highly complex internal conformational dynamics of proteins in solution under ambient conditions. Although the initial transition from a unfolded state to a global minimum representing the overall folded state is a thermodynamically driven, thus a fast process, but the relaxation from a point in the general bottom region of the funnel to the true global minimum may not necessarily occur on a short timescales that is comparable to those of most protein functions. In fact, as early as 1969, Cyrus Levinthal already concluded that the protein conformation right after folding does not necessarily to be the one with the lowest free energy. It suffices a meta-stable state in a sufficiently deep well in order to sustain all the perturbations that would otherwise lead to the unfolding of the protein [45]. The question as whether the structural relaxation of folded globular proteins will indeed reach the global minimum and converge to an stationary and ergodic process is the primary subject of the present dissertation.

2.1.3 Protein dynamics and function

Proteins are dynamical entities and its ability to move is essential for carrying out its function. The energy landscape is not merely a way to characterize the existence of many possible conformations and their probabilities, but moreover, it also determines the dynamics of the protein [42]. Just like protein folding can be considered as the motion a random peptide chain sampling the folding funnel by rolling down towards the bottom following a certain pathway, the thermal motion of the folded protein can be considered as the sampling of local structures within the global minimum associated with the native state. Since the ultimate protein behavior, *i.e.* its function, is driven by the changes in free energy ΔG consisting of the enthalpy change ΔH the changes in entropy ΔS, *i.e.*

$$\Delta G = \Delta H - T\Delta S, \tag{2.1}$$

where T is the absolute temperature. Therefore, the trinity of protein structure (H), dynamics (S) and function (G) are intimately related to each other through a fundamental law in the thermodynamics [65].

Structural dynamics of folded proteins are related to many functions such as enzyme catalysis [16, 78, 96], protein-protein or protein-ligand binding and recognition [26] crucial for signaling and regulation for cellular processes [126], allostery [63], or promiscuity and multi-functionality, which is important for evolution of new biological function and pathways [64, 132]. For the protein folding, the proper dynamics and pathways are important to ensure that proteins are folded correctly, and conversely, improper dynamics can lead to misfolding [65] and is related to many serious diseases.

The first clue of protein's flexible nature has already been provided by the early x-ray scattering data on protein crystals via the so-called Debye-Waller factor (DWF). Named after the physicists

duo Peter Debye and Ivar Waller, who have shown that the intensity of the scattered x-ray of the wave length λ by a harmonic oscillator at a scattering angle Θ will decrease by a factor of

$$f_{\mathrm{DW}} = \exp\left(-16\pi^2 \left\langle x^2 \right\rangle \sin^2(\Theta)/\lambda^2\right) \tag{2.2}$$

where $\langle x^2 \rangle$ mean-square fluctuation of the harmonic oscillator. If protein atoms in the protein crystals are indeed rigid, then DWF observed in protein crystal would be uniformly small and close to zero. However, this is not the case. In the pioneering protein x-ray crystallography experiments in the 1960s, scientists have observed that the DWFs adopt significant values for atoms at different locations along the protein primary sequence in the crystals [45], indicating that the vibrational motions, even in the solid crystalline proteins, are not homogeneous through out the protein and different regions vary in terms of structural flexibility. The significant DWFs observed in protein crystals where interpreted as supporting evidence for the existence of conformational substates for well-folded protein [45].

Despite the vibrational flexibility, frozen proteins at cryogenic temperatures do have a well-defined time-averaged conformation, despite small vibrational motions, that poses a constant barrier for reactions, such as the CO rebinding after photo-dissociation of myoglobin. Experiments by Nienhause et al. [103] have demonstrated that the CO rebinding rate to individual proteins randomly frozen in solution follows a single exponential Arrhenius behavior, although the rate differs significantly among individual proteins due to the different conformational substates in which they are frozen. Even in the crystalline solid state environment, atomic fluctuations in proteins are not always harmonic vibrations. Protein crystals exhibit modes of motions beyond the vibration like in a simple crystalline solids, *e.g.* copper crystal. Such complex motions in a solid state environment is best reflected by a phenomenon referred to as the *dynamical transition* observed in an series of early experiments of the F_{57} Mössbauer spectroscopy of heme containing myoglobin [62, 66], or via neutron scattering [27, 33] and x-ray crystallography [43, 130] on hydrated protein crystals, powders or frozen solutions [28]. Dynamical transition essentially describes a discontinuous increase in the atomic mean square displacement (MSD) with the increasing temperature. Around 180-230 K (depending on the experimental techniques used to observe the transition), an non-linear, sudden increase in the MSD as a function of temperature is observed. This sudden increase is interpreted as the onset of anharmonic structural relaxations, such as transition between two or more local minima on the energy landscape, rather than those motions within a single local minimum approximated by a harmonic potential, as seen in the temperature range below the dynamical transition. In these regions, the MSD increases roughly linearly with temperature, *i.e.* $\langle x^2 \rangle \sim k_B T/(m\omega^2)$, where k_B is the Boltzmann constant T the temperature, m the atomic mass and ω the angular frequency of the harmonic potential [66].

The situation can be even more complicated for proteins in solution at higher, physiological temperatures. In this case, the full spectrum of of available modes of structural relaxation sets on

and the protein can move from one conformational substate to another on all accessible regions of the energy landscape at a given temperature. Therefore, more slower modes of motions can be carried out corresponding to transitions between distant states or states separated by high barriers. Theoretically, one can divide the motions into two categories, *i.e.* the "equilibrium fluctuations" (EF) and the "functionally important motions" (FIM) [45]. The former is associated with the intrinsic thermal fluctuation when protein is dwelling inside a local minimum on the energy landscape representing a conformational substate, while the latter describes the transitions over the barriers separating one substate to anther. As we shall see later, these two types of motions can become related and hard to differentiate if the system is close to equilibrium. Due to the hierarchical and self-similar organization of the energy landscape [44], at any given tier of the hierarchy, motions that are considered as EF may become FIM after the protein moves into a new, smaller well belongs to a lower tier and starts to explore the local structure of this smaller well in order to move towards another new local minimum with eventually lower free energy.

Protein dynamics in solution at ambient temperatures can be measured via a variety of different experiments, most common ones include nuclear magnetic resonance (NMR) [54], quasi-elastic neutron scattering (QENS) [46], and more recently (since *circa* 2000), single-molecule fluorescence spectroscopy based techniques, such as light-induced electron transfer [143]. On the numerical front, molecular dynamics (MD) simulations have advanced tremendously in the past decades and is capable of directly revealing the full dynamical picture of proteins on timescales up to milliseconds. We will discuss MD simulations in greater details in Sec. 3.1 and MD is often used to complement experiments in order to provided a detailed dynamical and functional picture of the protein.

NMR is a measurement of the ensemble averaged relaxation behavior of the nuclear spins of certain natural protein atoms, or their isotopes (artificially engineered into the protein, *e.g.* deuterium (H_2), N_{15}, C_{14}, etc.) in an external magnetic field. Although it cannot directly reveal the protein dynamics in the same fashion as a computer simulation does by offering an explicit picture of how each individual atomic coordinates change, however, it can detect motions at different sites in the protein simultaneously together with the corresponding time scales, amplitudes, and energetics of these motions [35] processes occurring on short-time sub-nano second timescale and a long timescales from μs up to hours [54, 65]. We referred to recent review articles, such as such as ref. [65], for details on how NMR can be used to study protein dynamics.

Neutron scattering can reveal protein dynamics by detecting the energy changes of the neutrons scattered by the protein sample at different scatter vectors Q. The changes in the energy ΔE reveals the timescale of the probed protein motion while the scattering vector Q is associated with the length scale. The term quasi-elastic refers to the focus on scattering signals close to the elastic peak ($\Delta E = 0$) in the scattering profile, since these neutrons with small energy loss (thus quasi-elastic) reveals the slower motions on timescales longer than nanoseconds, which are more functionally

relevant for proteins. The resulting data are often presented in the form of coherent or incoherent intermediate scattering function [46]. The first can be interpreted as the density correlation functions (both self- and cross-correlations) of predominantly protein hydrogen atoms on the length scale $\sim 2\pi/Q$, while the latter represents only the self-correlation corresponding to the ensemble averaged diffusive motions over the length scale given by the scattering vector Q. Neutron scattering can be used to study dynamical characteristics of local motions such as methyl-group rotation [51] on short ps timescales or the dynamics of inter-domain motions on timescales of 10–100 ns [53]. One limitation posed by neutron scattering is that the internal dynamics of the protein probed is convoluted with the global translational and rotational motion. It is still a challenge to clearly and unambiguously separate these different contributions. Here, MD simulation comes in as a helpful tool for interpreting the neutron scattering data. We refer to papers such as ref. [46] for detailed review on how neutron scattering can be used to study protein dynamics.

2.2 Some basic concepts in statistical mechanics

Thermodynamics concerns with the relationships between certain macroscopic properties, *i.e.* the thermodynamic variables or functions, of a system in equilibrium [49]. Statistical mechanics goes beyond these relationships and allows one to study the connections between the properties of the molecules that made up the system and the observed values of thermodynamic functions [49]. In this section, we will briefly discuss some basic definitions and concepts from the statistical mechanics that are relevant for the present dissertation.

2.2.1 Phase space and Gibbs ensemble

Consider a dynamical system containing N moving Newtonian particles of the masses m_i, $i = 1, ..., N$ with the total degrees of freedom $f = 3N$. The positions and momenta of these particles can be expressed in $6N$ canonical coordinates in the form of two vectors, the generalized position coordinates $\mathbf{q} = (q_1, q_2, ..., q_{3N})$ and the conjugated momenta $\mathbf{p} = (p_1, p_2, ..., p_{3N})$, together spanning a $6N$ dimensional space. This space is referred to as the *phase space*, commonly denoted with the Greek letter Γ. As the system evolves with the time t, the corresponding phase space point will move from one location to anther forming a phase space trajectory $\mathbf{s}(t) = (\mathbf{q}(t), \mathbf{p}(t))$.

The equation of motion of the phase space points are governed by the Hamiltonian dynamics, or the canonical equations of motion

$$\dot{q}_i = \frac{\partial H}{\partial pi} \tag{2.3}$$

and

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}, \tag{2.4}$$

where H is the Hamiltonian function that represents the total energy of the system, *i.e.*

$$H = T(\mathbf{p}, \mathbf{q}) + U(\mathbf{q}) \tag{2.5}$$

with T is the total kinetic energy and U the total potential energy of the system.

A phase space point can be considered as a particular realization of the system, or a *state*. However, for any real physical system, one may never know its exact state, since it would require the complete knowledge about the position and momenta of all particles. Therefore, it is more practical to treat a state as a stochastic random variable associated with a probability density $\rho(\mathbf{p}, \mathbf{q}, t)$. Alternatively, as originally proposed by J. W. Gibbs, one could imagine of $\mathcal{N}$ independent and identical N-particle systems each having a realization according to a point in the phase space, forming a "cloud" in the phase space at the time t. For sufficiently large number $\mathcal{N}$, a continuous density of phase space points can be introduced and it is convenient to normalized the density in such way that it becomes a probability density $\rho(\mathbf{p}, \mathbf{q}, t)$, suggesting he likelihood of finding a realization of the system within a particular region in the phase space. Such a collection of the systems defined by a probability distribution ρ is referred to as an *ensemble* (or Gibb's ensemble) and members of the ensemble are referred to as *microstates* and the probability density ρ represents a *macrostate*. With the elapsing time t, the collection of the phase space trajectories starting from the initial cloud of phase space points representing the ensemble can be considered as a probability fluid that flows according to the Hamiltonian dynamics.

Note that the motions of the phase space trajectories are governed by the Hamiltonian dynamics, i.e. Eqs. 2.3 and 2.4. However, the time evolution of the probability density $\rho(\mathbf{p}, \mathbf{q}, t)$ is determined by the *Liouville equation*

$$i\frac{\partial \rho(\mathbf{p}, \mathbf{q}, t)}{\partial t} = \hat{L}\rho(\mathbf{p}, \mathbf{q}, t) \tag{2.6}$$

with the Liouville operator (a differential operator)

$$\hat{L} = -i\sum_{i=1}^{N}\left(\frac{\partial H}{\partial p_i}\frac{\partial}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial}{\partial p_i}\right) \tag{2.7}$$

In general, a macrostate may not be representative for the system depending on whether probability density is stationary and no longer changes with time. Closed (no particle exchange with the exterior of the boundaries that defines the system) and (thermally) isolated systems in nature tend to evolve towards the so-called *thermodynamical equilibrium*, in which their macroscopic properties become time-independent and the system may be fully described by a few state variables. In equilibrium, the total energy is constant and entropy reaches maximal according to the second law of thermodynamics. However, even the if system is in equilibrium with a known constant total energy, in general, one can not distinguish between different microstates with the same energy. However, if the system is *ergodic*, one can assign a probability to different microstates based on

its mechanical properties. The concepts of ergodicity has great implication in the praxis, since for ergodic systems, one will be able to construct unique equilibrium probability distributions that well agree with experiments [116]. We will discuss the concept of ergodicity in more details in Sec. 2.2.2.

2.2.2 Ergodicity

The concept of ergodicity plays an important role in the work presented in this dissertation. The term "*ergodic*" was originally coined by the physicist Ludwig Boltzmann in 1871 based on two Greek words: *ergon* (work) and *odos* (path). He introduced this term to name the *ergodic hypothesis* he proposed, which basically states that for a system with a large number of dynamical and interacting Newtonian particles in equilibrium, the time averaged properties of a single particle is the same as the ensemble average. Boltzmann assumed if the total energy is conserved, a Hamiltonian system will eventually pass every accessible point of its phase space that is in accordance with the constraints and boundary conditions.

Considered a N-particle system where the total energy is conserved, *i.e.* $H(\mathbf{p}, \mathbf{q}) = E$, a dynamical state $x = (\mathbf{p}, \mathbf{q})$ with the energy E must lie on the $6N - 1$ dimensional energy hypersurface S_E determined by the Hamiltonian H. As the state x evolves with time, it will move alone the hypersurface S_E due to the energy conservation. For an equilibrium ensemble of systems with the identical Hamiltonian H and same total energy E, one would expect that the states of the ensemble also follows an time-independent distribution on the energy hypersurface. In general, the equilibrium probability density $\rho(x)$ may or may not describe the distribution on the energy surface, *i.e.* the fraction of the microstates from the ensemble lying in the region R of the total energy hypersurface $S_R \in S_E$ is determined by the integral

$$S_R = \int_R \rho(x) dx. \tag{2.8}$$

The simplest form of $\rho(x, t) = C$ for $x \in S_E$, where C is a constant. Such an ensemble is also referred to as an *microcanonical ensemble*. Such an time-independent ensemble density $\rho(x)$ is also referred to as an *invariant ensemble*, implying the fraction of the microstates located on the region R of the energy hypersurface is time-independent as well.

For an ensemble of $\mathcal{N}$ classical dynamical N-particle systems, the equilibrium value of a quantity that is dependent on the microstates, or a phase function, $\phi(x)$, can be calculated from the *ensemble average* defined as

$$\langle \phi(x) \rangle = \int \phi(x) \rho(x) dx. \tag{2.9}$$

Often, the limit of $\mathcal{N} \to \infty$ is considered. In general $\rho(x)$ may not be the *only* invariant ensemble that has the same total energy E. In case that there is more than one $\rho(x)$, it can be problematic

since one has to chose which ensemble to use for the calculation of the equilibrium values via Eq. 2.9. An example of this ambiguity is given in Fig. 2.1.

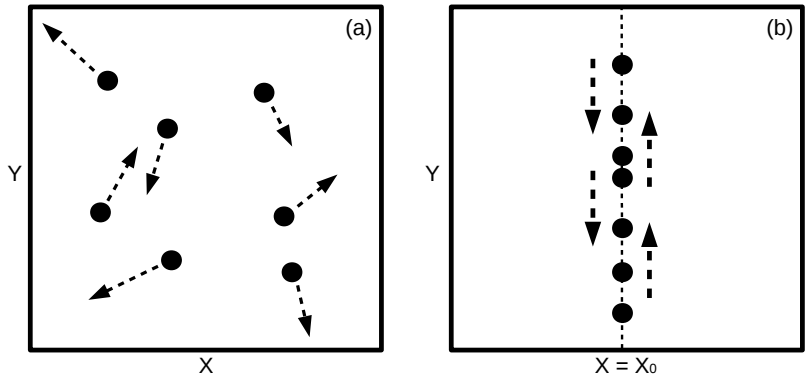

Figure 2.1: Two identical, closed and isolated systems of hard spheres with the reflective boundary conditions and the same total energy E. System (a) is a microcanonical ensemble with $\rho(x) = C$ with C being a constant. System (b) can be a realization of the same system that has an invariant ensemble on the energy hypersurface S_E but without an ensemble density.

In Fig. 2.1, system (a) represents randomly moving particles in a isolated and closed box corresponding to an microcanonical ensemble. The system (b) represents a system with perfectly aligned hard spheres that only bounce off the walls of the box and collide with each other alone a perfectly straight line. Although the probability of finding such a realization of the system is practically zero, theoretically, it can be established if the initial conditions can be initiated and realized perfectly. System (b) does not have ensemble density on the energy hypersurface, since all states will concentrate on a single point with an area of zero [74]. The question of that whether there are other invariant probability densities on the energy hypersurface S_E is equivalent to the question whether the system is ergodic.

In 1931, G. D. Birkhoff introduced the ergodicity theorem [15], in which he showed that the time average of a property $\phi(x)$ of a system, given by

$$\overline{\phi} = \lim_{\tau\to\infty} \frac{1}{\tau} \int_{t_0}^{t_0+\tau} \phi(\mathbf{p}(t), \mathbf{q}(t))dt, \tag{2.10}$$

exists for all integrable phase functions of physical interest [116]. In terms of averages, the ergodic theorem may be formulated as follows: *A system is ergodic if for all phase functions ϕ: (i) the time average, $\overline{\phi}$, exists for almost all phase space points (all but a set of measure zero), and (ii) when it exists it is equal to the ensemble phase average [116], i.e.*

$$\overline{\phi} = \langle\phi\rangle\,. \tag{2.11}$$

The phrase "almost all" here means that if there are a subset of points $M \in S_E$ on the energy hypersurface for which Eq. 2.11 is not valid, the integral over M is zero, *i.e.* $\int_M dx = 0$ [74]. This

formulation of the ergodicity is also often referred to as the *quasi-ergodic hypothesis*, since ergodic systems cannot sample every single point of the energy hypersurface, but they come very close. This is due to the fact that, for topological reasons, a trajectory obeying the canonical equations of motion cannot intersect with itself [116]. If put it in terms of lower dimensions, a one-dimensional line that cannot intersect with itself is incapable of covering a 2-dimensional surface, it will always miss some points. In this sense, Birkhoff's ergodicity theorem implies that a system is ergodic and a unique equilibrium provability density exists if almost all points on the energy hypersurface is sampled.

In practical terms, ergodicity means that, given sufficient time, a system can practically access any region of the available phase space allowed by the constraints and boundary conditions regardless of its initial condition, *i.e.* from which point in the phase space the trajectory started. Therefore the phase space trajectories of the ensemble will be similar to each other and thus the dynamical behaviors will share a high degree of similarities as well, therefore the ensemble average agrees with the time average. However, this is not guaranteed for non-ergodic systems. Certain system can have a phase space that are separated in mutually inaccessible domains. If starting from a point in one domain, it will never reach to anther domain. This type of behavior is referred to as *strong ergodicity breaking* [92]. However there are also systems, such as glasses, where the phase space dose not necessarily contain any inaccessible region *per se*, but the time required to fully sample the accessible phase space is practically infinite. This type of behavior is referred to as *weak ergodicity breaking* (WEB) [18, 92]. In both cases, the ergodicity condition given by Eq. 2.11 is violated. One direct consequence of that is the behaviors of the individual systems within the ensemble will be vastly different, despite that they are practically identical systems with the same Hamiltonian. They will not display a converged and time-independent equilibrium behavior, but rather population splitting [90, 120] regardless how long the system is being observed.

2.3 A brief historical review on diffusion and Brownian motion

Brownian motion is an important class of stochastic processes frequently applied in many different areas of science and engineering for stochastic modeling. It was also one of the earliest models for internal protein dynamics [82]. Here we dedicate a small section to review its importance and significance scientific history.

Brownian motion was named after the observations reported by the Scottish botanist Robert Brown (1773-1858) in 1828 [19]. While studying pollen grains suspended in water under a microscope, he noticed that the pollens do not stay still but carry out fast, erratic motions. Although he could not determine the cause of these motions, he was able to exclude the possibility that these

motions were caused by any "life forces"[2].

Little as Brown knew, as early as in 1807, French mathematician Jean-Baptiste Joseph Fourier (1768–1830) has already provided one important piece to the puzzle of these mysterious motions Brown observed during his studies on the heat conductance and propagation in solids, a seemingly utterly unrelated subject. In that year, Fourier introduced the much celebrated heat equation which describes the spread of the temperature increase in solids when exposed to an external heat source with a parabolic partial differential equation

$$\frac{\partial u}{\partial t} = \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right), \tag{2.12}$$

where $u = u(x, y, z, t)$ is a temperature field given in the Cartesian coordinates x, y, z at the time t.

This equation simply states that the temperature change per time unit is proportional to the second derivative of the temperature field in the spatial coordinates with the constant of proportionality α, although it was not completely clear at that time what actually heat is [99, 102]. Some thought that heat is a fluid that permeating and moving through the bodies wile others believed heat was motions and vibrations of matter at elementary level [99]. However, one could accurately measure heat in experiments with instruments such as closed-tube mercury thermometers perfected by German physicist Gabirel Daniel Fahrenheit and commercially produced in 1717, after whom the Fahrenheit temperature scale was named. It was until 1857 that German physicist Rudolf Clausius who proposed the idea that heat is a form of kinetic energy and gives rise of the motions of the molecules as he described in a paper published in 1857 [24] with the title *"Ueber die Art der Bewegung die wir Wärme nennen"* (About the type of motion we call heat), although at that time, the precise concepts of atoms and molecules were still in the infancy. One only had the rough idea that "atoms" and "molecules" are some forms of tiny, basic constituents of matters.

It was until the publication of Fourier's book *"Théorie analytique de la chaleur"* (The analytical theory of heat) [40] in 1882 that the broad scientific community started to realize the significance of the heat equation, not just for the science of heat conductance, but as a more general mathematical framework to conceptualize challenges encountered in other fields of science. Shortly afterwards, the analogy of heat conductance inspired German physicist George Simon Ohm to study the flow of electricity throw conducting matter. Around early 19th century, experimental chemists began to notice the equivalence between molecular diffusion and Fourier's heat equation which provided a great theoretical framework to interpret their experimental observations. Among many, the phenomenon of osmosis attracted much attention, which was the starting point for Albert Einstein to derive his diffusion equation in the landmark paper roughly one century later in 1905. Already a century earlier in 1752, French physicist Jean-Antoine Nollet (1700-1770) reported

[2] In the sense that pollen grains do not have an active mean of motion, such as a molecular motor driven flagella in certain bacteria

the selective movement of different liquids across an animal bladder (which is a semipermeable membrane) [99]. Between 1825-1850, French physician René Joachim Henri Dutrochet (1776-1847) and, more recognizably, Scottish chemist Thomas Graham (1805-1869) have carried out systematic experimental studies of osmosis and collected a large wealth of data. Particularly, Graham's work on diffusion of salt in water inspired and motivated German physiologist (who had also substantial trainings in mathematics and physics) Aldof Fick to derive the Fick's diffusion equation for liquids [39] by using the exact analogy to Fourier's heat diffusion equation, as Fick wrote in his paper in 1855 [39]: *"Genau nach dem Muster der Fourier'schen Entwickelung für den Wärmestrom leitet man aus diesem Grundgesetze für den Diffusionstrom die Differentialgleichung her"* (Following the pattern of Fourier's development for the heat flux one derives the differential equation as the fundamental law of the diffusion flux)

$$\frac{\partial c}{\partial t} = -k\left(\frac{\partial^2 y}{\partial x^2} + \frac{1}{Q}\frac{dQ}{dx}\frac{\partial y}{\partial x}\right). \tag{2.13}$$

Eq. 2.13, also known as Fick's second law, is mathematically equivalent to Fourier's heat equation. The molecular concentration $y = f(x,t)$ replaced the temperature field and the additional term on the left-hand side takes the changes of the area of the cross-section of the vessel into account in which the diffusion takes place (e.g. the diffusion flux through a cylindrical tube or a cone-shaped funnel). He was also the first to introduce the diffusion coefficient D as the proportionality constant between the rate of the diffusion and concentration gradient. Despite the successes, Fick's laws remain a rather phenomenological description and was often hard to verify experimentally at the time due to technical limitations [111]. From mathematical point of view, Fick's diffusion model is a continuum formulation that applies to diffusion currents of a large number of molecules, but for a single Brownian particle, or discrete atomic or colloidal systems, this model fails to provide an adequate description. On the other hand, molecular kinetic theory developed by, among others, Ludwig Boltzmann and James Clark Maxwell in the mid of 19th century have successfully described the macroscopic, collective behavior of gases under the explicit assumption that gases are composed of molecules, which are tiny, discrete rigid bodies that move according to the Newtonian laws of motion and collide with each other in an elastic fashion. In 1889, French physicist Louis Georges Gouy published the results from a series of carefully conducted experiments on Brownian motion in different liquids [111]. He noticed that motion is independent from external forces, such as vibration or electromagnetic field but depends on the viscosity of the liquid in which the test particles are suspended, *i.e.* the lower the viscosity the liquid, the higher the intensity of the motion. He concluded that the Brownian motion is a visible manifestation of the intrinsic thermal molecular motions [111].

It was Albert Einstein who unified these two approaches and provided a physical description for the diffusion of the pollen grain in water as observed by Brown and coined the term "Brownsche

Bewegung" or "Brownian motion" in his 1905 landmark paper [34]. The same description was also proposed independently by Marian von Smoluchowski in 1906 [134]. Einstein assumed that the stochastic motion of the larger Brownian particle is caused by the continuous random collisions with the much smaller solvent molecules from all directions. Mathematically, the so-called Einstein-Smoluchowski picture of the Brownian motion is a continuum description of the single particle, however, the continuous quantity here is not anything physical related to the actual particle but its probability density function (PDF) $P(x,t)$ of finding the diffusing particle at position x at time t, as stated by the Einstein-Smoluchowski diffusion equation

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2} \tag{2.14}$$

Eq. 2.14 has the same form as Fourier's heat equation, were the PDF replaces the temperature field with the proportionality constant D as the diffusion constant. More importantly, from the dynamical equilibrium of the Brownian particle suspended in solvent, Einstein (and independently by Smoluchowski in 1906) derived the Stokes-Einstein relation that can be written as

$$D = \frac{RT}{N_a}\frac{1}{6\pi\eta a} \tag{2.15}$$

where a is radius of the Brownian particle, η is solvent viscosity, R is the ideal gas constant and N_A the Avogadro's number. One can easily see that the diffusion constant D is only dependent on the solvent viscosity and size of the Brownian particle, all the other terms involved are fundamental natural constants, exactly as observed by Gouy in his experiments. It is noteworthy that Eq. 2.15 is only valid in the limit of small Reynold number, *i.e.* when the inertial force on the solute is much greater than the viscous force posed by the solvent. In this case, the friction constant ς, also often referred to as the drag, or damping constant, is given by $\varsigma = \gamma m \approx 6\pi\eta a$, where γ is the friction coefficient and m the mass of the Brownian particle. In general the diffusion constant can be written as

$$D = \frac{k_B T}{\gamma m}. \tag{2.16}$$

In 1908 shortly after Einstein and Smoluchowski, French physicist Paul Langevin introduced the Langevin equation – an intuitive, Newtonian-equivalent formulation for the Brownian motion [73].

$$m\frac{dv}{dt} = -6\pi\eta a v + F_r(t). \tag{2.17}$$

where v is the velocity of the Brownian particle, $F_r(t)$ is a fluctuating Gaussian random force with $\langle F_r(t)\rangle = 0$ and the spectral property of a white noise, *i.e.*,

$$\langle F_r(t)F_r(t')\rangle = 2D\delta(t-t') \tag{2.18}$$

with the diffusion constant D and $\delta(t-t')$ is Dirac's delta function. This equation simply states that the net force acting on the Brownian particle is the result of the competing frictional force and

the random forces exerted by the solvent molecules. Eq. 2.18 is also referred to as the *dissipation-fluctuation relation*, implying that the *systematic* part of the microscopic forces, *i.e.* the frictional force is determined by the correlation of the random force, and conversely the random forces also have to satisfy this relation and therefore dependent on the frictional forces [69]. Unlike the classic Newtonian equations of motion, which are ordinary differential equations, Eq. 2.17 is a stochastic differential equations. An equivalent stochastic differential equation for Brownian motion was already introduced by French mathematician Louis Bachelier in his doctoral thesis with the title *"Théorie de la spéculation"* (Theory of speculation) in 1900 to model the stochastic fluctuations of stock and bond prices [6]. Bachelier's work has been frequently considered as the birth of financial mathematics.

The significance of Einstein's diffusion theory lies beyond the mere explanation of the Brownian motion. Until the early 20th century, the atomic theory which states that all matter are constituted of discrete elementary particles, *i.e.* atoms - a theory originally proposed by the English chemist John Dalton roughly a century earlier, was not generally accepted in the scientific community at the time. It is until French physicist Jean Baptiste Perrin experimentally verified Einstein's diffusion theory, which is explicitly formulated based on the assumption of discrete "atomic" particles. The always broken "zig-zag" and self-similar diffusion trajectories obeying the Einstein's diffusion theory observed by Perrin was considered as the definitive evidence for the existence of discrete atoms and molecules. In 1926, Perrin was awarded with the Nobel price in physics for settling the century-long debate over Dalton's atomic theory.

2.4 Relationship between Brownian motion and random walk

Random walk is often used interchangeably with Brownian motion in the literature. The term "random walk" was coined by the famous English biostatistician Karl Pearson. In 1905, roughly 2 month after Einstein published his paper on Brownian motion, Pearson submitted a letter to the magazine *Nature* with the title "The Problem of the Random Walk" [109]. The letter itself is not a scientific paper but rather a call for an answer for the following problem, as Pearson wrote: *"A man starts from a point O and walks l yards in a straight line; he then turns through any angle whatever and walks another l yards in a second straight line. He repeats this process n times. I require the probability that after these n stretches he is at a distance between r and $r + \delta r$ from his starting point, O."* Unknown to Pearson at the time, the answer to this question was already answered by Louis Bachelier 5 year earlier in his doctoral thesis. Nevertheless, Pearson did receive several answers from the readers, among them an response from Lord Rayleigh (John W. Strutt) who provided a correct solution using an equivalent mathematical formulation he studying the problem of estimating the amplitude and intensity of the resultant of the mixing of n vibrations of the same

period and amplitude but of arbitrarily chosen phase [99]. He demonstrated the probability $P(r,t)$ of finding the random walker at position r at time t obeys a differential equation that has the same form as Fourier's heat equation (or Eq. 2.14) and the solution to that equation is a Gaussian centered at the starting point of the walker. As Pearson responded to Lord Rayleigh: "*The lesson of Lord Rayleigh's solution is that in open country the most probable place to find a drunken man who is at all capable of keeping on his feet is somewhere near his starting point!*" [99]

Einstein's diffusion equation can be derived from the random walk in the continuous limit in the following way. Assuming an one-dimensional Pearson's random walk along the coordinate x. After each time step Δt, the walker will move one step $\pm\Delta x$ and the probability of finding the walker at position x the time t is $W(x,t)$. We further assume that the walk process is fully Markovian, *i.e.*, the probability of finding the walker at the position x and time t *only* depends on the position of the previous position at time $t-\delta t$, $W(x,t)$ obeys the master equation

$$W(x,t+\Delta t) = \frac{1}{2}W(x+\Delta x,t) + \frac{1}{2}W(x-\Delta x,t). \tag{2.19}$$

The factor 1/2 suggests the probability for the walker to come from $x+\Delta x$ or $x-\Delta x$ prior arriving at the current position x is equal.

The Taylor expansion of Eq. 2.19 around Δt and Δx in the continuum limit of $\Delta x \to 0$ and $\Delta t \to 0$ yields

$$W(x,t+\Delta t) = W(x,t) + \Delta t\frac{\partial W(x,t)}{\partial t} + \mathcal{O}(\Delta t^2) \tag{2.20}$$

and

$$W(x\pm\Delta x,t) = W(x,t) \pm \Delta x\frac{\partial W(x,t)}{\partial x} + \frac{(\Delta x)^2}{2}\frac{\partial^2 W(x,t)}{\partial x^2} + \mathcal{O}(\Delta x^3), \tag{2.21}$$

respectively. Inserting Eqs. 2.20 and 2.21 into Eq. 2.19, one obtains the same equation as Einstein's diffusion equation

$$\frac{\partial W(x,t)}{\partial t} = D\frac{\partial^2 W(x,t)}{\partial t^2} \tag{2.22}$$

with a positive constant

$$D = \frac{(\Delta x)^2}{2\Delta t} \tag{2.23}$$

as the diffusion constant.

Later on, Pearson's random walk is further generalized by the works of, among others, E. Montroll, H. Scher, G. H. Weiss and M. Schlesinger [98, 118, 119, 138] to the so-called *continuous time random walks*. In which the fixed walk step length and constant time step are replaced by a random numbers drawn from certain distributions. Depending on the PDFs of these distributions, the resulting random walks can have vastly different properties and Brownian random walk is only a special case among the continuous space of different random walks. This subject remains a very active research area until the present day.

Random walk illustrates a generic concept and as well as a universality shared by a variety of different and unrelated phenomenons. As in the later chapters in this dissertation, we will see that a particular type of random walk can be used model the thermal fluctuation of protein structure.

2.5 From normal to anomalous diffusion

One most remarkable achievement brought by the theory of Brownian motion is the predictability of seemingly random and stochastic processes. Although it is impossible to tell, at any given time, where the next step of the random walker will be, but one can tell with great certainty about the *probability* $p(x+dx) = \int P(x,t)dx$ of finding the walker somewhere between x and $x+dx$ at the time t, and as well as the spread of an ensemble of such walkers given by the *mean squared displacement* (MSD).

The solution to the diffusion equation Eq. 2.14 is a normalized Gaussian [92]

$$P(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{2.24}$$

assuming the initial position of the walker $x(t=0) = 0$.

One important realization made by Einstein in his work on Brownian is that the average position or velocity of the walker do not yield much useful information to describe the stochastic process. A more meaningful quantity for characterization of the diffusion is the so-called mean squared displacement (MSD) $\langle x^2(t)\rangle$ or the second moment of the PDF $P(x,t)$. By solving the integral $\langle x^2(t)\rangle = \int x^2 P(x,t)dt$, one obtains the relation

$$\left\langle x^2(t)\right\rangle = \frac{1}{2}Dt. \tag{2.25}$$

Eq. 2.25 is often referred to as the Einstein-Smoluchowski relation and illustrate a trademark properties of the Brownian diffusion, *i.e.* the MSD increases with the time t in a linear fashion and in thus the diffusion constant can be considered as the velocity at which the variance of the distribution increases with the time. This linear behavior is a direct result from the central limit theorem and Markovian nature of the stochastic process. However, there are many other random processes in the nature or in many other systems whose trajectories, on the first glance, resembles a Brownian motion but deviate significantly from the linear scaling of MSD as a function of time. These processes are referred to as anomalous diffusions.

2.5.1 Anomalous diffusion and complex system

Anomalous diffusion is frequently observed in complex systems. These systems are characterized by a large number of heterogeneous elementary units, strong coupling and interactions between

these units. Examples of complex systems occur in many different and diverse fields of studies, such as glasses, polymers, semi-conductors, proteins, biological organism or socioeconomic systems such as markets or macroeconomics. Due to the complexities, various underlying assumptions for Brownian motions may be violated, for example, the Markovian nature of the stochastic process, large timescale difference between the motion of the Brownian particle and the solvent molecules such that the successive collisions between these are statistically independent, and as well as the spatial isotropy of the displacements. As result, the central limit theorem can no longer be applied to these processes but replaced by the Lévy-Gnedenko generalized central limit theorem [93]. In this case, the PDF for the underlying anomalous diffusion obeys a distribution from the family Lévy-stable distributions, from which the Gaussian is a special case in the family.

In general, the time-dependence of MSD of dynamical processes can be described by a power-law

$$\text{MSD} \propto t^{\alpha} \tag{2.26}$$

with the exponent $\alpha > 0$. Evidently, for Brownian motion the exponent α adopts the unity. Diffusion processes with $0 < \alpha \leq 1$ are referred to as subdiffusion, since the MSD increases slower with the time comparing to regular Brownian motion. Diffusions with $\alpha > 1$ are generally referred to as superdiffusion. Superdiffusions can be further divided into different categories; $\alpha = 2$ is referred to as ballistic process, in which the object moves with a constant velocity like a billiard ball or a bullet fired from a rifle if neglecting the frictional and gravitational forces. Processes with $1 < \alpha < 2$ are often referred to as sub-ballistic processes, since they are somewhere between a Brownian particle and a bullet. In the present dissertation we concentrate on the subdiffusive motions which were observed in the protein structural fluctuations.

2.5.2 Ensemble and time-averaged mean squared displacements

Before further discussions of physical mechanisms and modeling of the anomalous diffusion, we first introduce the definitions of two different ways to define MSD. For simplicity's sake, we consider an ensemble of N particles performing one-dimensional motion each characterized by a trajectory $x(t)$. In this dissertation, we use the angular brackets $\langle A \rangle$ to denote *ensemble average* and overline $\overline{A}$ for the *time average*.

The ensemble averaged MSD (EA-MSD) is simply the second moment of the probability density function (PDF) $P(x,t)$ describing the ensemble

$$\left\langle x^2(t) \right\rangle = \int_0^{\infty} x^2 P(x,t) dx \tag{2.27}$$

which is a function of the running time t of the stochastic process. Often in experiments or numerical simulations, one directly obtain N individual the time series of $x(t)$ and the PDF is not known *a*

prior. In this case, the ensemble averaged MSD can be calculated as the following

$$\left\langle x^2(t) \right\rangle = \frac{1}{N} \sum_{i=1}^{N} x_i^2(t) \tag{2.28}$$

Alternatively, for a single particle, one can defined the *time averaged* MSD (TA-MSD) by averaging the displacements of the particle over a sliding time window. In this case, the TA-MSD is a function of the *lag-time* Δ, given by

$$\overline{\delta^2(\Delta)} = \frac{1}{t_{\max} - \Delta} \sum_{t'=0}^{t_{\max}-\Delta} \left[x(t' + \Delta) - x(t') \right]^2 , \tag{2.29}$$

where $t_{\max}$ is the total length of the experimental observation time or simulation trajectory. If an ensemble of N independent trajectories are available, each TA-MSD calculated from one of the trajectories of the ensemble can be considered as a stochastic (multi-dimensional) random variable and analytical models often fit to the additionally averaged TA-MSDs, defined as the average over N independent times series, *i.e.*

$$\left\langle \overline{\delta^2(\Delta)} \right\rangle = \frac{1}{N} \sum_{i=1}^{N} \overline{\delta_i^2(\Delta)}. \tag{2.30}$$

For ergodic processes, such as Brownian motions in an harmonic potential, the EA-MSD is equal to the TA-MSD, *i.e.*

$$\left\langle x^2(t) \right\rangle = \left\langle \overline{\delta^2(\Delta)} \right\rangle . \tag{2.31}$$

Deviations from Eq. 2.31 can be considered indicative for weak ergodicity breaking [92].

2.6 Modeling the structural dynamics of globular proteins using anomalous diffusion

Phrases such as "diffusion" or "Brownian motion" are intuitively associated with transport processes such as the movement of a pollen grain in water. However, this concept can be generalized to describe any stochastic fluctuating quantities with time. Years before Einstein and Langevin, French mathematician Louis Bachelier already applied the one-dimensional Brownian diffusion to model the price movement of stocks and bonds in a market place. In this analogy, the price movement represents the trajectory of the pollen (in one dimension) and the different buy- and sell-orders placed in the market place long the trading hours are equivalent "random forces" that drives the prices up or down.

When observing the protein motion from MD simulation trajectories, the distance fluctuation between the centers-of-mass of two separate protein domains over time also manifests into a

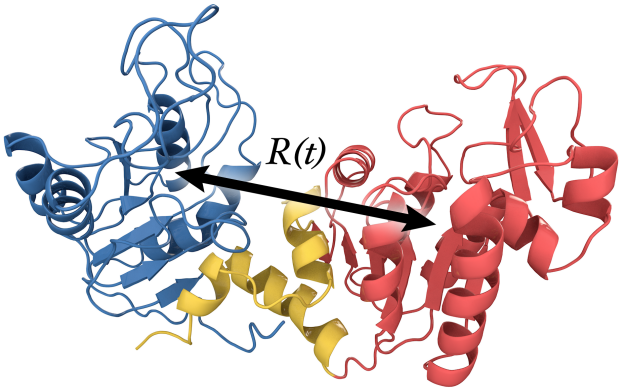

Figure 2.2: Inter-domain motion of the yeast enzyme phosphoglycerate kinase (PGK). Two protein domains, the N- and C-terminal domains, are colored in red and black, respectively. The hinge region separating both domains are colored in yellow. The arrow represents the distance $R(t)$ between the two domains.

stochastic fluctuation, resembling the one-dimensional diffusion of a fictive (see Fig. 2.2). However, this diffusion is not free and unbiased like the pollen grain suspended in water but confined within an external potential. As long as the protein remains folded in its native state, the inter-domain distances will be limited and further more there is a restoring forces pushing the walker back to a stationary position due to the intrinsic elastic properties of the protein. Early models proposed that internal domain motions can be modeled by a Brownian motion within an harmonic potential via the Langevin equation [61, 82, 107], *i.e.*

$$m\frac{d^2x(t)}{dt^2} + \gamma\frac{dx(t)}{dt} + \frac{dU(x)}{dx} = F_r(t) \tag{2.32}$$

where γ is the friction coefficient, $U(R) = \frac{1}{2}kR^2(t)$ the harmonic potential with the spring constant k and $F_r(t)$ is the random force has the spectral properties of a Gaussian white as discussed in Eq. 2.18. However, with the wealth of experimental and numerical simulation data accumulated in the last two decades [47, 52, 72, 75, 90, 95, 136, 143], especially due to the advancements made in the single-molecule spectroscopy techniques and massive parallel supercomputing, clearly demonstrated that this Brownian picture of protein dynamics is oversimplified and the anomalous dynamics is a more appropriate description of the phenomenon.

However, the modeling of anomalous diffusion processes is not as straight forward as the Brownian motion which represents the convergence of the central limit theorem. A variety of generalized versions of Einstein-Smoluchowski diffusion picture exist and they can be conceptually very different from each other [92] with profoundly different physical implications. Therefore, great care is needed when one attempts to identify the underlying mechanism for an observed subdiffusive process. In the following, we will briefly discuss two most frequently used approaches but with very different physical implications - the fractional Langevin equation (FLE) and the continuous time random walk (CTRW).

2.6.1 Fractional Langevin equation

Given $x(t)$ is the trajectory of a freely diffusing particle, the fractional Langevin equation can be written as the following [92]

$$m\frac{d^2x(t)}{dt^2} = -\gamma_\alpha \int_0^t (t-t')^{\alpha-2}\left(\frac{dx(t')}{dt'}\right)dt' + A\xi_{\text{fGn}}(t) \tag{2.33}$$

where γ_α is the fractional friction coefficient of the dimension g·s$^{-\alpha}$, A a fixed amplitude of the noise obeying the generalized Kubo fluctuation dissipation relation [92]

$$A = \sqrt{\frac{\gamma_\alpha k_B T}{\alpha(\alpha-1)D_\alpha}} \tag{2.34}$$

and $\xi_{\text{fGn}}(t)$ is the fractional Gaussian noise with $1 < \alpha < 2$, characterized by a standard normal distribution and a power-law correlation [92] for any $t > 0$

$$\langle \xi_{\text{fGn}}(t)\xi_{\text{fGn}}(t')\rangle = \alpha(\alpha-1)D_\alpha|t-t'|^{\alpha-2}. \tag{2.35}$$

Eq. 2.33 is special version of Häggi-Kudo generalized Langevin equation [92] and an extension of the classic Langevin equation describing the motion of a freely diffusive particle in a viscoelastic medium. Due power-law correlation of the noise $\xi_{fGn}(t)$ and convolution with the same power-law $(t-t')^{\alpha-2}$ in the frictional force term, the dynamical process has explicit dependence on the past and thus the system has long term memory (decays with the power-law). The term "fractional" refers to the fact that the convolution integral in Eq. 2.33 can written with the Caputo time fractional derivative, defining a non-integer ($2-\alpha$-th) order of differentiation in time, *i.e.*

$$\frac{d^{2-\alpha}x(t)}{dt^{2-\alpha}} = \frac{1}{\Gamma(\alpha)}\int_0^t (t-t')^{\alpha-2}\left(\frac{dx(t')}{dt'}\right) \tag{2.36}$$

and yielding

$$m\frac{d^2x(t)}{dt^2} = -\gamma_\alpha\Gamma(\alpha-1)\frac{d^{\alpha-2}x(t)}{dt^{\alpha-2}} + A\xi_{\text{fGn}}(t) \tag{2.37}$$

and in the overdamped limit, *i.e.* $d^2x(t)/dt^2 = 0$, Eq. 2.37 reduces to

$$\gamma_\alpha\Gamma(\alpha-1)\frac{d^{\alpha-2}x(t)}{dt^{\alpha-2}} = A\xi_{\text{fGn}}(t) \tag{2.38}$$

The FLE describes an ergodic subdiffusive process thus, the EA and TA-MSDs are identical for free, unbiased diffusion [92], *i.e.*

$$\left\langle x^2(\Delta)\right\rangle = \lim_{t\to\infty}\overline{\delta^2(\Delta)} = \frac{2k_BT\Delta^2}{m}E_{\alpha,3}(-\Gamma(\alpha-1)\frac{\gamma_\alpha}{m}\Delta^\alpha), \tag{2.39}$$

where $E_{a,b}(z)$ is the generalized Mittag-Leffler function

$$E_{a,b}(z) = \sum_{n=0}^{\infty}\frac{z^n}{\Gamma(na+b)} = -\sum_{n=1}^{\infty}\frac{z^{-1}}{b-na}. \tag{2.40}$$

As results, for short-time scales of $t \ll (m/\gamma_\alpha)^{\frac{1}{\alpha}}$, the MSD scales as $\langle x^2(t)\rangle \propto t^2$, indicative of ballistic dynamics [92]. On longer time scales, *e.g.* $t \gg (m/\gamma_\alpha)^{\frac{1}{\alpha}}$, the MSD scales as a power-law $\langle x^2(t)\rangle \propto t^{2-\alpha}$ with $1 < \alpha < 2$, therefore transitions into subdiffusive dynamics.

However, if confinement via boundary conditions or external potential is present, the equivalence between EA- and TA-MSDs can become slightly more complicated. As discussed in details in ref. [57], even for fully ergodic processes such as Brownian motion or FLE governed processes, the confinement can cause a transient inequivalence between EA- and TA-MSDs for fully ergodic processes such as Brownian diffusion, FLE or FBM governed motions. For these processes, the presence of confinement means that the system can relax towards a time-independent, equilibrium state, characterized by a stationary probability distribution $P_{\text{st}}(x)$ given by the Boltzmann distribution

$$P_{\text{st}}(x) = \exp\left(-\frac{V(x)}{k_B T}\right) \tag{2.41}$$

where $V(x)$ is an external confining potential. This stationary distribution also implies that EA-MSD, defined as the second moment of the distribution, will also become independent after sufficient time and reaches a constant value $\langle x^2\rangle_{\text{th}}$, often referred to as the thermal plateau. In general, the n-th moment of the Boltzmann distribution is given by

$$\langle x^n\rangle_{\text{th}} = \frac{1}{\mathcal{N}} \int_{-\infty}^{\infty} x^n \exp\left[-V(x)/(k_B T)\right] dx \tag{2.42}$$

with the normalization constant $\mathcal{N}$

$$\mathcal{N} = \int_{-\infty}^{\infty} \exp\left[-V(x)/(k_B T)\right] dx. \tag{2.43}$$

For ergodic processes such as Brownian motion, FBM or FLE governed processes, an single exponential relaxation towards the thermal plateau is expected [57]. Within the confinement of an external potential , both EA- and TA-MSDs of FLE processes show the same ballistic behavior $\propto \Delta^2$ for small Δ. For large Δ, the TA-MSD is give by [56, 57]

$$\overline{\delta^2(\Delta)} = 2\langle x^2\rangle_{\text{th}} \left(1 - E_\alpha\left[-\frac{k}{\gamma\Gamma(\alpha-1)}\Delta^{2-\alpha}\right]\right) \tag{2.44}$$

Note here that the TA-MSD converges towards $2\langle x^2\rangle_{\text{th}}$ rather than $\langle x^2\rangle_{\text{th}}$ as the case of EA-MSD.

In the limit of $\Delta \to \infty$, Eq. 2.44 can be approximated by a power-law [57]

$$\overline{\delta^2(\Delta)} \approx 2\langle x^2\rangle_{\text{th}} \left(1 - \frac{\gamma}{k\Delta^{2-\alpha}}\right), \tag{2.45}$$

indicating a slow, power-law convergence towards $2\langle x^2\rangle_{\text{th}}$. Such power-law convergence was observed observed in experiments of single particle tracking in wormlike micellar solution using optical tweezers [56]. This interesting discrepancy between EA- and TA-MSDs has many important

implications for the analysis and interpretation of experimental and simulation data. We refer to ref. [57] for more details and discussion on this subject. Besides various transport processes, FLE was applied to model the structural dynamics of globular protein under physiological conditions obtained from *in vitro* single-molecule spectroscopy experiments on the extremely long observation time limit up to several minutes [95].

2.6.2 Continuous time random walk

Continuous time random walk (CTRW) is a generalization of the Pearson's random walk, originally introduced by E. Montroll, M. Schlesinger and H Scher [98, 118, 119] to described the anomalous diffusion of charge carriers in amorphous semi-conductor. For each step in the Pearson's random walk, the walker moves a fixed walk step size δx with in a fixed time intervals, or time step δt, which can be considered as an instantaneous jump of the constant length δx after a fixed waiting period δt. In the CTRW, δx and δt are no longer constants but continuous, stochastic random variables distributed according to the PDFs $\phi(\delta x)$ and $\xi(\delta t)$.

An important feature of the CTRW is its renewal character [92, 93], *i.e.* at each time step, a new pair of δx and δt are randomly selected from their respective distributions and it is independent for each step. Therefore the CTRW itself still has the Markovian character, however, depending on the properties of the PDFs determining the jump lengths and waiting times, the resulting dynamical process $x(t)$ may or may not be Markovian.

CTRW can produce subdiffusive, normal or superdiffusive random walk trajectories, depending on the properties of $\phi(\delta x)$ and $\psi(\delta t)$. One decisive criterion is whether the mean waiting time $\langle \delta t \rangle$ and the variance of the jump length $\langle \delta x^2 \rangle$ given by

$$\langle \delta t \rangle = \int_0^\infty \delta t \phi(\delta t) d(\delta t) \tag{2.46}$$

and

$$\left\langle \delta x^2 \right\rangle = \int_0^\infty \delta x^2 \psi(\delta x) d(\delta x). \tag{2.47}$$

For any distribution where the $\langle \delta t \rangle$ and $\langle \delta x^2 \rangle$ are finite, the resulting dynamics will be a Brownian motion with the diffusion constant $D = \langle \delta x^2 \rangle / (2 \langle \delta t \rangle)$. For any diverging mean waiting time and jump length variance, the resulting processes will be governed by the Lévy-Khintchine generalized central limit theorem, the resulting PDF for such process is a Lévy-stable distribution [93].

The CTRW is subdiffusive if the jump length follows a distribution with finite variance, while the waiting times following a fat-tail PDF of which the first moment is infinite, such as a asymptotic power-law [92, 93]

$$\psi(\delta t) \propto \frac{\delta t_0^\alpha}{\delta t^{1+\alpha}} \tag{2.48}$$

in the limit of $\delta t \to \infty$ and $0 < \alpha < 1$. δt_0 is a scaling factor associated with some characteristic time scale of the process. In this case, the anomalous diffusion constant D_α is given by

$$D_\alpha = \frac{\langle \delta x^2 \rangle}{2\delta t_0^2}, \tag{2.49}$$

thus it no longer has the dimension of m^2/s as in the Brownian case, but dimension m^2/s^α.

Free, unbiased CTRW

In the Einstein-Smoluchowski picture (in the diffusion limit), the free, unbiased subdiffusive CTRW is governed by the fractional diffusion equation (FDE) [92]

$$\frac{\partial}{\partial t} P_\alpha(x,t) = D_\alpha {}_0\mathcal{D}_t^{1-\alpha} \frac{\partial^2}{\partial x^2} P_\alpha(x,t) \tag{2.50}$$

where ${}_0\mathcal{D}_t^{1-\alpha}$ is the Riemann-Liouville fractional operator

$${}_0\mathcal{D}_t^{1-\alpha} P_\alpha(x,t) = \frac{1}{\Gamma(\alpha)} \frac{\partial}{\partial t} \int_0^t \frac{P_\alpha(x,t')}{(t-t')^{1-\alpha}} dt', \tag{2.51}$$

where $\Gamma(\alpha)$ is the gamma function. For $\alpha = 1$, Eq. 2.50 becomes the regular Einstein diffusion equation. The closed form of the solution for Eq. 2.50 is the Fox H-function [93]. If an external potential $V(x)$ is present, the propagator for the subdiffusive CRTW $P_\alpha(x,t)$ is determined by the fractional Fokker-Planck equation (FFPE) [91, 93]

$$\frac{\partial}{\partial t} P_\alpha(x,t) = {}_0\mathcal{D}_t^{1-\alpha} \left(\frac{\partial}{\partial x} \frac{V(x)}{m\gamma_\alpha} + D_\alpha \frac{\partial^2}{\partial x^2} \right) P_\alpha(x,t) \tag{2.52}$$

where γ_α and D_α are the generalized friction coefficient and diffusion constant, respectively.

Although both FLE and FDE/FFPE both describe subdiffusive behavior, however, there is a crucial difference. Namely, FDE and FFPE describe a non-ergodic, non-equilibrium process while the FLE is fully ergodic. For FDE and FFPE, the EA-MSD and TA-MSD are different from each other and exhibit some very interesting properties. The EA-MSD for unbiased subiffusive CTRW (Eq. 2.50) is given by

$$\langle x^2(t) \rangle = \frac{2D_\alpha}{\Gamma(\alpha+1)} t^\alpha \tag{2.53}$$

with $0 < \alpha < 1$. While the TA-MSD (with additional averaging over independent trajectories), for $\Delta \ll t$, is given by

$$\left\langle \overline{\delta^2(\Delta)} \right\rangle \sim \frac{2D_\alpha}{\Gamma(\alpha+1)} \frac{\Delta}{t^{1-\alpha}}, \tag{2.54}$$

displaying a linear increase with the lag-time Δ, which resembles that Brownian motion, although the underlying physics is completely different. Therefore, in studies for which only a small number of trajectories are available, caution is advised and further analysis are needed to makes sure that the observed linear MSD is indeed from a Brownian random walk.

Confined CTRW

Due to the non-ergodic and non-equilibrium properties of the subdiffusive CTRW, the thermal plateau in the MSD will be never reached, despite the spatial confinement. The EA-MSD for subdiffusive CTRW within confinement is given by [91]

$$\left\langle x^2(t)\right\rangle = \left\langle x^2(t)\right\rangle_{\text{th}} \left(1 - E_{\alpha,1}\left[-(t/\tau_c)^{\alpha}\right]\right) \tag{2.55}$$

where $E_{\alpha,1}(z)$ is the Mittag-Leffler function (Eq. 2.40) and τ_c a characteristic time scale of the system, assuming $\langle x^2(0)\rangle = 0$. The Mittag-Leffler function interpolates between a stretched exponential and a power-law [93], *i.e.*

$$E_{\alpha,1}(-t/\tau_c) \propto\sim \exp\left[-(t/\tau)^{\alpha}\right] \quad \text{for} \quad t \ll \tau_c \tag{2.56}$$

and a power-law

$$E_{\alpha,1}(-t/\tau_c) \propto \frac{1}{\Gamma(1-\alpha)} \left(\frac{t}{\tau_c}\right)^{-\alpha} \quad \text{for} \quad t \gg \tau_c. \tag{2.57}$$

Thus due to the power-law relaxation, the thermal plateau $\langle x^2(t)\rangle$ is only reached for $t \to \infty$ [91]. In the limit of $\Delta \ll t$, the TA-MSD of the subdiffusive CTRW is given by [22]

$$\left\langle \overline{\delta^2(\Delta)} \right\rangle \sim \left(\left\langle x^2\right\rangle_{\text{th}} - \langle x\rangle_{\text{th}}^2\right) \frac{2\sin(\pi\alpha)}{\pi\alpha(1-\alpha)} \left(\frac{\Delta}{t}\right)^{1-\alpha} \tag{2.58}$$

clearly deviates from its ensemble averaged counter part. The power-law scaling with $0 < \alpha < 1$ indicates the TA-MSD will only approach the thermal plateau asymptotically, and as long as $\Delta \ll t$, the TA-MSD never exceeds the value $\left\langle x^2\right\rangle_{\text{th}} - \langle x\rangle_{\text{th}}^2$.

Chapter 3

Methods

In this chapter details of the numerical techniques used to generate the simulation data for this dissertation and as well as some analysis methods are discussed. The main focus here remains on the molecular dynamics (MD) simulation of bio-macromolecules. We first provide a primer on theories behind MD simulation, followed by details about the potential energy force field for bio-macromolecules and the algorithm for the integration of the Newtonian equations of motion.

3.1 Molecular dynamics simulation

Generally speaking, molecular dynamics (MD) is computational method for study of the time evolution of a classical N-particle systems by solving the Newtonian equations of motion for each individual particle numerically. In this way, one can gain the knowledge about the collective behavior of the individual atoms and as well as the "macroscopic" ensemble behavior of the whole system. Ever since the first notable application by A. Rahman in which the MD simulation was used to study the correlated atmoic motions in liquid argon in 1964 [113], and thanks to the advancement in computing algorithms and computer hardwares in the past decades, MD simulations have evolved to a complex and sophisticated numerical method that can study a large numbers of different systems, ranging from complex biological macromolecules to galaxy formation.

Although the molecular motions are ultimately governed by the laws quantum mechanics, however due to the environmental noise, quantum effect can often be approximated by its classical counterpart [102]. In MD simulation for biomolecules, the Born-Oppenheimer approximation is assumed [17] and a molecule is represented by hard spheres at the positions of the nucleus, where

each hard sphere is assigned an electric partial charge. The validity of the Born-Oppenheimer approximation is based on the fact that the electrostatic forces between the electrons and nuclei are on the same orders of magnitude due to the similar charges (on the same order of magnitude) they possess, therefore, they have similar momenta as well. However, the huge difference in their masses (on the order of factor $\sim 10^4$) lead to large difference in their velocities such that the electrons are moving with a speed many orders of magnitude higher comparing those for the nuclei. Therefore they can be assumed to follow the motion of the nuclei instantaneously and always in the ground state (instantaneous relaxation).

If the interaction potential energy $U(\mathbf{x})$ between the atoms are know, the forces can be calculated via

$$M\frac{d\mathbf{x}}{dt} = -\nabla U(\mathbf{x}) \tag{3.1}$$

where $\mathbf{x}$ is the $3N$-dimensional atomic coordinates vector and M is a $3N \times 3N$ diagonal matrix containing the atmoic masses, i.e. $M_{ij} = m_j$ for $i = j$ and 0 if $i \neq$j. Given the initial positions $\mathbf{x}_0$ and velocities $\mathbf{v}_0$, the numerical integration of Eq. 3.1 yields the time evolution of positions and velocities $\mathbf{x}(t)$ and $\mathbf{v}(t)$ which are referred to as MD trajectory. When performing MD simulation of biomacromolecules, such as a protein, the initial position coordinates of protein atoms are mostly provided by the x-ray crystal structures. The velocities can be chose as zero or randomly assigned based on the Boltzmann distribution. In the following, some details on the potential interaction force field and numerical integration algorithm are provided.

3.2 The interaction force field for bio-macromolecules

A core component of MD simulation is the empirical *force field* describing the interaction potential energy between the atoms. In this section, we will base our discussion on the CHARMM force field [13, 59, 83] as an representative example, which is also the force field used to carry out MD simulations in this dissertation.

As discussed earlier, individual atoms are represented by charged mass points in accordance with the Born-Oppenheimer approximation. However, the atoms are not completely free but coupled to each other via different types of constraints, such as bonds between pairs or angles between triplets of atoms, etc., to ensure the correct molecular structure. In typical the force fields for bio-macromolecules, the total potential energy of inter-atomic interactions, $U(\mathbf{x})$, is divided into two components, a *bonded* and a *non-bonded* term, *i.e.*

$$U(\mathbf{x}) = U_{\text{bond}}(\mathbf{x}) + U_{\text{non-bond}}(\mathbf{x}) \tag{3.2}$$

The bonded term considers interactions between any group of up to five atoms that are connected to each other via covalent bonds in the molecular topology. In CHARMM, the energy function for

the bonded term is given by

$$U_{\text{bond}}(\mathbf{x}) = \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\phi[1 + \cos(n\phi - \delta)] + \sum_{\text{improper}} K_\omega(\omega - \omega_0)^2 + \sum_{\text{cross-terms}} K_{n,m}[1 + \cos(n\Phi + m\Psi - \delta_{n,m})] \tag{3.3}$$

An schematic illustration for each different terms in Eq. 3.3 are shown in Fig. 3.1. For the bonded interactions, typically only up to 4-bodies interaction are (with the exception for the dihedral angle term) harmonic approximations that maintain the geometry of the molecule. For dihedral angles, a cosine function with multiple minima are used to reproduce various reproduce commnon geometrical configurations such as *cis*, *trans*, etc. In the special case of peptides, an additional 5-body interaction is introduced to improve the backbone dihedral angles Ψ, Φ between two adjacent amino acids [83].

Each individual term in the bonded interaction can be further split into a geometrical component and a force field parameter component. The geometrical component, such as bond distance or angle, can be directly calculated from the atomic coordinates in real time during the simulation, while the force field parameters give resulting forces an magnitude based on the geometry. Force field parameters, *e.g.*, harmonic constants K_b, K_Θ, etc. in Eq. 3.3 are a set of predetermined parameters and are different for different types of bonds, angles, dihedrals, etc. depending on the specific chemical environment. These parameters are usually obtained either through experiments or *ab initio* quantum mechanical calculations. All together, the bonded interaction energy terms are able to maintain an overall correct molecular geometry while allowing certain flexibility to accommodate the necessary degrees of freedom.

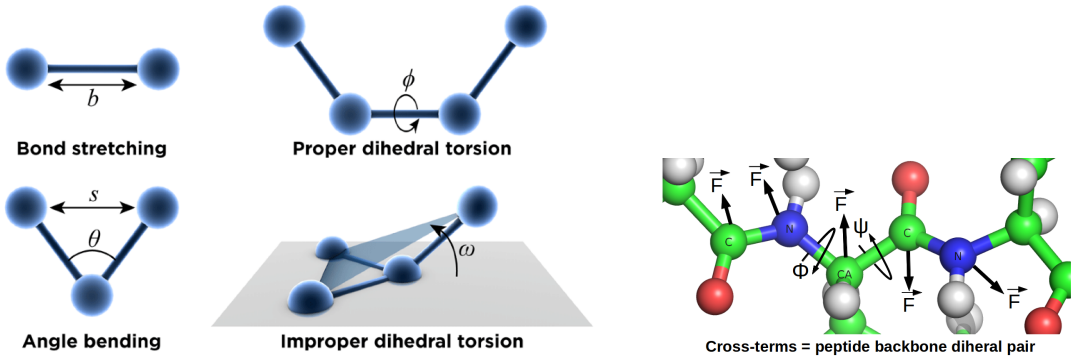

Figure 3.1: Schematic illustrations for different bonded ener1gy terms in Eq. 3.3. Left sub-figures containing the bond, angle dihedral and improper provided to the author due to the courtesy by Dr. Thomas Splettstösser (www.scifistyle.com).

The *non-bonded interactions* accounts for forces acting between all pairs of atoms, without the presence of covalent bonds between them. These forces are especially important for the

collective behavior of the system, such as phase transition or large scale conformational changes in proteins. Therefore, the non-bonded term is a function of the pair-wise inter-atomic distances r_{ij}. In typical force field of biomolecules, only Lennard-Jones (LJ) (also referred to as the van-der-Waals interactions) and electrostatic interactions are considered. For example, the non-bonded term in CHARMM force field is given by

$$U_{\text{non-bond}}(r_{ij}) = \sum_i \sum_{j,i\neq j} \left[\underbrace{\epsilon[(r_{ij}^{\min}/r_{ij})^{12} - (r_{ij}^{\min}/r_{ij})^{6}]}_{\text{Lennard-Jones}} + \underbrace{\frac{q_i q_j}{\varepsilon_1}\frac{1}{r_{ij}}}_{\text{Electrostatic}} \right]. \tag{3.4}$$

Due to the pair-wise nature of these forces, the non-bonded interactions cannot be computed locally on a parallel computer, since in order to compute the total force on one atom allocated to a particular computer processor would require the knowledge of the positions of, in theory, all the other atoms allocated to other processors, therefore requiring significant data exchange and communication between different processors, causing a bottle neck for the overall MD simulation performance. In the praxis, the LJ-interactions are calculated based on cut-offs around 1.0 to 1.2 nm due to its fast decay $\propto r^{-6}$. For electrostatic forces scaling with $\propto r^{-2}$, a twin-range approach is usually applied. On short-range within 1.0-1.2 nm, pair-wise Coulombic interactions are considered. Beyond the short range, particle mesh Ewald method [29, 37] is usually applied to account for the long range electrostatic interaction.

3.3 Numerical integration

Once the potential energy $U(\mathbf{x})$ is properly defined, the forces acting on each individual atoms can be calculated from the first derivative of the potential energy

$$\mathbf{F} = -\nabla U(\mathbf{x})$$

yielding the Newtonian equations of motion

$$\frac{d\mathbf{v}}{dt} = \frac{\mathbf{F}}{m} \quad \text{and} \quad \frac{d\mathbf{r}}{dt} = \mathbf{v}.$$

The time evolution of the positions and velocities of all atoms in the simulation systems can be obtained by solving the Newtonian equations of motion numerically. Here, we provide an example for a single particle system in one-dimension described by the position $x(t)$ and velocity $v(t) = dx(t)/dt$ for the simplicity. The same approach can be easily generalized to N-dimensional systems.

A common numerical integration scheme used in the MD simulation is the so-called velocity Verlet algorithm [50], a modified version of the original Verlet algorithm [133] named after its

inventor Loup Verlet. Assuming initial conditions $x_0 = x(0)$ and $v_0 = v(0)$, the position and velocity at the time $t + \Delta t$ can be approximated by the corresponding second order Taylor expansions

$$x(t + \Delta t) = x(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 \tag{3.5}$$

$$v(t + \Delta t) = v(t) + \frac{a(t) - a(t + \Delta t)}{2}\Delta t \tag{3.6}$$

where $v = dx/dt$ is the velocity and $a = d^2x/dt^2$ the acceleration, given Δt is sufficiently small (typically $\propto 10^{-15}$ s). At each time step, *i.e.*, $t = \Delta t$, $t = 2\Delta t$, $t = 3\Delta t$ ..., the following operations are carried out repeated throughout the entire simulation:

1. Calculate the velocity at $\frac{1}{2}\Delta t$: $v(t + \frac{1}{2}\Delta t) = v(t) + \frac{1}{2}a(t)\Delta t$
2. Calculate the temporal position $r'(t + \Delta t) = x(t) + v(t + \frac{1}{2}\Delta t)\Delta t$
3. Calculate the new acceleration $a(t + \Delta t) = -\frac{1}{m}\frac{dU(r'(t+\Delta t))}{dr'}$
4. Calculate new velocity $v(t + \Delta t) = v(t) + \frac{(a(t)+a(t+\Delta t))\Delta t}{2}$
5. Calculate new position $x(t + \Delta t) = x(t) + v(t + \Delta t)\Delta t$

The results from steps 4. and 5. will be written into a so-called trajectory files on the hard disk which are essentially the phase space trajectory of the system as discussed earlier in Sec. 2.2.1

3.4 p-variation test to distinguish different types of subdiffusion

The p-variation test, also often referred to as the "p-var" or "p-sum" test [85, 87], is a numerical statistical test that is capable of distinguishing between different types of ergodic or non-ergodic subdiffusive processes. The concept of p-variation is based on the generalization of the total variation of a stochastic signal. Given $x(t)$ is a stochastic time series, the p-variation corresponding to $x(t)$ observed over the time interval $[0, t_{\text{max}}]$ is defined as [87]

$$V^{(p)}(t) = \lim_{n\to\infty} V_n^{(p)}(t) \tag{3.7}$$

where

$$V_n^{(p)}(t) = \sum_{i=0}^{2^n-1} \left| x\left(\min\left[(\frac{t_{\text{max}}(i+1)}{2^n}, t\right]\right) - x\left(\min\left[(\frac{it_{\text{max}}}{2^n}, t\right]\right)\right|^p, \tag{3.8}$$

is finite sum of pth powers of the increments of $x(t)$, and for $p = 1$, Eq. 3.7 becomes the total variation [87]. The p-var test takes the advantages of the fact that $V^{(p)}(t)$ adopts very different

analytical forms for different types of diffusions, such as regular Brownian motion, fractional Brownian motion or subdiffusive CTRW. For sufficiently large n, the partial sum (p-sum) $V_n^{(p)}(t)$ offers a good approximation for the true p-variation and is particularly easy to compute. In the context of this dissertation we are particularly interested in the behavior $V_n^{(p)}(t)$ for different types of diffusions within confinement.

In case of $p = 2$, the partial sum for the non-ergodic, subdiffusive CTRW is given by [85, 86]

$$V^{(2)}(t) = 2D_\alpha S_\alpha(t) \tag{3.9}$$

where D_α is the fractional diffusion constant from of the corresponding FFPE and $S_\alpha(t)$ is the inverse-time, α-stable subordinator defined as [86]

$$S_\alpha(t) = \inf\{\tau : U_\alpha(\tau) > t\} \tag{3.10}$$

where $U_\alpha(\tau)$ is a strictly increasing, α-stable Levy motion [86]. Therefore, $S_\alpha(t)$, and thus $V^{(2)}(t)$, should appear to be a monotonic increasing, discontinuous, step-like function. For an ergodic subdiffusive motion, such as fractional Brownian motion (FBM), the $V^{(2)}(t)$ increases in a linear fashion. It needs to be noted too that this test does have certain limitations in the presence of noisy. If the actual stochastic motion is masked by an external noise, such as a Brownian drift or a stationary noise such as Ornstein-Uhlenbeck stochastic fluctuation, depending on the amplitude of the noise comparing to those of the actual stochastic signal, the p-variation test may be rendered inconclusive and ambiguous, as demonstrated in ref. [55].

3.5 Compact box-burning algorithm for the estimation of the fractal dimension of a graph

In order to determine the fractal dimension of the conformational cluster transition network (see Sec. 4.8 for details), a version of the box covering algorithm, the so-called compact-box-burning (cbb) algorithm [128], was implemented. Here, we provide a brief description of this algorithm based on ref. [128]. The underlying idea of this algorithm is relatively simple, one chooses a random node on the network as the center of the box and select its neighbors with within the distance of the box size l_b through a breath-first search until the number of the nodes inside box is maximal. The algorithm can be implemented in iterative steps as the following:

(1) Choose a random node in the network which is not covered by any box and use it as the seed for the center of a new box

(2) Obtained all uncovered nodes connected to the seed node and test each node whether they are within the box size l_b of any other nodes that are currently in the box, if a node fulfills this criterion, it will be included in the box.

(3) Repeat (2) until there are no more nodes can be added to this current box

(4) Repeat (1), (2) and (3) until all nodes are covered.

This procedure can be repeated for different box length l_b. The fractal dimension of the graph d_f can be then determined by using the scaling relationship

$$\frac{N_b(l_b)}{N_v} \propto l_b^{d_f}, \tag{3.11}$$

where $N_b(l_b)$ is the number of boxes required to cover the entire graph as a function of the box size l_b and N_v is the total number of the vertices in the graph.

Chapter 4

Results

In this chapter, we present the results from a series of MD simulations of three different globular proteins, (i) phosphoglycerate kinase (PGK) from the yeast *Saccharomyces cerevisiae*, (ii) aminopeptidase N (ePepN) from the bacterium *Escherichia coli* and (iii) the human K-Ras protein. The simulations of proteins (i) and (ii) were carried by the author himself, while the MD trajectory of protein (iii) was provided to the author by courtesy of Dr. Micholas Dean Smith from the University of Tennessee for analysis. These three proteins here serve as model systems for the structural dynamics. The data presented in this chapter is independently published in ref. [52].

The central finding of the study presented in this dissertation constitutes a novel view of protein dynamics, namely the collective structural relaxation in protein appears to be a non-ergodic and non-equilibrium process, contrary to the common assumption that protein structural fluctuation under physiological conditions is in the thermodynamical equilibrium. Moreover, our data indicates that the protein structural dynamics may never reach equilibrium on time timescales of most protein functions or even on the typical lifespan of globular proteins *in vivo*. Although fluctuations of distances between atoms in folded proteins are necessarily spatially confined by its native conformation, it is conceivable that, as the timescale of observation is increased a protein may incorporate into these fluctuations slower pathways over its energy landscape therefore increase the overall timescale for the structural relaxation. The question then arises as to whether there exists at all a finite characteristic time associated with any given structural change, or maybe, instead, that the timescale on which a structural fluctuation is observed determines the apparent characteristic relaxation time for the motion that will be obtained. To examine this question we have performed molecular dynamics (MD) simulations to characterize the internal dynamics of three

globular proteins mentioned above. They were chosen due to their markedly different sizse and structures: K-Ras is a single domain protein, PGK a two-domain protein and is roughly twice the size of K-Ras, and finally the ePepN, which has four domains and roughly twice the size of PGK.

4.1 Protein structures and simulation setups

4.1.1 PGK

PGK (see Fig. 2.2) consists of N- and C-terminal domains of nearly equal mass connected by a small α-helix, which, together with roughly the last 25 residues in the C-terminal tail, is sandwiched between the two domains [12]. PGK catalyzes a reaction step in the glycolysis pathway, in which a phosphate group is transferred from 1,3-bisphosphoglycerate (3,1BPG) to adenosine-diphosphate (ADP) to produce adenosine triphosphate (ATP). 3,1BPG and ADP bind to the N- and C-terminal domains, respectively. The inter-domain motion that brings 3,1BPG and ADP close to each other facilitates the transfer of the phosphate group. Therefore, the relative motion between the N- and C-terminal domains is considered to be particularly important for the function of the protein [12, 48]. The molecular mass is about 45 kD.

All-atom MD simulations were performed on four different observation timescales (simulation lengths), *i.e.* $t = 100$ ps, 10 ns, 500 ns and 17 μs. For the 100 ps, 10 ns and 500 ns timescales, five independent production runs were performed from different initial configurations and atom velocity distributions with at least 500 ns equilibration time. For the 100 ps and 10 ns simulations coordinates of the system were saved every 1 fs and 100 fs, respectively. For the 500 ns simulations the system was saved every 10 ps. Finally, one 17 μs MD trajectory was generated after 13 μs equilibration from which the coordinates were saved every 150 ps.

All simulations were carried out starting from the same crystal structure of yeast (*Saccharomyces cerevisiae*) PGK (PDB ID 3PGK). All sub-microsecond PGK simulations were carried using GROMACS 4.5.6 [112] with the CHARMM27 force field for protein with the CMAP correction [59, 83] on a local computing cluster. The 17 μs simulation was carried out on the ANTON supercomputer [122] using the CHARMM36 force field [13] for this enzyme.

For the GROMACS simulations of PGK, the system was solvated in a cubic water box (edge length ~ 10 nm) with periodic boundary conditions (PBC) leading to a total system size of about $\sim 91,000$ atoms. For the ANTON simulation, a rectangular water box with dimensions of $9.5 \times 8.8 \times 11.5$ nm^3 was used, leading to $\sim 92,000$ atoms in total. All simulations were carried out using TIP3P [58] water model in the NVT ensemble using a Nosé-Hoover thermostat [105] at temperatures of $T = 283$ and 298 K for the GROMACS and ANTON simulations, respectively. For the GROMACS simulations, the non-bonded Van der Waals (VdW) interactions were represented

using a switch algorithm with 1 nm as the switch onset distance and 1.2 nm as the distance at which the VdW interactions reach zero. The short-range electrostatic interactions within the cut-off distance of $r_c = 12$ Å were treated as Coulombic. For distances beyond r_c, the Particle Mesh Ewald (PME) method [29, 37] was used. The default outputs of the pre-processor scripts guess_chem, refine_sigma and subboxer (included in the ANTON software package) were used to set simulation parameters such as cut-off distances, electrostatic settings and the spatial domain decomposition for the ANTON simulations.

4.1.2 K-Ras

Human K-Ras (as a single chain in the PDB ID 3GFT, previously 2PMX) is a single domain GTPase involved in cellular signaling and is particularly biomedically interesting due to its involvement in cancer development. It has a molecular mass of roughly 21 kD, roughly half that of PGK.

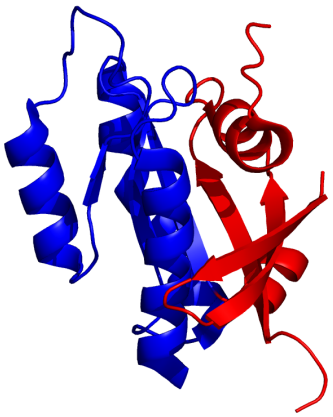

Figure 4.1: Structure of human K-RAS. The colors indicate the two segments defined, i.e. residues 1-76 (red) and residues 77-167 (blue).

Here, we use this protein as a model for studying the conformational dynamics of small, single-domain proteins. In order to examine the overall intra-domain structural dynamics in a manner similar to the inter-domain motions in the larger proteins examined here, we divided the protein into two nearly equal-sized segments, i.e. residues 1-76 and 77-167 (see Fig. 4.1), calculated the center-of-mass distance trajectories from the MD simulation and compared the results to the inter-domain distance dynamics of PGK and ePepN. Five independent 10 ns simulations (using different initial coordinates and velocities) and a single 500 ns simulation were performed in the NPT ensemble (constant number of particles, pressure and temperature) using the MD software GROMACS (version 5.0.2) [1] with the CHARMM36 force field on the Titan super-computer at Oak Ridge National Laboratory. The temperature was set to 310 K using a Berendsen thermostat [11] and the pressure to 1 bar using Parrinello-Rahman barostat [108]. The system was solvated using

the TIP3P explicit water model [58] and 8 sodium ions to neutralize the total charge. Electrostatic interactions were treated with a cut-off of 1 nm, beyond which the PME method [29, 37] was used. The VdW interactions were treated using a cutoff of 1 nm. For the 10 ns and 500 ns simulations, system coordinates were written into the trajectory file every 0.1 and 2 ps, respectively. All simulations were equilibrated for at least 0.5 μs.

4.1.3 ePepN

The *E. Coli.* aminopeptidase N (ePepN, PDB ID 2HPO) is a zinc-dependent metalloenzyme. Its primary function is protein degradation into smaller peptides at certain cleavage sites [2]. ePepN is a relatively large single-chain enzyme consisting of four distinct structural domains with a primary sequence length of 870 residues (see Fig. 4.2). The molecular mass of ePepN is ~ 101 kD, roughly twice that of PGK. Domain IV has the largest size and makes up roughly half of the total protein. Domains I and II are roughly equal in size, contributing about 20% each to the total protein mass, while domain III is the smallest with $\sim 10\%$ of the total protein mass. It has been proposed that the activity of this enzyme is related to open and closed states, involving coordinated movements of the domains [2]. Here, we use this protein as a model system for studying the inter-domain dynamics of large, multi-domain proteins.

Figure 4.2: Structure of the *E. coli.* aminopeptidase N (ePepN). The protein consists of four distinct domains; Domain I: residues 1-193 (red), domain II: residues 194-443 (blue), domain III: residues 444-545 (magenta) and domain IV: residues 546-870 (green). The orange sphere represents a Zn^{2+} ion located in the catalytic center of the protein.

We carried out 5 independent 10 ns simulations (with different initial atomic coordinates and

velocities) and a single 800 ns simulation of ePepN solvated in a cubic water box (12.4 nm edge length) using TIP3P water [58] with 0.1 M ion concentration (Na and Cl). The full system contains slightly over 180,000 atoms. All simulations were carried out on the Hopper super-computer at the National Energy Research Scientific Computing Center (NERSC) using the MD software GROMACS (version 4.6.7) [112] with the CHARMM36 force field [13] at 298 K in NVT ensemble (constant particle number, volume and temperature). The Nosé-Hoover [105] thermostat was used to control the temperature. The non-bonded Coulombic electrostatic interactions were cut off at 1.2 nm. PME method [29, 37] was used to treat electrostatic interactions beyond the cut-off. The VdW interactions were treated using a simple cutoff of 1.2 nm. All systems were equilibrated for at least 0.2μs prior data collection. We evaluated the dynamics of the inter-domain COM distance trajectory of four different domain pairs; domains I-II, II-III, II-IV and the distance between domain IV to the COM of the rest atoms of the protein, consisting of domains I, II and III.

4.2 Global collective internal protein dynamics

In this section we present the results from our analysis on global collective internal dynamics. Since full protein structural dynamics is described by a $3N$-dimensional vector $\mathbf{x}(t)$, where N is the total number of the protein atoms, certain projection of $\mathbf{x}(t)$ onto a lower dimensional coordinate is required in order to make the problem tractable. For multi-domain proteins, e.g. PGK and ePepN, we chose the distances between the center-of-mass (COM) coordinates of individual domains as the indicator for the global collective dynamics. For the single domain K-Ras protein, we chose COM coordiates of two protein segments that roughly divide the protein into two halves (see Fig. 4.1). From here on we refer to these time series of these COM-distances obtained from the MD simulations $R(t)$.

To characterize the dynamics of collective protein motion captured by $R(t)$, we use two quantities that can be directly calculated from $R(t)$: (i) the TA-MSD as defined in Eq. 2.29 and (ii) the normalized autocorrelation function (ACF) given by

$$C(\Delta) = C'(\Delta)/C'(0) \tag{4.1}$$

where

$$C'(\Delta) = \frac{1}{t_{\max} - \Delta} \sum_{t'=0}^{t_{\max}-\Delta} \delta R(t')\delta R(t'+\Delta) \tag{4.2}$$

with $\delta R(t) = R(t) - \langle R \rangle$. Analogously to the TA-MSDs, ACFs obtained from different independent MD trajectories with the same total length t can be averaged as

$$\langle C(\Delta) \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i(\Delta). \tag{4.3}$$

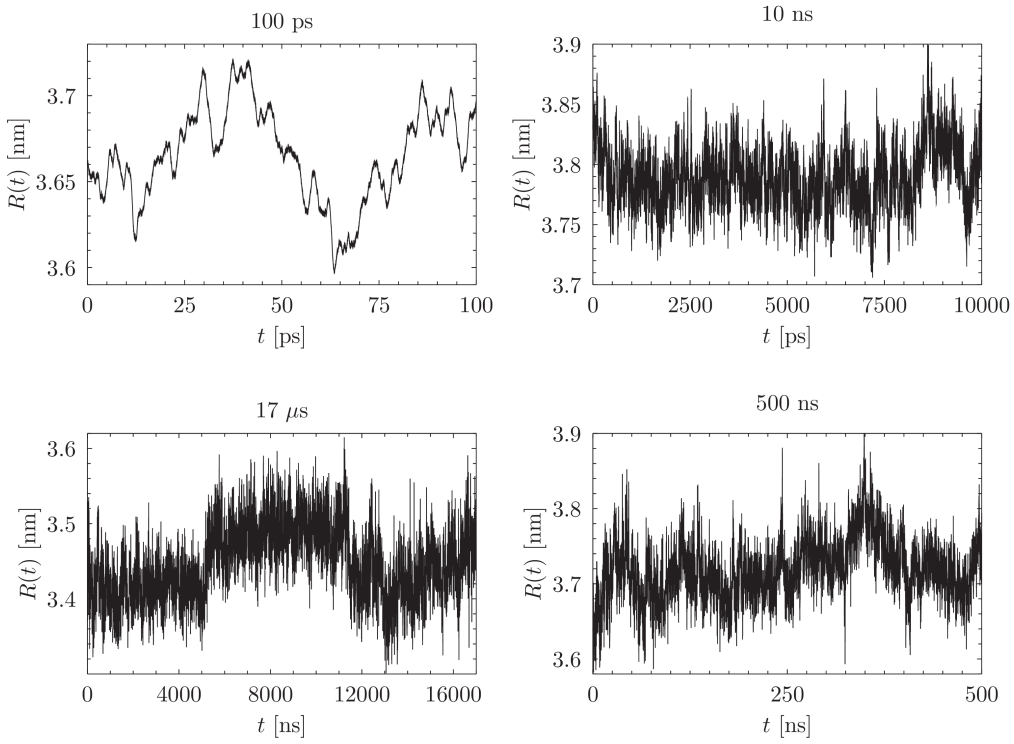

Figure 4.3: Examples of the inter-domain COM distance fluctuation of PGK observed on different timescales (total length of the trajectory).

Both quantities $\overline{\delta^2(\Delta)}$ and $C(\Delta)$ are functions of the lag-time Δ. If $R(t)$ is a fully stationary time series, the $\overline{\delta^2(\Delta)}$ can be directly related to the $C(\Delta)$ via the relationship

$$\overline{\delta^2(\Delta)} = 2\left\langle dR^2 \right\rangle [1 - C(\Delta)] \tag{4.4}$$

where $\langle dR^2 \rangle$ is the variance of $R(t)$. However, stationarity cannot be assumed here *a priori* without precaution. A general relationship between TA-MSD and ACF is derived in Appendix A. Moreover, both quantities also share a common parameter, namely, the total length of the times series t. If the protein dynamics has indeed reached the equilibrium state and the dynamical properties, such as the inter-domain distance fluctuation, are stationary, the time averaged statistical quantities such as TA-MSD and ACF should become independent from t, given that t is sufficiently large. Therefore, we have carried out MD simulations of different lengths were performed in order to study the convergence behavior.

The dynamics observed in our simulation on all three proteins exhibits the same quantitative behavior. Here, we use the two-domain protein PGK as a representative model system for the inter-domain dynamics to showcase the characteristic of the observed dynamics in greater details. For PGK, the inter-domain hinge bending motion is considered to have direct functional importance [12]. The data obtained form the analysis of the PGK inter-domain distance fluctuations are presented in Fig. 4.4.

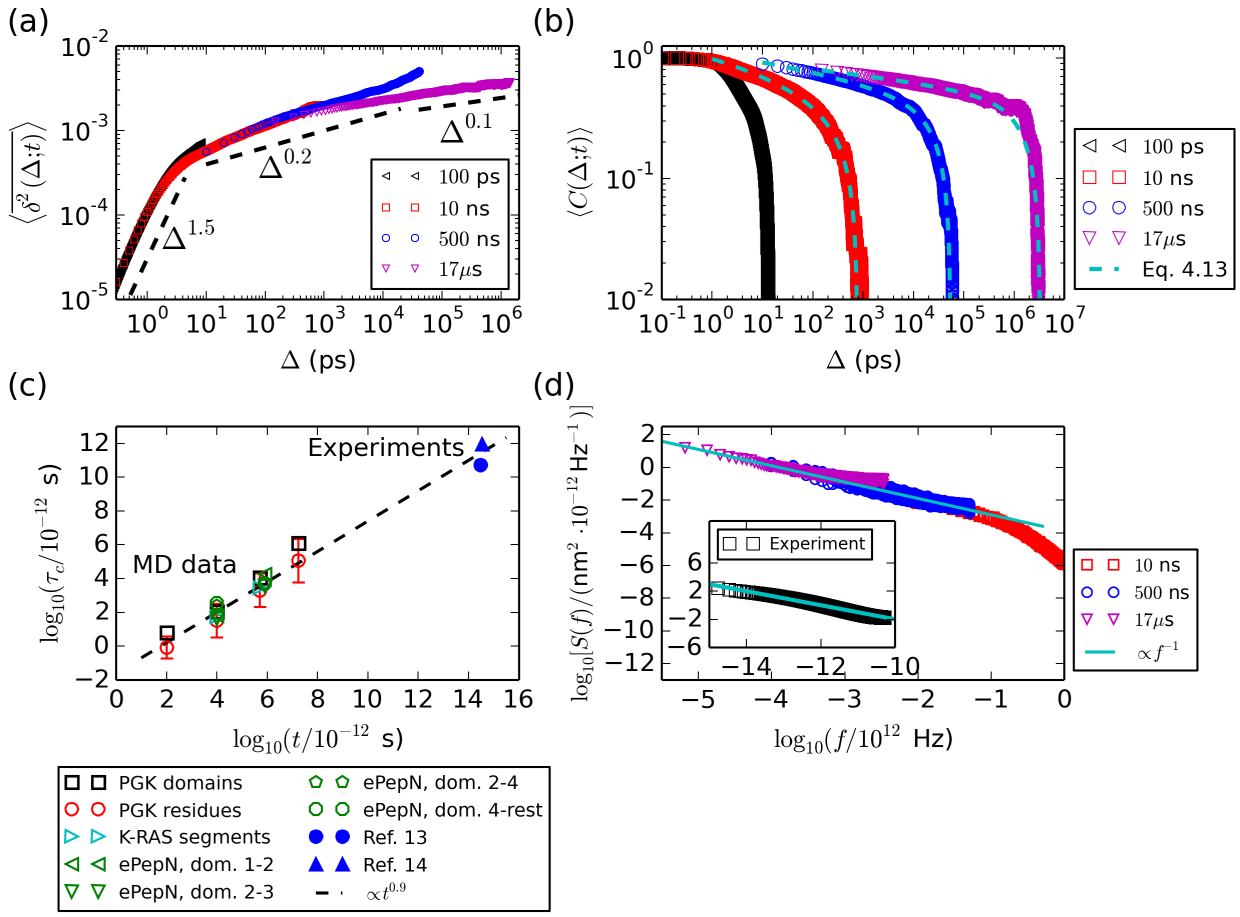

Figure 4.4: Nonequilibrium inter-domain dynamics of PGK. **(a)** TA-MSD averaged over five independent trajectories for $t = 100$ ps, 10 ns, 500 ns, together with the TA-MSD for $t = 17\mu$s. Dotted reference lines indicating power laws with different exponents are plotted as a visual guide. **(b)** ACFs of the inter-domain distance trajectories, calculated from different independent MD trajectories with the same legend as sub-figure (a). **(c)** Scaling behavior between the observed characteristic time τ_c and the observation time t. The logarithm (to base 10) of characteristic relaxation time τ_c of the inter-domain distance fluctuation of PGK, ePepN, of intra-domain structural fluctuation within the single domain protein K-Ras (see SI), and of average for the distance fluctuations between residue side-chain pairs in PGK, are plotted against the logarithm (to base 10) of the observation time, t. τ_c obtained from MD simulations is defined as the time at which the normalized autocorrelation function decays to $1/e$. A reference line for the power-law relationship $\tau_c(t) \propto t^{0.9}$ is plotted as a visual guide. The error bars shown with the red circles represent the standard deviation of $\log_{10}(\tau_c)$ associated with individual residue pairs. **(d)** Power spectral density, $S(f)$ of the inter-domain distance fluctuation of PGK *versus* frequency, f ($[f] = 10^{-12}$ Hz), calculated using the Welch algorithm [140]. Different colored symbols indicate different observation times; black: $t = 100$ ps, red: $t = 10$ ns, blue: $t = 500$ ns and magenta: $t = 17$ μs. The inset shows the estimated PSD of protein structural fluctuation based on the experimental single molecule data published in Ref. [95], obtained by numerical Fourier transform of an analytical fit to the experimentally measured autocorrelation function.

We first looked at the TA-MSDs of the inter-domain distance time series obtained on different observation time scales. For timescales below pico-seconds, the TA-MSD scales as t^{α} with $\alpha \sim 1.5$, indicating sub-ballistic dynamics. For timescales beyond 1 ps, the slope of the MSD quickly transitions from sub-ballistic to subdiffusive within $\sim$ 10 ps, where α continuously decreases to $\sim$ 0.1 at μs timescale (see Fig. 4.4(a)). Due to the continuously decreasing exponent α with increasing observation time, the MSD beyond a few ps cannot be described by a single power law. With the increasing length of the observation timescale, the increase of the MSD is extremely slow, consistent with the behavior of asymptotic approach towards the thermal plateau. Furthermore, TA-MSDs obtained at different observation times (length of the trajectory) are shifted relative to each other, with the slope becoming increasingly smaller with increasing trajectory length. This behavior has been observed in the numerical simulations of the continuous time random walk in confinement and is considered a manifestation of the aging effect [101].

Next, we calculated the normalized autocorrelation function $C(\Delta; t)$ of the respective times series of the inter-domain dynamics and the results are shown in Fig. 4.4(b). The most striking feature of this figure is that $C(\Delta; t)$ shifts towards longer lag-times with increasing t, *i.e.* the tail of the ACF becomes increasingly fatter with the longer observation time. If we introduce a phenomenological characteristic time τ_c for the observed dynamics of the inter-domain motion at a given t as the lag-time Δ at which $C(\Delta; t)$ decays to $1/e$, the shifting $C(\Delta; t)$ means that the *observed* characteristic time is rather an arbitrary quantity, since it varies with the observation time length itself. This behavior is again consistent with aging and also observed in the other two proteins, K-Ras and ePepN as well. The TA-MSDs and ACFs for the inter-segment and inter-domain distance fluctuations of K-Ras and ePepN are provided in Appendies B.2 and B.3, respectively.

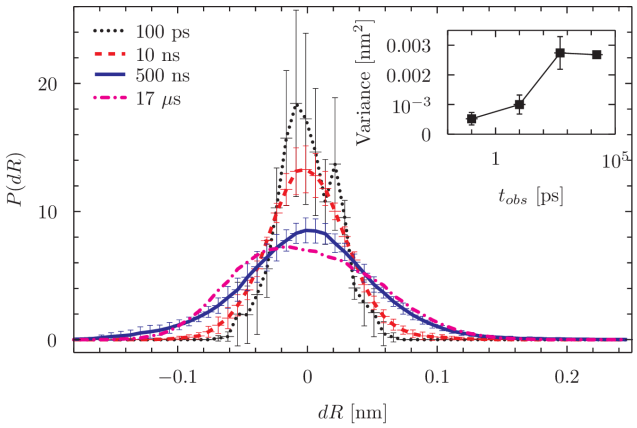

Figure 4.5: The inter-domain COM distance distributions centered at the average distance R_0 averaged over independent simulations at different observation time scales. The variance of the inter-domain distance distributions on different observation times are shown in the inset of figure

The observed aging is indeed intriguing, because the spatial magnitude of the motion appears

to have converged on the microsecond timescale. We have calculated the normalized inter-domain COM distance distributions from different trajectories obtained at different observation time lengths and the results are shown in Fig. 4.5. The data here indicates that the variance of the distance fluctuation no longer increases after hundreds of nanoseconds. This observation is fully consistent with the fact that protein structural fluctuation is a confined motion restricted by the folded structure. The variance of such motion can not increase indefinitely. However, spatial confinement does not automatically guarantee that the dynamics will reach stationarity and thermodynamical equilibrium as discussed in Sec. 2.6. The absence of the temporal convergence of the dynamics despite the finite spatial confinement is a strong circumstantial evidence for that the protein dynamics observed here are non-ergodic, non-equilibrium processes.

4.3 Distance fluctuations between residue pairs

Data presented so far concern distance fluctuation between two points representing the average positions of a substantially large group of atoms, such as a protein domain, thus indicative for global, collective dynamics. An important question is how general the non-equilibrium, aging dynamics observed for the inter-domain motion in PGK is also valid for local structural fluctuations.

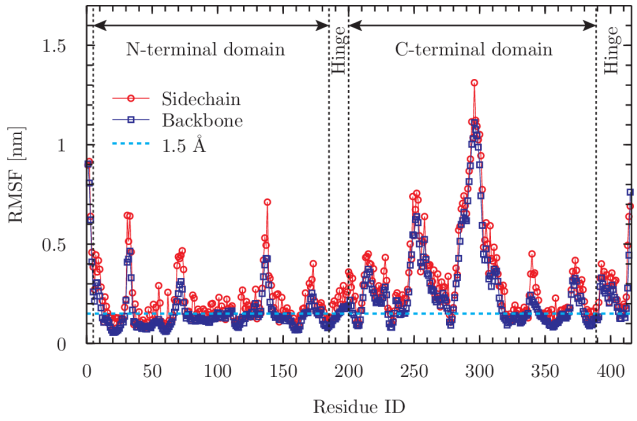

Figure 4.6: Backbone and sidechain root mean square fluctuations (RMSF) averaged over each residue. A cyan dashed line at RMSF = 1.5 Å is plotted to serve as a visual reference.

In the experimental single-molecule studies [95, 143], distance fluctuations between close-by residue sidechains are probed, which are typically $\sim 0.3 - 0.4$ nm apart from each other [95, 143]. In contrast to the inter-domain motions in PGK presented here which averages at a distance of ~ 3.8 nm. Clearly, distance fluctuations between covalently bonded atoms will be in equilibrium, as these are close to harmonic in the sense of the Born-Oppenheimer approximation. However, it is not clear *a priori* whether small groups of atoms, such as atoms constitute individual amino acid side chains coupled to the fluctuating protein matrix, may or may not exhibit similar behaviors

as those of the inter-domain motions. Therefore, we carried out more detailed investigations of inter-residue distance fluctuations by using 32 selected residue sidechains pairs from a variety of structural environments in PGK and performed the same analysis as it was done for the inter-domain motions.

Of the 32 residue pairs, eight were selected based on their time averaged residue side-chain root-mean-square fluctuations (RMSF) per residue calculated from the longest MD trajectory (Fig. 4.6). Four residues from each domain - those two with the highest and those two with the lowest and second lowest RMSF values, representing the distance fluctuations between the two most flexible and two most rigid residue side chains in each domain, yield four intra-domain pairs. We also calculated the distance fluctuations between the two most and two least flexible side-chains on the two different domains, yielding four more pairs of inter-domain residues. The remainder of the 24 residue pairs were selected based on their positions in the protein and the secondary structural motifs and their locations in the protein. Both the N- and C-terminal domains of PGK have a Rossmann-fold tertiary structural motif, containing six parallel β-strands forming the central core of each domain surrounded by four parallel α-helices connected by loops [137]. We selected nearby residues on the same β-strands in the core regions of the domains, on the surrounding $\alpha-$helices closer to the protein surface and on the loops that are directly exposed to the solvent, and calculated distance fluctuations between residues on both the same and different secondary structural motifs and both within each and between the domains. The selection yields in total 480 individual autocorrelation functions (due multiple independent trajectories for the same observation time t, with the exception of $t = 17\mu$s, where only a single trajectory was available.

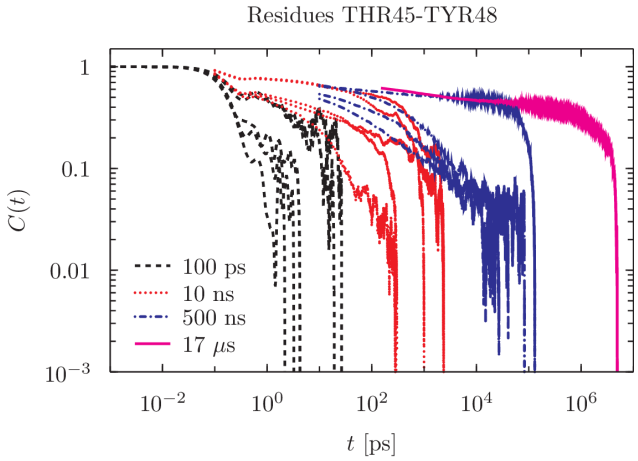

Figure 4.7: Normalized autocorrelation function of the distance fluctuation between the sidechains of the residue pair THR45-TYR48. Both residues are located on the same $\alpha-$helix. Despite the close proximity between the two residues (~ 0.7 nm apart from each other), the autocorrelation functions still exhibit highly non-equilibrium behavior.

The distance fluctuations between residue pairs analyzed exhibit heterogeneous dynamical

behavior and substantial variations were observed: for some residue pairs both the spatial and temporal aspects of the dynamics converge quickly, while others show no sign of convergence at all, *i.e.* the variance of the distance fluctuation and the characteristic time of the motion both increase continuously with observation time up to at least the $\sim 10\mu$s timescale. However, the average behavior of the inter-residue distance fluctuations over all pairs analyzed shows the same quantitative behavior as is seen for the inter-domain COM motion, *i.e.*, the corresponding characteristic time continuously increases with the observation time, following practically the power-law as for the inter-domain dynamic (see the open red circles in Fig. 4.4(c)). Thus, the non-equilibrium dynamical behavior appears to hold for both for global (*e.g.*, inter-domain) protein motion and a substantial fraction of local, inter-residue motions, and in some cases even for distance fluctuations between adjacent residues on the same α-helix, an example of which is shown in Fig. 4.7.

4.4 Power spectral density of the PGK inter-domain motion

Besides TA-MSD and ACF, we also investigated the frequency spectrum of the observed dynamics in PGK by computing the power spectrum of the PGK inter-domain distance fluctuation. The data in Fig. 4.4(b) indicate that the characteristic relaxation time of protein motions prolongs when the observation time is extended. To further analyze this finding, we computed the power spectra $S(f)$ of the PGK inter-domain COM distance trajectories $R(t)$, given by

$$S(f) = \left|\tilde{R}(f)\right|^2, \tag{4.5}$$

where f is the frequency and $\tilde{R}(f)$ is the Fourier transform of $R(t)$. The results are shown is shown in Fig. 4.4(d).

Interestingly, $S(f)$ obtained on different observation timescales can be concatenated onto a single spectral profile. For frequencies $f \geq 0.1$ THz, $S(f)$ scales approximately as f^{-1} over nearly five frequency decades is observed in the MD simulation data for PGK, indicating a $1/f$-noise. Moreover, we have estimated the power spectrum of inter-residue distance time series of the single molecule experimental data published in ref. [95] by performing the numerical Fourier transform of the analytical fit function of the for the experimentally observed autocorrelation function obtained on the timescale of 10^2 seconds. The estimated PSD from the experimental data follows roughly the same $1/f$ behavior (inset in Fig. 4.4(d)). Hence, the $1/f$ dependence of the power spectrum is likely to extend much more beyond micro-second timescale as observed in our MD simulations and may extend up to timescales of the single molecule experiments up to minutes. This "$1/f$-noise", often also referred to as "flicker-noise" or "pink noise"[1], indicates self-similarity of the corresponding

[1]The color "pink" originates from fact that the PSD of $1/f$ noise, *i.e.* $\propto f^{-1}$ lies between the red noise $\propto f^{-2}$, which is the PSD of Brownian motion, and the white noise, which has a constant PSD, *i.e.* $\propto f^0$.

dynamics on different timescales [139]. The lack of a characteristic frequency associated with a $1/f$ spectrum is consistent with the observation of the time-dependent relaxation time, *i.e.*, motions with ever lower frequencies are sampled as the observation length is increased.

In proteins, $1/f$-noise spectra have been reported in MD simulations for the gorge gating motion of acetylcholinesterase [123, 124] and in the fluctuations of the current conductance in ion-channel proteins in patch-clamp experiments [14, 125]. For these systems it was suggested that the $1/f$ spectrum is due to complex dynamics in the structural fluctuations related to the gated conductance behavior of the proteins [14, 125].

It has also been proposed [31] that $1/f$-noise behavior in protein conformational dynamics and reaction kinetics may be related to self-organized criticality (SOC) [8, 9], in which the self-similar, fractal dynamics indicated by the $1/f$ power spectra is the result of a self-organized critical state. This "critical state" in the context of SOC is not to be confused with the critical state in the theory of phase transitions, but rather represents an attractor of a non-linear dynamical system, towards which the system naturally evolves, and is insensitive to the adjustments of system parameters, such as temperature, pressure, etc. [8, 9]. In contrast, in the case of a phase transition, system parameters need to be carefully adjusted in order for the system to be able to reach the critical state. Further studies and data are required in order to determine as whether the protein might be in such a self-organized critical state, but it is nonetheless an interesting notion that worth the attentions of future studies.

4.5 Aging and observation time dependent dynamics as a general phenomenon in globular proteins

One most crucial observation made in the study presented in this dissertation is an intriguing commonality is found in the dynamics examined on different timescales: Fig. 4.4(c) shows that τ_c, the characteristic times of the inter-domain motions, large-scale intra-domain motions (*i.e.* relative motion between the segments in the single domain K-Ras) and as well as the inter-residue dynamics increase in a power-law fashion with the observation time, t, where $\tau_c(t) \propto t^{\theta}$, with $\theta \approx 0.9$, showing no sign of convergence. Remarkably, data from single-molecule experiments on the distance fluctuations between side-chain pairs [95, 143] fall close to the same power-law relationship (Fig. 4.4(c)). These data were obtained from two completely different globular proteins with an observation time up to ~ 300 seconds – more than 7 orders of magnitude longer than the longest MD simulations carried out in the present study. Together, Figs. 4.4 (a)-(c) reveal strong non-stationarity (aging) of the inter-domain dynamics and suggest a power-law dependence extending from 10^{-12} s to 10^{2} s. The data points on the power-law dependence between τ_c and t originated from independent

studies of five different globular proteins and observed on timescales spanning over roughly 13 decades. This commonality suggests that the dynamics of globular proteins in general may be intrinsically out of equilibrium and non-ergodic, which is in the contrary to the common belief: given sufficiently long observation time, protein dynamics maybe will reach its equilibrium and the time averaged properties will converged to the ensemble averaged counterparts.

This may be indeed the case, assuming one could establish ideal conditions, for example, an *in vitro* sample of purified protein in an optimal buffer, shielded from any chemical or radiation damages, such that the protein's physical and chemical integrities are preserved under physiological conditions over a long period of time, such as over weeks or months. Under such circumstances, the protein may be able to fully sample the accessible phase space points and its dynamics may be able to fully relax to the thermodynamical equilibrium and become fully ergodic, but such systems have no biological relevance. Proteins are synthesized by the cell for very specific purposes and all proteins *in vivo* have a finite life time, typically ranging between ~ 30 min to a day depending on the type of the protein and the organism to which the cell belongs to [10, 23, 131]. Especially the cellular concentrations of globular enzymes, such as kinases, are strictly regulated [10, 23]. These proteins are translated based on the demand posed by the substate concentration or other biological relevance, and, actively degraded by the cell once they are no longer needed. As results, cellular enzymes, such as yeast PGK, may have a typical lifespan as short as 45 minutes [10]. If the power-scaling of the characteristic time *vs.* the observation time is indeed valid for another one or two orders of magnitude, it would imply that the protein dynamics may never reach its equilibrium on the timescale of typical lifespan *in vivo*. Thus, proteins will have to carry out its function in a dynamical non-equilibrium state. Besides the possibility of transient non-ergodicity, the system may be intrinsically non-ergodic in the sense of weak ergodicity breaking (see Chap. 2, Sec. 2.2.2). In this case, the time required for the protein to fully sample the phase space is infinite, therefore the aging will persists indefinitely. A conclusion of non-ergodic and non-stationary protein dynamics would have profound impact on our understanding of protein functions and biological processes in cells in general. We shall discuss these in greater details in Chap. 5.

4.6 Autocorrelation functions of the inter-domain dynamics and evidence of broken ergodicity

So far our data have clearly demonstrated that the protein dynamics are subdiffusive and exhibit aging. However, it is not yet clear what physical model is behind the observed dynamics. One major challenge lies in the fact that the deviation from Brownian dynamics means the strong convergence towards Gaussian as the consequence of the central limit theorem is broken [92]. Once entered

the realm of subdiffusion, there are many different generalizations of the Einstein-Smoluchowski diffusion that can give the rise of similar subdiffusive behavior but with very different statistical properties and physical implications.

Pioneering single molecule experiments have provided direct insight into the intrinsic protein structural dynamics at room temperature in aqueous solution, indicating that, even on timescales as long as 10^2 s, single protein structural dynamics still remains highly subdiffusive and non-Markovian [95, 143]. However, the results from these experiments did not yield a clear answer as to whether single protein internal dynamics is ergodic, since relaxation functions derived from both ergodic (*e.g.*, the generalized Langevin equation) and non-ergodic models (*e.g.*, the fractional Fokker-Planck equation) seem to fit the experimental data equally well and both models have been applied to interpret the experimental data [95, 143].

To address the question as to whether global, function-related protein internal dynamics are indeed ergodic, we analyzed the dynamics of the distance fluctuations using the TA-MSD and ACF on different observation timescales (trajectory lengths), t. Data for PGK, K-Ras and ePepN are in details in Figs. B.1, B.3 and B.2 in the Appendices, respectively.

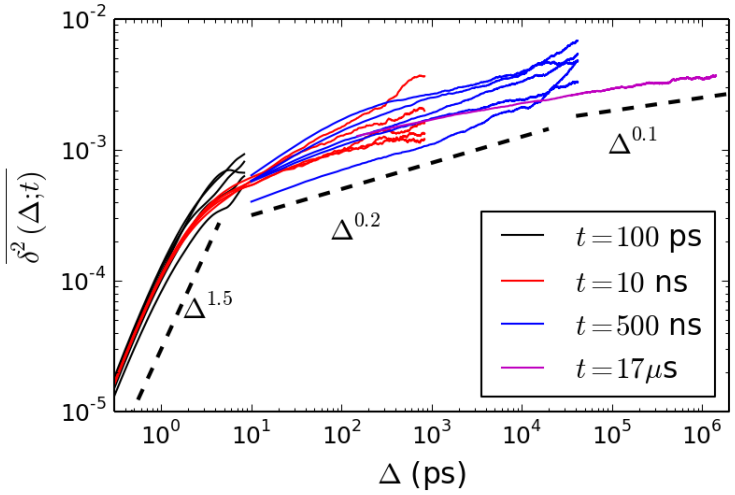

Figure 4.8: Individual TA-MSDs calculated from the inter-domain distance time series on different observation timescales t.

For the TA-MSDs calculated from individual MD trajectories, power-laws can be used to assess the dynamics over limited timescale ranges. As shown previously, a transition from the sub-ballistic into the subdiffusive dynamics occurs around the timescale of several picoseconds. For timescales larger than a few picoseconds, the TA-MSDs cannot be described by a single power-law but decrease its slope continuously, which is consistent with the asymptotic nearing of the thermal plateau. The ballistic region of the TA-MSD on sub-ps timescales stems from the fact that these timescales are too short for inter-atomic frictional forces in MD simulations to take significant effect, and therefore

the atomic motions on this timescale behave as friction-free. The transition into the overdamped limit occurs quickly for timescales larger than a few ps. Although the system is subjected to overall confinement, arising from the well-defined average protein structure, convergence of the TA-MSD to a time-independent value was not observed here, including in the longest simulation of 17 μs.

The lag-time dependence of the TAM-MSD does not yield much clue for question regarding to the ergodicity since TA-MSDs for subdiffusive dynamics from FLE or subdiffusive CTRW can yield similar behavior of slow approaching towards the plateau defined by the confining potential. Although if looking at the TA-MSDs calculated from individual trajectories, there are some significant spread among those obtained on the same observation timescale. A large spread between individual time averaged quantities is a characteristic behavior for weak ergodicity breaking. The distribution of the spread between individual TA-MSDs can be estimated via a dimensionless parameter ξ, defined as $\xi = \overline{\delta^2(\Delta;t)} / \left\langle \overline{\delta^2(\Delta;t)} \right\rangle$ where t is the length of the observation [21, 92]. In the case of ergodic process and for sufficiently large number of independent TA-MSDs, ξ will be distributed sharply around unity following a delta function $P(\xi) = \delta(\xi - 1)$ for $t \to \infty$ [21]. For certain slowly converging but still ergodic subdiffusive process such as fractional Brownian motion, $P(\xi)$ follows approximately a Gaussian distribution where the variance is determined by the intrinsic timescale of the process, the length of the observation t and the lag-time Δ, but independent of the Hurst exponent $H = \alpha/2$ [21]. For non-ergodic subdiffusive processes, $P(\xi)$ follows [21]

$$P(\xi;t) \to \frac{\Gamma^{1/\alpha}(1+\alpha)}{\alpha \xi^{1+1/\alpha}} L_\alpha \left(\frac{\Gamma^{1/\alpha}(1+\alpha)}{\xi^{1/\alpha}} \right) \quad \text{for} \quad t \to \infty \tag{4.6}$$

with $L_\alpha(x)$ as the one-sided Lévy stable distribution, where all moments diverge. Unfortunately, due to limitations in the computing capacity, we were unable to acquire sufficient number of trajectories to provide reliable statistics to accurately estimate $P(\xi)$.

Interestingly, the form of the ACFs obtained from different MD simulation trajectories yields an important clue for the dynamics. Given the Gaussian-like inter-domain distance distribution (see Fig. 4.5), we assume that the confining potential for the inter-domain motion is approximately harmonic, *i.e.* $V(x) = Kx^2/2$ with k is the spring constant. For ergodic subdiffusive processes within an harmonic potential, such as those governed by the FLE within a confining potential, the corresponding two time position-position autocorrelation function in the overdamped and stationary limit follows is given by [57, 67]

$$\langle x(t_1)x(t_2)\rangle_{\text{st}} = \frac{k_B T}{K} E_{2-\alpha}\left[-\frac{K}{\gamma \Gamma(\alpha-1)} |t_2 - t_1|^{2-\alpha} \right] \tag{4.7}$$

with $E_a(z) \equiv E_{a,1}(z)$ is the Mittag-Leffler function. As shown previously in Sec. 2.6.1, for large $\Delta = |t_2 - t_1|$, Eq. 4.7 scales as a power-law $\propto \Delta^{-(2-\alpha)}$. However, if looking at the ACF data (see Appendices B.1, B.2 and B.3) calculated from all protein dynamics time series, the tail of the ACFs

strongly deviate from a power-law, which would appear as a straight line in the log-log scaled plots. Therefore, the FLE cannot be used to interpret the ACFs. Combined with the clear evidence of aging, the data rather pointing towards a non-ergodic dynamics. A natural starting point here is the subdiffusive CTRW with a diverging waiting time distribution and a jumping distance distribution with finite variance.

From physical point of view, subdiffusive CTRW processes obeys the fractional Fokker-Planck equation [91, 93, 94] (FFPE) and describes a non-ergodic, non-stationary stochastic process out of the dynamical equilibrium. It is indeed counterintuitive, that even within an harmonic potential, the spatial confinement alone does not always guarantee stationarity and equilibrium such as in the case of subdiffusive CTRW [92]. In the absence of ergodicity and stationarity, the Wiener-Khinchin theorem that provides direct relationships between TA-MSD, ACF and power-spectrum of a stationary stochastic process is no longer valid.

For non-ergodic and non-stationary stochastic processes obeying FFPE, the corresponding position-position autocorrelation function has the form of [22]

$$C(t_2, t_1) \equiv \langle x(t_2)x(t_1)\rangle \sim \left(\langle x^2\rangle_{\text{th}} - \langle x\rangle^2_{\text{th}}\right) \frac{B(t_1/t_2, \alpha, 1-\alpha)}{\Gamma(\alpha)\Gamma(1-\alpha)} + \langle x\rangle^2_{\text{th}} \tag{4.8}$$

where

$$B(z, a, b) = \frac{1}{\Gamma(a)\Gamma(b)} \int_0^z y^{a-1}(1-y)^{b-1} dy \tag{4.9}$$

is the incomplete beta function, $\Gamma(x)$ is the gamma function and $1-\alpha < 1$ is the power-law exponent of the subdiffusive TA-MSD. By setting $\Delta = |t_2 - t_1|$ we can simplify Eq. 4.8 to

$$C(\Delta; t)_{\text{CTRW}} = C_1 B(\Delta/t, \alpha, 1-\alpha) + C_2 \tag{4.10}$$

where C_1, C_2 are constants.

Eq. 4.10 captures the overall shapes (especially for large Δ) of all ACFs observed in the MD data of all proteins simulated. However, there are still deviations in the short time ranges where Δ is small. Clearly, a simple, generic confined CTRW model cannot fully explain the protien conformational dynamics. If looking at the inter-domain distance fluctuation trajectories directly, such as in Fig. 4.3 or Fig. 4.12(a) (top panel), these time series do not exhibit the characteristic long waiting periods typical of subdiffusive CTRW such as displayed in Fig. 4.9(a). Also, the normalized distribution of the fluctuation about the mean of the distance $P(dR)$ for PGK , where $dR = R - \langle R\rangle$ (see Fig. 4.5), converges within statistical uncertainty on the μs timescale, and the Gaussian-like form of the distribution is consistent with those from ergodic subdiffusive modes, such as fractional Brownian motion (FBM) or processes governed by FLE [92]. Furthermore, we have conducted the p-variation test [85, 87] and the result of the raw protein dynamics trajectories. The results shown a relatively linear increase increase of the quadratic partial sum $V_n^{(2)}(t)$ with the increasing trajectory

run time. An example using the longest, 17 μs PGK trajectory, is shown in Figs. 4.12(a) and (b). In the following, we show that these observed discrepancies and the apparent ergodic-like behavior results can be explained by from the noise superposed on the CTRW, which itself is a non-ergodic, non-equilibrium process using the so-called noisy CTRW model [56].

4.7 The noisy CTRW picture of protein conformational dynamics

The subdiffusive CTRW is the results of the restriction and trapping in the temporal aspect of the diffusion. In picture of the conformational energy landscape picture [41, 44], the protein structural conformation can be considered as a random walk of a fictive particle over a rugged energy landscape, on which each point represents a certain protein conformation. Energy landscape for proteins can have a very complex topology and geometry on which deep wells can be found in which the fictive walker can be trapped over extended period of time. Such long trapping within deep, local wells and transitions from one to another can be captured by the CTRW, however, while dwelling inside of a trap, the system is assumed to be standing still as indicated by long horizontal flat sections in the trajectory (see Fig. 4.9). However, real proteins *always* carry out structural fluctuations and therefore introducing thermal noise to the overall relaxation signal. This is a feature of the protein dynamics that cannot be captured by th generic subdiffusive CTRW model. This thermal fluctuations while trapped in local wells will obviously contribute to the protein dynamics on shorter timescales.

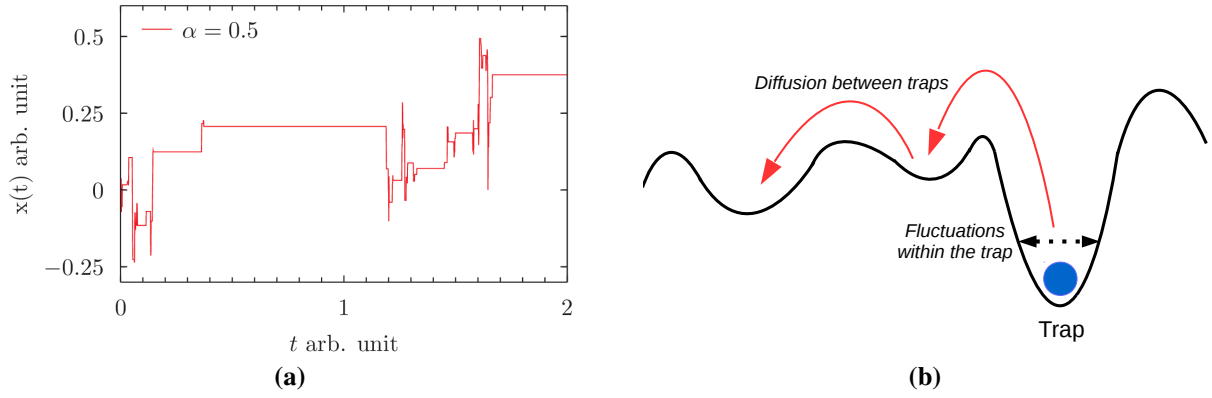

Figure 4.9: The limitation in the generic CRTW model. (a) An example of subdiffusive CTRW with a power-law waiting time distribution (Eq. 2.48) with the exponent $\alpha = 0.5$, which exhibits the characteristic long waiting time period. The trajectory is generated using the algorithm described in ref. [86]. (b) While the generic CTRW model can capture the jumps from one trap to another, the thermal fluctuation within the traps, which themselves can be complex non-Brownian motions, are totally neglected and replaced by stationary flat line, as shown in the example in (a).

In the noisy CTRW model introduced in ref. [55], motivated by that fact that real systems always

exhibits thermal fluctuations regardless of being trapped or not. Therefore, noise is superimposed on the generic subdiffusive CTRW motion. Uhlenbeck-Ornstein process of fixed amplitude was proposed as a potential model for stationary thermal noise [55]. This concept is particularly useful to describe protein structural motion together with the idea of energy landscape with many deep wells [44]. In the combined model, the subdiffusive CTRW captured the dynamics from trap to trap and the thermal fluctuations within the traps can be seen as independent processes, therefore, the total TA-MSD and ACF are additive. Based on this idea, an ACF of the form

$$C(\Delta;t) = C(\Delta;t)_{\text{CTRW}} + C(\Delta)_{\text{noise}}, \tag{4.11}$$

can be constructed, where $C(\Delta;t)_{\text{CTRW}}$ is an aging relaxation function with explicit observation time dependence resulting from the subdiffusive CTRW, and $C(\Delta)_{\text{noise}}$ represents the average relaxation behavior within the traps. A reasonable choice for $C(\Delta;t)_{\text{CTRW}}$ is the Eq. 4.10 as discussed in Sec. 4.6. As shown in ref. [55], the added noise may be fully stationary and ergodic, however, the total resulting dynamics is still non-ergodic due to the long-term contribution of the subdiffusive CTRW.

Protein energy landscape is a complex environment the fictitious walker can be trapped locally in deep wells, however, the surface of the landscape within these wells is not smooth but may contain many smaller wells of various depths. Therefore, a simple Uhlenbeck-Ornstein process (Brownian motion confined in a harmonic potential) with a single exponential relaxation function can not be assumed to be sufficient to address the complexity. Instead, we use a more generalized and empirical Kohlraush-Williams-Watts (KWW) relaxation function, also often referred to as the stretched exponential, to describe $C(\Delta)_{\text{noise}}$, which can be considered as the average over an ensemble of single exponentials. In this case, the ACF for the full protein domain motion has the form

$$C(\Delta;t) = c_1 \exp\left[-(\Delta/\tau)^{\beta}\right] + c_2 B(\Delta/t, \alpha, 1-\alpha) + c_3 \tag{4.12}$$

where the exponent α can be determined from the fit of the subdiffusive section of the TA-MSD at larger lag-times where $\overline{\delta^2(\Delta;t)} \propto \Delta^{1-\alpha}$, and t is the length of the simulation trajectories, leaving c_1, c_2, c_3, τ and β as fit parameters. The stretched exponential parameters β and τ can be interpreted as descriptors of the average interior, local properties of the traps. Eq. 4.12 also captures one of the most crucial features in the observed dynamics, namely the aging effect is explicitly incorporated through observation time dependency in the form of Δ/t, as expected for the underlying fractional Fokker-Planck equation [22]. For the increasing observation time t, the ACF will shift towards longer timescale, exactly as observed in the MD simulations.

ACFs of the inter-domain/segment time series calculated from all simulated proteins were fitted using Eq. 4.12. The results are shown together with the corresponding TA-MSDs in Fig. 4.4(b) in Sec. 4.2, and Figs. B.1, B.3 and B.2 in the Appendices. We found that all ACFs obtained from observation timescales beyond picoseconds for all proteins considered can be well fitted by Eq.

4.12. The KWW parameters β and τ vary between different types of distance fluctuations (*e.g.*, different pairs of domains in ePepN and the inter-segment motion in K-Ras) in different proteins. The resulting fit parameters are shown in Tables B.1, B.3-B.5 and B.2 in the Appendices.

4.8 Fractal organization of conformational substates on the free energy landscape

The protein conformational dynamics discussed so far concern the structural fluctuation of folded proteins situated at the bottom of the folding funnel representing a general global free energy minimum. As discussed in Chap. 2, the bottom of the funnel is not a single point but its a relatively broad area with many local minima representing the conformation substates associated with the native state and separated by free energy barriers of different heights. [44, 135]. The dynamics of folded proteins can thus be considered as a fictitious particle diffusing on a rugged landscape possessing many wells of various depths, whose detailed features ultimately determine the resulting protein dynamics [42]. Therefore, it is important to characterize the topology and geometry of the energy landscape. However, due to the high dimensionality of the free energy landscape, it is very difficult to picture detailed geometrical and topological features in a intuitive manner.

One usual approach is reduce the high number of dimensions via projection of the full landscape onto a few or even a single reaction coordinate. However, such a radical reduction of the dimensionality can significantly conceal the topological characteristics of the true landscape. Even a simple example from the geometry of a 3-dimensional saddle surface function $f(x) = x^2 - y^2$ demonstrates, that depending which coordinate (x or y) one picks to represent the whole function, the reduced representation with either x or y can be very different and not representative for the actual full function. In the case of protein domain dynamics, given a reduced reaction coordinate, such as the COM distance between the domains R, there are many different possible pathways that can be taken by the protein to achieve the same changes in R. Different pathways can lead to very different dynamics depending on the length of the pathway over the landscape and free energies of the barriers between the individual conformational substates encountered along the pathway. On the other hand, some reductions have to be made in order to make the problem to be tractable at all.

To understand the features of the free energy landscape that give rise to the observed self-similar and non-equilibrium dynamics, we used a different approach; instead of predefining reduced reaction coordinates, we coarse grained the free energy landscape instead by considering only a set of distinct conformational substates and then projecting the coarse grained version of the landscape on to a network (or a graph) based on the transitions between different conformational substates. The protein configurations sampled during a simulation trajectory are sorted into discrete clusters based on

structural similarity. Similar configurations are located close to each other on the energy landscape and therefore a structural cluster represents a metastable local well on the landscape which form the vertices (nodes) of the network. Whenever the protein transits between two clusters during the simulation, an edge (bond) is added between the two vertices representing the clusters, thus forming a network we referred to as the *conformational cluster transition cetwork* (CCTN). Similar networks have been used previously to describe complex dynamics, *e.g.*, refs. [100, 104].

This network-based approach offers many advantages comparing to the generic reaction coordinate with highly reduced dimensions. First of all, each node of the network represented by the cluster average conformation remains a $3N$-dimensional object. In comparison, if the system is projected onto a simple reaction coordinate, such as the inter-domain distance R, two nodes representing two different local minima with same or similar R cannot be distinguished. Furthermore, the network itself is a high-dimensional construct, whose dimension is defined by the highest number of edges a single node has, but still can be illustrated and visualized on a 2-dimensional plane. The network can be seen as the analogous to the road map of a real landscape where the nodes represent cities and edges the highways connecting them. One may not know what exact landscapes are between the cities, *e.g.* flat plateau, rugged mountains or even a uncrossable river separating the land, however, based on the information provided by the map, one can tell that as long as there is a connecting highway between two cities, there are no uncrossable obstacles such as a river without bridges between these cities, and based the "traffic flow" between the cities, *i.e.* the frequency of the transitions associated with an edge, one can take a good guess, whether there is flat plateau that offers an easy travel between the cities or a rugged mountain where the traffic is slowed down.

To construct the CCTN from the simulations, we started by grouping the protein structural snapshots sampled during a simulation trajectory into discrete clusters based on structural similarity using a root mean square deviation (RMSD) between the protein heavy atom coordinates as a cut-off, typically between 1.5-2.0 Å. The cluster analysis was carried out by sampling a sufficiently large set of protein snapshots from the trajectory, equidistant in time. This analysis was performed on the 500 ns and 17 μs MD trajectories. 25000 snapshots, equivalent to a sampling rate of ~ 0.05 ps^{-1}, were used for the 500 ns trajectories, and for the 17 μs trajectory, 37995 snapshots were used for the analysis, equivalent to a sampling rate of ~ 0.002 ps^{-1}. These snapshots were sorted into clusters using an algorithm introduced in ref. [30]. Networks similar to CCTN have been used to describe the structural dynamics of complex systems, *e.g.*, refs. [100, 104]. The graphical illustrations of two representative examples of the networks are shown in Figs. 4.13 and 4.14 at the end of this chapter.

In Figs. 4.13 and 4.14, the obtained CCTNs are visualized as the following: Each node, or conformational cluster, is depicted by a circle labeled with an integer indicating the rank of the cluster in terms cluster size. The cluster size is define by the number conformations sampled during the simulation that belong to the cluster. In this sense, the size of cluster is inverse proportional to

the free energy of the conformational substate represented by the cluster, *i.e.* the larger the cluster size, the lower its free energy. For the purpose of better visualization, the cluster size is further visually enhanced by the diameter of the circle and its color tune in grey scale, *i.e.* large diameter and dark color tune indicate large cluster size and *vice versa*. The edges in the network are depicted by the arrows representing the transitions between the clusters. Here, the thickness of the arrow is proportional to the frequency of the particular transition it represents, *i.e.*, the more frequent a transition, the thicker the arrow. Furthermore, the transition frequency is visually enhanced via the color tune associated with the arrows which a diverging color scale from colder cyan to warmer magenta. Here, a warmer color tune is associated with a high transition frequency and *vice versa*. One most striking feature of these networks is the inhomogeneous connectivity between nodes. Nodes tend to form densely connected hubs around a few nodes with the highest ranks, while outside the hub, the connectivity is rather sparse and there is a relative large distance between the hubs. If look at the network obtained from the longest simulation (Fig. 4.13), within the major hub formed around the largest cluster (rank 1), certain hierarchical organization can be observed, *i.e.* the node with the lowest free energy (inverse proportional to its rank) can be see to be heavily connected to a few nodes with higher rank, *e.g.* nodes 3, 4 and 6, which again are densely connected to a big number of nodes with lower ranks. We will see later in this section that this inhomogeneity can be characterized by two distinct geometrical fractal scaling regions over different distances.

Since the energy landscape is a function of protein atomic coordinates [42], we assume that similar protein structures are located close to each other on the energy landscape thus a structural cluster as defined above represents a local minimum. The network formed by the transitions between these minima, *i.e.* the CCTN, can be used to characterize the geometrical and topological organization the energy landscape which corresponds to the equivalent properties of the graph. First, we examined the topological properties of the CCTN by calculating the degree distributions of the networks, defined as the probability, $P(d)$ of finding a vertex connected to d direct neighbors. In Fig. 4.11(a) we show the degree distribution $P(d)$ calculated from the four independent 500 ns and one 17 μs long MD simulation trajectories. The data can be well fitted by a log-normal distribution

$$P(d) = \frac{1}{\sqrt{2\pi}\sigma d} \exp\left\{-\frac{[\ln(d) - \mu]^2}{2\sigma^2}\right\}, \tag{4.13}$$

where the fit parameters μ and σ are the mean and standard deviation of the distribution, respectively. The numerical values of μ and σ from different data sets, obtained by fitting the data points with Eq. 4.13, are listed in the figure legend of Fig. 4.11(a). Interestingly, all $P(d)$s fully overlap on different timescales within the statistical errors, indicating topologically self-similarity, despite the huge sampling time difference between the 0.5 and 17 μs trajectories. A log-normal degree distribution is characteristic of random multiplicative processes [97], indicating that the probability $P(n)$ of finding a vertex with n neighbors can be written as the product $P(n) = \prod_{i=1}^{n} p_i$, where p_i

is the probability of vertex i being a direct neighbor of the vertex under consideration and the p_i are independent of each other [97].

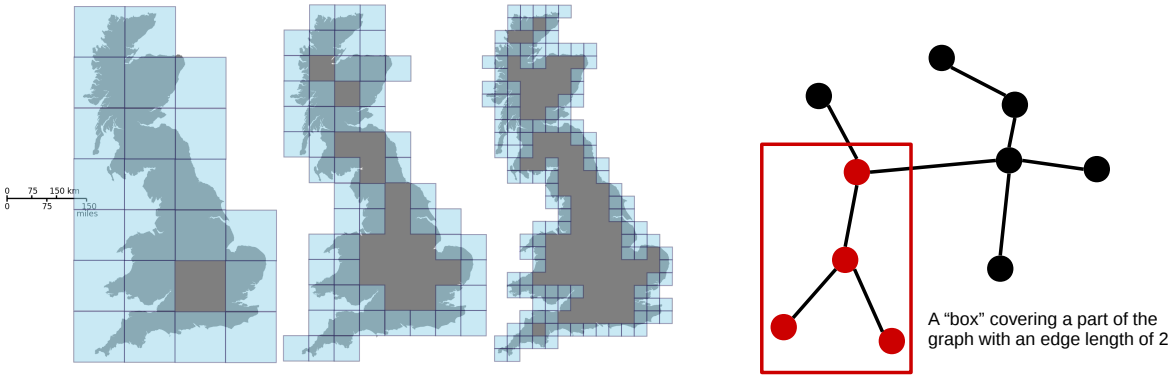

Figure 4.10: (a) An schematic illustration of the idea behind the box covering method. To determine the fractal dimension of the coast line of Britain, one could use a set of quare-shaped boxes with identical edge length to cover the coast line. Evidently, one will need an increasing number of boxes to fully cover all portions of the coast line with decreasing box size. Assuming the "mass" (portion of the coast line) contained within each box is roughly the same, for a fractal object, such as the coast line, the scaling behavior between the number of the boxes N and the edge length of the box l will follow a power law, *i.e.* $N \propto l^{-d_f)}$, where d_f is fractal dimension. Figure adopted from Wikipedia. (b) An schematic illustration of a "box" in the context of graph with the "edge length" 2 covering a subset of nodes (colored in red) in the graph.

Next, we investigated the geometrical properties of the graph. Especially, we are interested to see whether the network will also exhibit a fractal, self-similar geometry. In the context of a graph, it is important to distinguish between the self-similarity in the *topology* and the *geometry* of a network, which were often confused with each other and used interchangeably in the literature [129]. As pointed out in ref. [129], a so-called scale-free network, indicated by a power-law degree distribution [3] may not necessarily have a fractal geometry and, *vice versa*, a network with a fractal geometry may not have a degree distribution following a power-law.

The geometry of a graph is rather an arbitrary quantity because it requires the definition of a metric, such as the distance between two points in the Euclidean space is the length of the shortest path that directly connecting between them, *i.e.* a straight line. One common way to define the "distance", between two nodes in a graph is the shortest number of edges connecting them [129]. Once the metric is defined, one can determine the fractal geometrical scaling behavior of the graph, which implies a power-law relationship between the average “mass” $\langle M_f \rangle$ of a network its geometrical dimension r_f, *i.e.* $\langle M_f \rangle \sim r_f^{d_f}$, where the exponent d_f is the fractal dimension. In the context of a graph, the mass is constituted by the nodes of the network, *i.e.* the individual vertices are considered as mass points.

We applied a box covering algorithm [128, 129] (also see Sec. 3.5 for details) to the networks obtained from the MD simulations to determine the fractal dimension of the graph. In this approach

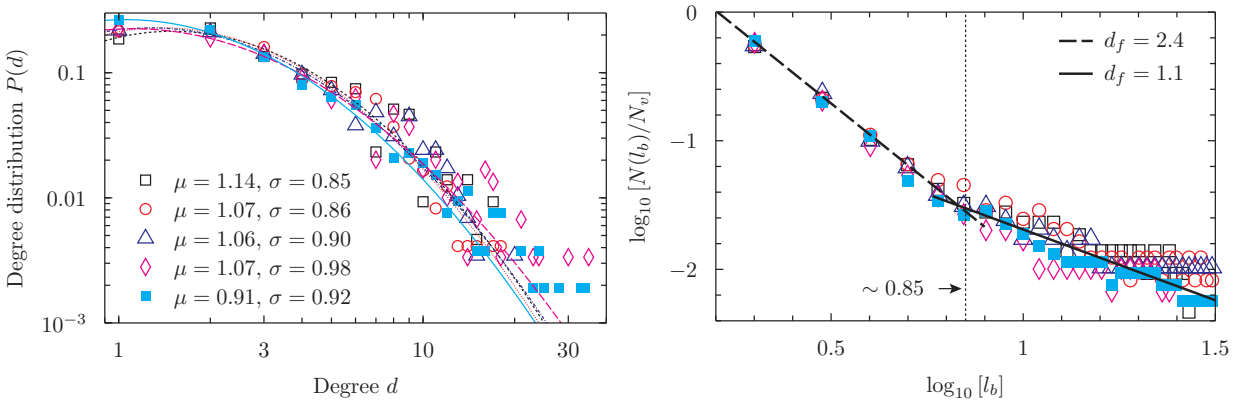

Figure 4.11: Properties of PGK transition networks. **(a)** Degree distributions $P(d)$ of the PGK transition networks (Figs 4.13 and 4.14) obtained from four independent 500 ns MD simulations (open symbols) and one 17 μs MD simulation (solid symbol). Different lines represent fits using log-normal distribution (Eq. 4.13). The μ and σ values for different data sets are determined from the fit of Eq. 4.13 and displayed in the figure legend. **(b)** Fractal scaling of different transition networks obtained using compact box covering algorithm [128]. The number of the boxes required to cover the network, N_b, normalized by the number of the vertices in the network, N_v, is plotted against the edge length of the box, l_b. Four different open symbols represent data obtained from four different transition networks generated from independent 500 ns MD simulations. The solid squares represent the data from the 17 μs MD simulation. The dashed line represents the average linear fit over all data sets.

is equivalent to the box covering method for determine the fractal dimension of geometrical fractals, such as the example of the coast line of Great Britain, as shown in Fig. 4.10(a). For the analysis of the CCTN obtained from the MD simulations, the network is covered using "boxes" of the dimension, or box edge length l_b, such that the distances, l, between all vertices within a "box" are smaller than the "box edge" i.e. $l < l_b$ [129]. An schematic example of such "box covering" is shown in Fig. 4.10(b). The same concept can be applied to a network; if the graph has a fractal geometry, the average "mass", $\langle M_b \rangle$, covered by a box with a geometrical dimension l_b, should follow the power-law scaling as

$$\frac{N_v}{N_b} \sim l_b^{d_f}, \tag{4.14}$$

where N_v is the number of the vertices in the graph, N_b are the total number of boxes that are required to cover the graph, thus the quotient $\frac{N_v}{N_b}$ represents the average mass per box $\langle M_b \rangle$ which scales as a power-law with the dimension of the box l_b with the power-law exponent d_f as the fractal dimension.

All networks obtained from observation timescales (500 ns and 17 μs) show the same fractal scaling behavior; up to distances of ~ 7 edges (≈ 0.85 on the log-scale), the networks exhibit the same fractal scaling with a fractal dimension of roughly 2.4. Similar values have been observed in other biological networks; for example, the protein interaction networks for *E. coli* and humans have

both been found to have a fractal dimension of about 2.3 [129]. However, for longer distances, > 7 edges, a sudden, discontinuous change occurs and the fractal scaling abruptly decreases to about 1.1. This sudden change arises because at short distances vertices tend to cluster into hubs around highly populated nodes, forming highly inter-connected sub-graphs, whereas the hubs themselves, separated by larger distances, are less densely connected. Examples for such hubs are the dense local networks formed around clusters 1 and 2 in Fig. 4.11(a). Consequently, the transition point seen in the fractal scaling behavior in Fig. 4.11(d) corresponds to the typical diameter of a hub. Such a hub can be seen as a larger well on the energy landscape containing many smaller wells, i.e. the cluster within the hubs, reflecting a hierarchical structural order of the energy landscape [44].

The network as employed here is not ensemble averaged, and nor is it for a single protein averaged over infinite time, but rather is a representation of the dynamics of a single protein molecule over a finite time period. In this case, only some of the possible transitions that take place are sampled stochastic manner. As expected for systems exhibit weak ergodicity breaking, the network thus obtained from any two identical molecules over the same time period will never be identical. However, the specific transitions are sampled during a trajectory is not of interest in the present context; rather we are investigating the overall characteristics of the connectivity of the time-dependent CCTN obtained from the single molecule MD trajectories, which has been revealed as self-similar and fractal at sufficiently long timescales. This self-similarity is the main result from the network analysis, that revealed a fundamental physical property of the energy landscapes of single protein molecules which unifies the way we can conceptualize the functional internal dynamics on vastly different timescales.

4.9 Coarse-graining (CG) of the protein dynamics using conformational cluster transition network

One other useful feature of CCTN is that one can construct the coarse-grained trajectories of the dynamics of interest. In the CCTN, at any given time during the simulation the protein is in one of the "states" represented by the structural clusters. Thus, one can coarse-grain the original MD inter-domain COM distance trajectory down to a set of discrete inter-domain distance values represented by those of the individual cluster mean structures. The degree of coarse-graining can be controlled by varying the RMSD cut-off value. For example, if the cutoff is zero, the number of clusters will be the same as the total number of frames in the original trajectory, and one would then recover the original MD domain distance time series. *Vice versa*, if the cut-off has an extremely large value, all snapshots in the MD trajectory will be grouped into a single cluster and thus the CG trajectory would yield a flat line at the average domain distance.

This coarse-grained trajectory can be used to test the validity of the noisy CTRW model. As demonstrated in Ref. [56], noise of sufficiently high amplitude superimposed on a generic subdiffusive CTRW can have significant impact on the overall dynamics of the trajectory and mask signature properties of the underlying subdiffusive CTRW. This can, for example, render the p-variation test inconclusive. Therefore, it would be beneficial if the noise could be separated or filtered from the CTRW. Here, we use the CCTN to filter out the noise from the trajectory through coarse-graining.

In our assumptions the noisy CTRW model applied to protein dynamics, the source of the noise is the thermal fluctuation of the protein while being confined in local minima on the energy landscape acting as traps, while the subdiffusive CTRW describes the transitions from one trap to another. In the coarse-graining of the trajectory using the CCTN model, the thermal fluctuations are removed implicitly, since all structural fluctuations within a conformational cluster represented by a node in the network, in other words, inside a sufficiently deep minimum on the energy landscape, are coarse-grained out and only transitions between minima remain. In order to study the effect of the noise as defined above on the overall dynamics of the inter-domain motion, we performed the CG procedure on the longest trajectory, the 17 μs PGK simulation, with 3 different cut-offs. The results are shown in Fig. 4.12.

In Fig. 4.12(a), the original inter-domain distance time series from the MD is compared to the CG trajectories. The CG trajectories show the same overall trend but become more CTRW-like with the increasing degree of CG. The results of the p-variation tests of the MD and CG domain distance trajectories are shown in Fig. 4.12(b). In the original MD trajectory, quadratic partial sum $V_n^{(2)(t)}$ (with $n = 10^{12}$) as a function of running time displays a nearly linear increase, which would be consistent with an an ergodic subdiffusive process [85]. However, as the degree of coarse-graining increases, $V_n^{(2)(t)}$ deviates increasingly from a straight line, becoming a more discontinuous, step-wise monotonically increasing function, which is characteristic of a nonergodic subdiffusive CTRW [85]. The TA-MSDs and the ACFs of the original and CG trajectories are shown in Fig. 4.12(c) and (d), respectively. The slopes of the TA-MSD of the original MD data and various CG times series are similar to each other for Δ up to 100 ns, increasing as a power-law with $\propto \Delta^{1-\alpha}$, with $\alpha \sim 0.9$. The ACFs exhibit the same behavior for large Δ but differ clearly in their decay behavior for $\Delta < 1\mu$s. With increasing coarse-graining, equivalent to removal of the noise from the overall signal, the ACF exhibits slower decay on sub-μs timescales. We fitted the ACFs of the original MD and the CG domain distance trajectory (see Fig. 4.12(d)) with a simple incomplete beta-function of the form

$$f(\Delta) = c_1 B(\Delta/t, \alpha, 1-\alpha) + c_2 \tag{4.15}$$

without the noise term in Eq. 4.12, where α is the subdiffusive exponent estimated from the TA-MSD, $t = 17\mu$s is the length of the trajectory and c_1, c_2 are the fit parameters. As can be seen in

Fig. 4.12(d), with increasing degree of CG the ACF becomes more similar to the plain incomplete beta-function (solid black line) than the ACF calculated from the original MD trajectory. Eq. 4.15 has been shown to be the ACF of a non-ergodic subdiffusive CTRW process derived from the fractional Fokker-Planck equation [22]. Thus, with the increasing degree of CG, equivalent to stronger noise filtering, the trajectories exhibit more non-ergodic, subdiffusive CTRW behavior, confirming the noisy CTRW interpretation as the mechanism of diffusion over the energy landscape.

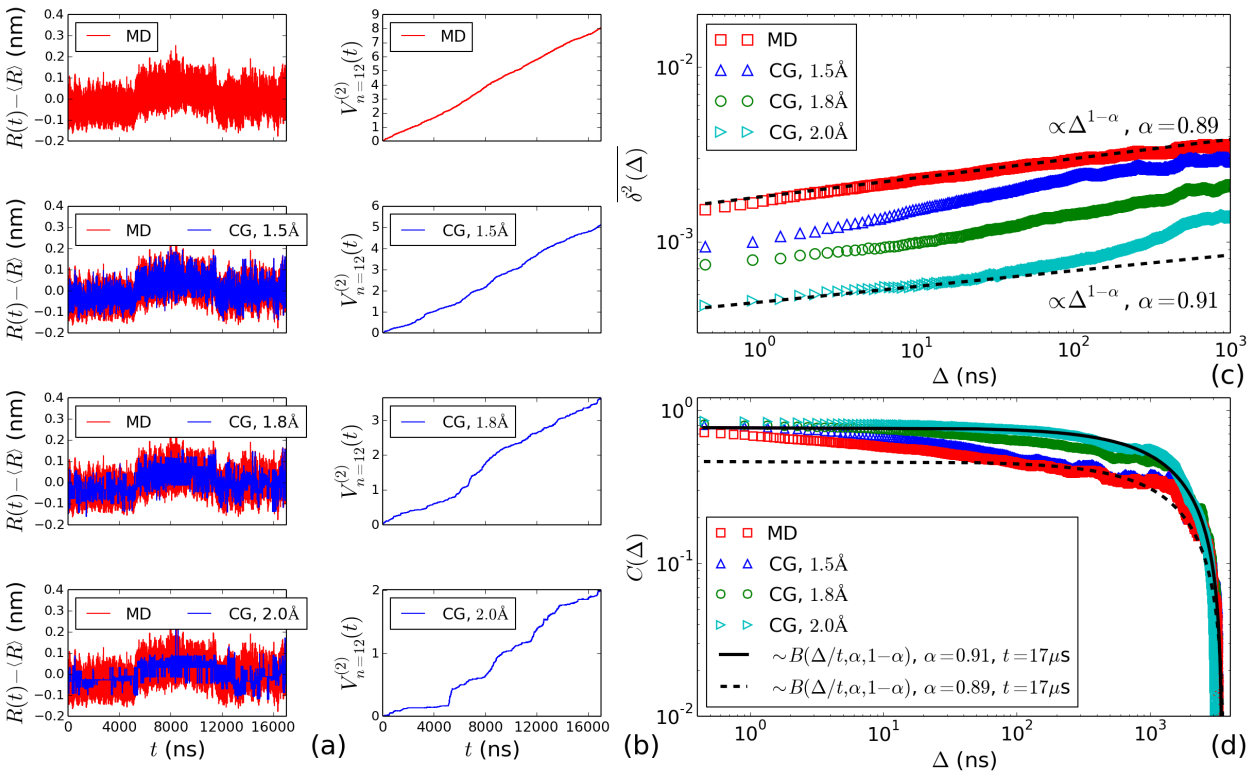

Figure 4.12: Comparison of Coarse-Grained and Atomistic models. Comparison between the MD (red) and CG trajectories (blue) generated with the CCTN model using the 17 μs PGK trajectory as an example. **(a)** Comparison between the original domain distance time series from MD simulation and CG time series with different clustering RMSD cut-offs. **(b)** Quadratic partial sum $V_n^{(2)}(t)$, with $n = 12$, as a function of the simulation run time. **(c)** TA-MSDs calculated from the MD and CG trajectories with different clustering cut-offs. **(d)** The ACFs of the MD and CG trajectories with different clustering cut-offs. The dashed black line is the fit to the ACF calculated from the MD trajectory (red open squares) using Eq. (4.15), while the solid black line is the fit to the ACF calculated from the CG trajectory with 2.0 Å cut-off (cyan open right triangle) using the same fit function.

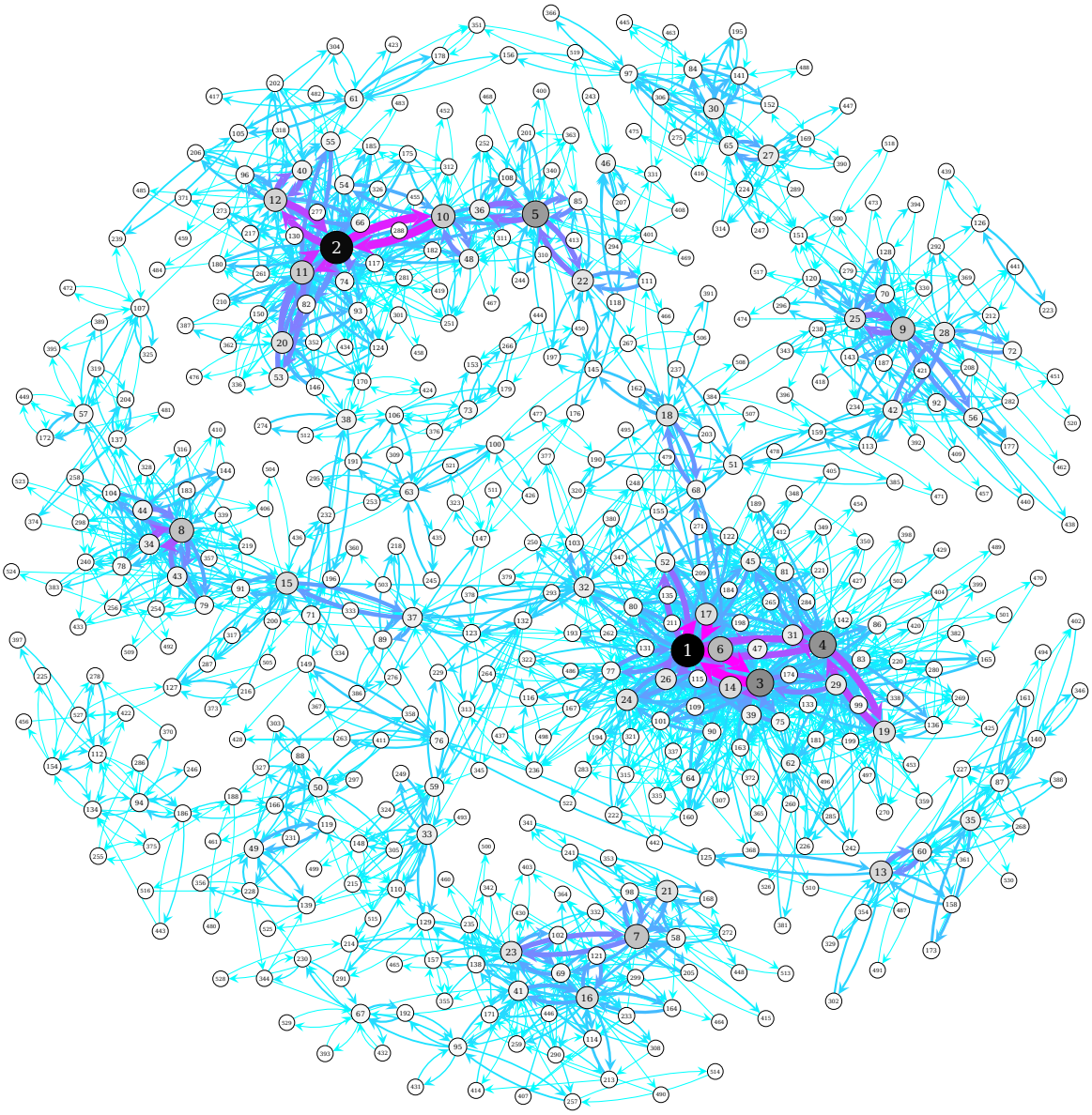

Figure 4.13: Network representation of conformational transitions in PGK. Conformational Cluster Transition Network obtained from a 17 μs simulation of PGK, containing 530 vertices and 2345 edges. The circles represent structural clusters, the diameter and the color scale of each circle indicate the cluster size, defined by the numbers of conformations belonging to the cluster. The integer label on each vertex indicates its index based on its rank in terms of the cluster size. The arrows represent the transitions between the clusters. The thickness of the arrow and warmness of color scale indicate transition frequency. The graphical representations of the networks were generated using the Python library graph-tool (http://graph-tool.skewed.de/).

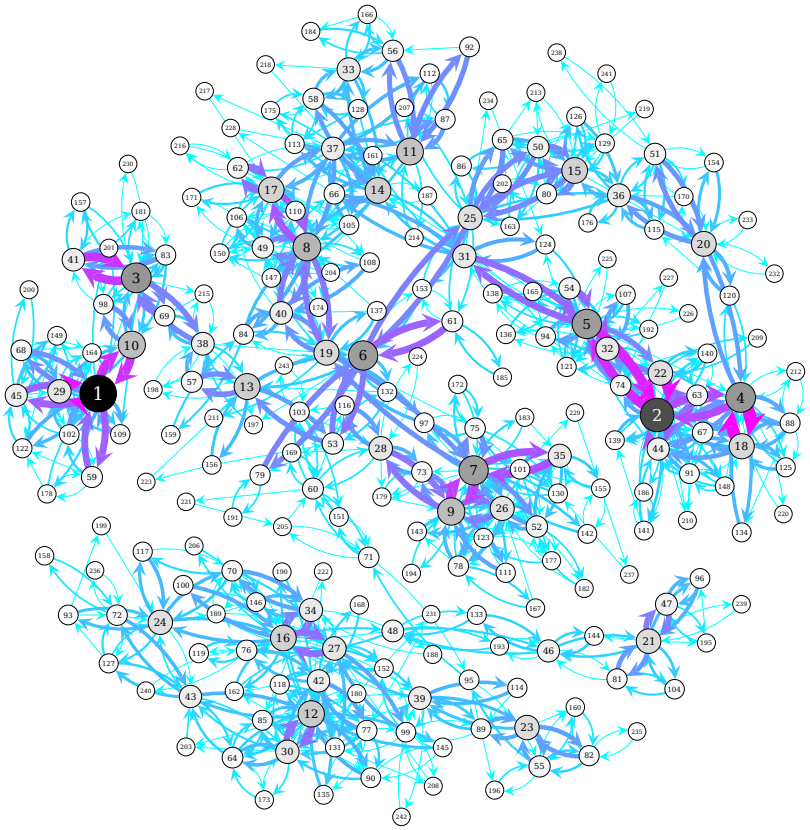

Figure 4.14: Network representation of conformational transitions in PGK. Conformational Cluster Transition Network from a 500 ns PGK simulation, contains 243 vertices and 951 edges. Colors and symbols indicate the same quantities as in Fig. 4.13. The graphical representations of the networks were generated using the Python library graph-tool (http://graph-tool.skewed.de/).

Chapter 5

Conclusions and Future Outlook

Proteins are essential building blocks of life that require proper structures and the appropriate modes of motion in order to perform their function. In this dissertation we presented the findings from our extensive molecular dynamics simulation study on the internal structural fluctuation of globular proteins. The essential findings of this dissertation can be summarized as followed: Data extracted from three globular proteins, which are vastly different in terms of their sizes and structural organizations, lead to a highly unintuitive and novel view to consider protein dynamics, namely, the dynamics of a single protein molecule behaves as a self-similar and non-ergodic stochastic processes over an enormous range of timescales. The combination of the existing single molecule experimental data [95, 143] with our MD results indicates that the observed non-equilibrium dynamics can span from picoseconds up to $\sim 10^2$ second, covering practically all timescales over which known protein functions occur. Aging effect, *i.e.* dynamics that depends on the length of the observation, appears to be omnipresent, and furthermore, this non-ergodic behavior in the sense of weak ergodicity breaking [92], appears to be a general phenomenon and not limited to specific proteins, which can eventually extend to timescales that are of typical protein lifespan *in vivo*.

We found that, empirically, a version of the noisy continuous time random walk (CTRW) model [55] can be applied to model the stochastic time series of collective protein structural fluctuation, such as inter-domain motions. In this model, aging is implicitly built in due to its origin in the fractional Fokker-Planck equation. Given the assumption that protein dynamics is determined by its underlying energy landscape [41, 42, 44], we applied a graph based model to access the characteristic of the energy landscape by mapping it onto a network based the clusters conformations sampled during the simulation and the transition between them. We found that the

resulting network is self-similar in terms of both topological organization and geometry. We showed that the observed non-ergodic protein dynamics can be understood as a subdiffusive continuous time random walk on the network representing the energy landscape due to extended trapping in the local minima on the landscape.

The findings presented here stand in the contrary to the common belief that the protein structural fluctuation is a fully ergodic process within the thermodynamical equilibrium, which serves as a basic assumption in many commonly used classical theories and models involving protein functions, such as transition state theory or Michaelis-Menten formalism [60], which are frequently used to study enzyme kinetics, which offers the convenience of a constant free energy barrier of the reaction. In the past two decades, a body of experimental studies have clearly demonstrated that the fluctuation of the catalytic rate displayed by a single enzyme molecule [96, 144] or an ensemble if enzymes [35] correlates directly with its structural dynamics. One interesting question here is whether these dynamics observed are indeed stationary, equilibrium processes? Although single-molecule experiments can offer crucial and unique insights on the dynamical behavior of individual protein molecules, they also have certain limitations, for example, processes happening on timescales faster than millisecond cannot be resolved due time resolution of current experimental techniques. However, for proteins, a millisecond is already a relatively long time window over which many dynamical processes can occur such as conformational changes, allosteric or enzymatic catalysis. Thus, dynamical quantities, such as autocorrelation function of protein structural dynamics, obtained from these single-molecule experiments yield rather the long-term limiting behavior of the real function. The rich dynamical information on intermediate timescales remains undiscovered in these experiments. Here, molecular dynamics (MD) simulation based techniques serve nicely as a complementary tool to examine the dynamics on these intermediate, sub-millisecond timescales, and with the ever improving computer hardwares and numerical algorithms, the MD simulations are expected to further push the existing boundaries and reach up to the experimental length and timescales in the near future.

One consequence of such a non-equilibrium and non-ergodic functional dynamics is that for an ensemble of identical enzymes under the same conditions, not all of them will behave the same, in fact, as pointed by ref. [90], there will be a population split, where different enzymes will show vastly different activity levels under the same environmental conditions. Such a behavior of have been actually observed in the early single molecule experiments [142], which the cumulative amount of the products by a single enzyme molecule was measured after a waiting time of *circa* 2 hours. The results showed that, despite the identical buffer environment, the activities between individual enzymes differ by many folds. However, the finding were interpreted as the supporting evidence for static disorder stating that the same protein species can fold into different functional states with various activity levels, a notion that has been rejected by the later single-molecule experiments,

which could resolve the time series of individual single enzyme catalysis events, in favor of dynamics disorder [36, 80, 136, 141]. Not surprisingly, the conclusion of static disorder was made based on the presumption that the single-molecule catalytic activity was fully stationary. However, for non-equilibrium systems in which Boltzmann-Khinchin ergodicity hypothesis is invalid, the dynamics can appear to be stationary over a long period of observation time, until on even longer timescales, another slower mode of the dynamics will enter the overall relaxation process and changes the time averaged quantities observed over the current observation time window.

Another practical implication of our results is that for any ensemble averaged experiments involving proteins. Often, the ensemble averages of the measured dynamical quantities are used to infer the corresponding time averaged behavior of a single protein molecule. However, if protein dynamics are indeed intrinsically non-ergodic and out of equilibrium, such measured ensemble average may not necessarily reflect the time averaged behavior of an individual protein molecules. However, in the praxis, such assumption of ergodicity is made implicitly in many ensemble based experiments, such as neutron scattering or NMR experiments of proteins. In this case, experimental measured dynamical quantities such as intermediate scattering function may not necessarily reflect the time averaged behavior of single protein molecules. Furthermore, findings presented in this dissertation also raises many open questions about protein dynamics and much future works can be carried out to address these questions, for example:

- In our noisy continuous time random walk (CTRW) model, the short-term noise is described by an empirical stretch exponential function. It would be interesting to further investigate the behavior of the noise that masks the underlying CTRW. As we suggested, the noises are results of the structural fluctuations of the protein while being trapped inside of a local minimum on the energy landscape. However, the detailed physical properties of the noise still remains elusive. The dynamical signatures of these non-Brownian fluctuations within a trap also can yield clues on the local properties of the energy landscape surface. Analytical models such as subordination of stochastic processes [89, 127] may be able to combine the subdiffusive CTRW with some other ergodic subdiffusive processes, such as FLE processes or fractional Brownian motion as the short time noise and capture the full protein dynamics in a consistent fashion.

- The network model offers a promising way to characterize protein dynamics in conjunction with the energy landscape. It is indeed interesting that the self-similar dynamics of the protein structural fluctuation is reflected by the self-similarities in the underlying topology and geometry of the energy landscape. However, further quantitative investigations are required in order to fully understand the relationships between these two observed self-similarities. Furthermore, a continuous time random walk scheme on the fractal [89] applied

to the conformational cluster transition network (CCTN) is expected to reproduce the protein dynamics, as observed in MD over the sampled portion of energy landscape. However, the detailed scheme as of how to exactly determine the parameters of the power-law waiting times and as well as the stochastic functions determining how the fictive walker should propagate across the network are still work in progress.

- There is still a gap on timescales from $10^2\mu$s to milliseconds over which the knowledge of single-molecule protein dynamics are still missing due to technical limitations in experimental techniques and the available speed of current computers. Thank to the development of more specialized computer hardwares, such as the ANTON machine [122], there are good chance that MD simulations on these timescales can be performed in the very near future.

The data presented in this dissertation calls for a new mindset to consider the protein dynamics and function. Moreover, what is the biological implication of such non-ergodic and non-stationary protein dynamics? In the case of the population separation as discussed above, how does the cell cope with an ensemble of vastly differently efficient enzymes? Assuming the energy expense to synthesize two enzymes of the same species are the same, it would mean for the same amount energy investment per enzyme, the cell cannot expect a well-defined average amount of work in return. This would be an obvious flaw in the cell's energy budget and allocation. However, one needs to keep in mind that *complex systems*, such as a cell, are a highly dynamical entities with many interacting components with well-defined operational procedures and various positive or negative feedback mechanisms, where activities are regulated in a *dynamical* fashion. If the cell happens to have synthesized a batch of "lazy" enzymes that do not catalyze enough downs-stream products, the low product concentrations can be detected and up-regulates the transcription and translation of the required enzyme, which then leads to an increase of the desired catalytic activity and level of the product. Once the sufficient product concentration is reached, the cell can start to degrade the enzyme back to smaller building blocks for other biochemical pathways. Such self-regulation and self-organization behaviors are characteristic signatures for complex systems [71]. When dealing with stochastic processes in real complex systems, one can no longer assume simple behaviors that fully obey the central limit theorem (thus Gaussian statistics) and Boltzmann-Khinchin ergodic hypothesis *a priori*. Rather, the Lévy-Gnedenko generalized central limit theorem leading to Lévy statistics for which not all moments of the distributions may exist. In this sense, the Gaussian behavior is just a special case among an infinite number of other possible probability distributions, towards one of them the system converges.

This picture offered by the generalized central limit theorem and the associated Lévy statistics is obviously much complicated to deal with compare to the simple and elegant Gaussian behavior, where all moments are defined and all properties predictable. There will be surely reluctance in

adopting the idea of non-equilibrium, non-ergodic approach to treat the dynamics observed in biological systems. However, blind application of the simple, and perhaps more "elegant", Gaussian assumption for complex systems can lead to disastrous results. One best example came from the complex system of financial market [71], the failure one of largest hedge funds ever existed, the long term capital management (LTCM) in the late 1990s. LTCM was famous for its high intellectual capacity – headed by two Nobel laureates in economy, Myron Scholes and Robert C. Merton, and supported by a small army of Ivy-league PhDs who applied sophisticated mathematical models to invest fund's capital. Their spectacular failure was attributed to their strict assumption of Gaussian statistics of the stock derivatives which led to gross under-estimations of the intrinsic market risks posed by events in the tail of the distribution, which would be virtually impossible according to Gaussian statistics [79]. At the end, U. S. Federal Reserve had to step in and orchestrate a rescue to ensure market stability. For protein dynamics, the non-equilibrium, non-ergodic picture is clearly more inconvenient to deal with comparing to the simple Brownian picture of the protein structural fluctuation [88]. However, in the proper context of complex systems, a deviation from simple behavior should be very much expected.

Bibliography

[1] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. in print.

[2] A. Addlagatta, L. Gay, and B. W. Matthews. *PNAS*, 103(36):13339, 2006.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[4] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.*, 47:1309–1314.

[5] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14(24):5355–5373, 1975.

[6] L. Bachelier. Théorie de la spéculation. 3(17).

[7] A. Bairoch and B. Boeckmann. The swiss-prot protein sequence data bank. *Nucl. Acids Res.*

[8] P. Bak, C. Tang, and K. Wiesenfeld. *Phys. Rev. Lett.*, 59:381, 1987.

[9] P. Bak, C. Tang, and K. Wiesenfeld. *Phys. Rev. A.*, 38(1):364, 1988.

[10] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. USA*, 103(35):13004–13009, 2006.

[11] H. J. C. Berendsen, J. P. M. Postma, A. DiNola, and J. R. Haak. *J. Chem. Phys.*, 81:3684, 1984.

[12] B. E. Bernstein, P. Michels, and W. Hol. *Nature*, 385:275, 1997.

[13] R. B. Best, X. Zhu, J. Shim, P. Lopes, J. Mittal, M. Feig, and A. D. MacKerell. *J. Chem. Theory Comput.*, 8:3257, 2012.

[14] S. M. Bezrukov and M. Winterhalter. *Phys. Rev. Lett.*, 85(1):202, 2000.

[15] G. D. Birkhoff. Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. USA*, 17(12):656–660, 1931.

[16] D. D. Boehr, H. J. Dyson, and P. E. Wright. An nmr perspective on enzyme dynamics. *Chem. Rev.*, 106:3055–3079, 2006.

[17] M. Born and R. Oppenheimer. Zur quantentheorie der molekülen. *Ann. Phys.*, 84:457–484, 1927.

[18] J. Bouchaud. *J.Phys. I France*, 2:1705, 1992.

[19] R. Brown. A brief account of microscopical observations on particles contained in the pollen of plants. *Phil. Mag.*, 4:161–173, 1882.

[20] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *PROTEINS: Structure, Function, and Genetics*, 21:167–195, 1995.

[21] S. Burov, J.-H. Jeon, R. Metzler, and E. Barkai. Single particle tracking in systems showing anomalous diffusion: the role of weak ergodicity breaking. *Phys. Chem. Chem. Phys.*, 13:1800–1812, 2011.

[22] S. Burov, R. Metzler, and E. Bakai. Aging and nonergodicity beyond the khinchin theorem. *Proc. Natl. Acad. Sci. USA*, 107(30):13228–13233, 2010.

[23] S. B. Cambridge, F. Gnad, C. Nguyen, J. L.Bermejo, M. Krüger, and M. Mann. Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *J. Proteome Res.*, 10:5275–5284, 2011.

[24] R. Clausius. Ueber die art der bewegung die wir wärme nennen. *Pogg. Ann. d. Phys. u. Chem.*, 100:353–380, 1857.

[25] D. B. Craig, E. A. Arriaga, J. C. Y. Wong, H. Lu, and N. J. Dovichi. Studies on single alkaline phosphatase molecules: Reaction rate and activation energy of a reaction catalyzed by a single molecule and the effect of thermal denaturation - the death of an enzyme. *J. Am. Chem. Soc.*, 118(22):5245–5253, 1996.

[26] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.*, 35:539–546, 2010.

[27] S. Cusack and W. Doster. *Biophys. J.*, 58:243, 1990.

[28] R. M. Daniel, J. L. Finney, and J. C. Smith. The dynamic transition in proteins may have a simple explanation. *Faraday Discuss.*, 122:163–169, 2002.

[29] T. Darden, D. York, and L. Pedersen. *J. Chem. Phys.*, 98:10089, 1993.

[30] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. *Angew. Chem. Int. Ed.*, 38:236, 1999.

[31] T. G. Dewey and J. G. Bann. *Biophys. J.*, 63:594, 1992.

[32] K. A. Dill. Folding proteins: Finding a needle in a haystack. *Curr. Opinion Struct. Biol.*, 3:99–103, 1993.

[33] W. Doster, S. Cusack, and W. Petry. *Nature*, 337:754, 1989.

[34] A. Einstein. Über die von molekularkinetischen theorie der waerme geforderte bewegung von in ruhrenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*, 17:549–560, 1905.

[35] E. Z. Eisenmesser, D. A. Bosco, M. Akke, and D. Kern. Enzyme dynamics during catalysis. *Science*, 295:1520–1523, 2002.

[36] B. P. English, W. Min, A. M. van Oijen, K. T. Lee, G. Luo, H. Sun, B. J. Cherayil, S. C. Kou, and X. S. Xie. Ever-fluctuating single enzyme molecules: Michaelis-menten equation revisited. *Nat. Chem. Biol.*, 2:87–94, 2006.

[37] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. *J. Chem. Phys.*, 103:8577, 1995.

[38] M. Ferrand, A. J. Dianoux, W. Petry, and G. Zaccaï. Thermal motions and function of bacteriorhodopsin in purple membranes: effects of temperature and hydration studied by neutron scattering. *Proc. Natl. Acad. Sci. USA.*, 90(20):9668–9672, 1993.

[39] A. Fick. Über diffusion. *Pogg. Ann. d. Phys. u. Chem.*, 44:59–86, 1855.

[40] J.-B. J. Fourier. *Théorie analytique de la chaleur*. Paris: F. Didot, 1822.

[41] H. Frauenfelder. *Nature Struct. Biol.*, 2:821, 1995.

[42] H. Frauenfelder and D. T. Leeson. *Nature Struct. Biol.*, 5(9):757, 1998.

[43] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou. *Nature*, 280:558, 1979.

[44] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motion of proteins. *Science*, 254:1598–1603, 1991.

[45] H. Frauenfelder, P. G. Wolynes, and R. H. Austin. Biological physics. *Rev. Mod. Phys.*, 71(2):S419–S430, 1999.

[46] F. Gabel, D. Bicout, U. Lehnert, M. Tehei, M. Weik, and G. Zaccai. Protein dynamics studied by neutron scattering. *Q. Rev. Biophys.*

[47] S. J. Hagen and W. A. Eaton. *J. Chem. Phys.*, 104(9):3395, 1996.

[48] G. Haran, E. Haas, B. K. Szpikowska, and M. T. Mas. *PNAS*, 89:11764, 1992.

[49] T. L. Hill. *Statistical Mechanics - Principles and Selected Applications*. Dover Publications, INC. New York, 1956.

[50] R. W. Hockney, S. P. Goel, and J. Eastwood. Quiet high resolution computer models of a plasma. *J. Comp. Phys.*, 14:148–158, 1974.

[51] L. Hong, N. Smolin, B. Lindner, A. P. Sokolov, and J. C. Smith. Three classes of motion in the dynamic neutron-scattering susceptibility of a globular protein. *Phys. Rev. Lett.*, 107:148102, 2011.

[52] X. Hu, L. Hong, M. D. Smith, T. Neusius, X. Cheng, and J. C. Smith. The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nat. Phys.*, 12:171–174.

[53] R. Inoue, R. Biehl, T. Rosenkranz, J. Fitter, M. Monkenbusch, A. Radulescu, B. Farago, and D. Richter. Large domain fluctuations on 50-ns timescale enable catalytic activity in phosphoglycerate kinase. *Biophys. J.*, 99:2309–2317, 2010.

[54] R. Ishima and D. A. Torchia. Protein dynamics from nmr. *Nat. Struct. Biol.*, 7(8):740–743, 2000.

[55] J.-H. Jeon, E. Barkai, and R. Metzler. Noisy continuous time random walks. *J. Chem. Phys.*, 139:121916, 2013.

[56] J.-H. Jeon, N. Leijnse, L. B. Oddershede, and R. Metzler. Anomalous diffusion and power-law relaxation of the time averaged mean squared displacement in worm-like micellar solutions. *New J. Phys.*, 15:045011.

[57] J.-H. Jeon and R.Metzler. Inequivalence of time and ensemble averages in ergodic systems: Exponential versus power-law relaxation in confinement. *Phys. Rev. E*, 85:021147, 2012.

[58] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. *J. Chem. Phys.*, 79:926, 1983.

[59] A. D. MacKerell Jr. et al. *J. Phys. Chem. B*, 102:3586, 1998.

[60] M. Karplus. Aspects of protein reaction dynamics: Deviations from simple behavior. *J. Chem. Phys. B.*, 104:11–27, 2000.

[61] M. Karplus and J. A. McCammon. Dynamics of proteins: elements and function. *Ann. Rev. Biochem.*, 53:263, 1983.

[62] H. Keller and P. G. Debrunner. *Phys. Rev. Lett.*, 1980.

[63] D. Kern and E. R. P. Zuiderweg. The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.*, 13:748–757, 2003.

[64] O. Khersonsky and D. S. Tawfik. Enzyme promiscuity: a mechanistic and evolutionary perspective.

[65] I. R. Kleckner and M. P. Foster. An introduction to nmr-based approaches for measuring protein dynamics. *Biochimica et Biophysica Acta*, 1814:942–968, 2011.

[66] E. W. Knapp, S. F. Fischer, and F. Parak. Protein dynamics from mössbauer spectra. the temperature dependence. *J. Phys. Chem.*, 86:5042, 1982.

[67] S. C. Kou. Stochastic modeling in nanoscale biophysics: Subdiffusion within proteins. *Ann. Appl. Stat.*, 2:501–535, 2008.

[68] H. A. Kramer. Brownian motion in a field of force and the diffusion model of chemical reaction. *Physica*, 7:284–304, 1940.

[69] R. Kubo. The fluctuation-dissipation theorem. *Rep. Prog. Phys.*, 25:255–284, 1966.

[70] K. Kuczera, J.-C. Lambry, J.-L. Martins, and M. Karplus. Nonexponential relaxation after ligand dissociation from myoglobin: A molecular dynamics simulation. *Proc. Natl. Acad. Sci. USA*, 90:5805–5807, 1993.

[71] J. Kwapień and S. Drożdża. Physical approach to complex systems. *Phys. Rep.*, 515:115–226, 2012.

[72] D. G. Lambright, S. Balasubramanian, and S. G. Boxer. *Chem. Phys.*, 158:249, 1991.

[73] P. Langevin. *C. R. hebd. Acad. Sciences*, 146:530–533, 1908.

[74] J. L. Lebowitz and O. Penrose. Modern ergodic theory. *Physics Today*, 26(2):23–29, 1973.

[75] M. Lim, T. A. Jackson, and P. A. Anfinrud. *PNAS*, 90:5801, 1993.

[76] D. J. Lipman, A. Souvorov, E. V. Koonin, R. Panchenko, and T. A. Tatusova. The relationship of protein conservation and sequence length. *BMC Evolutionary Biology*, 2:20, 2002.

[77] J. R. Loofbourow. Borderland problems in biology and physics. *Rev. Mod. Phys.*, 12:267–358, 1940.

[78] J. P. Loria, R. B. Berlow, and E. D. Watt. Characterization of enzyme motions by solution nmr relaxation dispersion. *Acc. Chem. Res.*, 41:214–221, 2008.

[79] R. Lowenstein. *When Genius Failed: The Rise and Fall of Long-Term Capital Management*. Random House Trade Paperbacks; Reprint edition, 2001.

[80] H. P. Lu, L. Xun, and X. S. Xie. Single-molecule enzymatic dynamics. *Science*, 282:1877–1882, 1998.

[81] G. Luo, I. Andricioaei, X. S. Xie, and M. Karplus. Dynamic distance disorder in proteins is caused by trapping. *J. Chem. Phys. B*, 110:9363–9367, 2006.

[82] J. A. MacCammon, B. R. Gelin, and M. Karplus. The hinge-bending mode in lysozyme. *Nature*, 262:325–326, 1976.

[83] A. D. MacKerell, Jr., M. Feig, and III. C. L. Brooks. *J. Am. Chem. Soc.*, 126:698, 2004.

[84] S. Magazù, F. Migliardo, and A. Benedetto. Puzzle of protein dynamical transition. *J. Chem. Phys. B.*, 115(24):7736–7743, 2011.

[85] M. Magdziarz and J. Klafter. Detecting origins of subdiffusion: P-variation test for confined systems. *Phys. Rev. E.*, 82:011129, 2010.

[86] M. Magdziarz and A. Weron. Fractional fokker-planck dynamics: Stochastic representation and computer simulation. *Phys. Rev. E*, 75:016708, 2007.

[87] M. Magdziarz, A. Weron, K. Burnecki, and J. Klafter. Fractional brownian motion versus the continuous-time randomwalk: A simple test for subdiffusive dynamics. *Phys. Rev. Lett.*, 103:180602, 2009.

[88] J. A. McCammon, B. R.Gelin, M. Karplus, and P.G. Wolynes. The hinge-bending mode in lysozyme. *Nature*, 262:325–326, 1976.

[89] Y. Meroz, I. M. Sokolov, and J. Klafter. Subdiffusion of mixed origins: When ergodicity and nonergodicity coexist. *Phys. Rev. E*, 81:010101(R), 2010.

[90] R. Metzler. Forever aging. *Nat. Phys.*, 12:113–114, 2016.

[91] R. Metzler, E. Bakai, and J. Klafter. Anomalous diffusion and relaxation close to thermal equilibrium: A fractional fokker-planck equation approach. *Phys. Rev. Lett.*, 82(18):3563–3567, 1999.

[92] R. Metzler, J.-H. Jeon, A. G. Cherstvy, and E. Barkai. Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.*, 16:24128–24164, 2014.

[93] R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Phys. Rep.*, 339:1–77, 2000.

[94] R. Metzler and J. Klafter. The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. *J. Phys. A: Math. Gen.*, 37:R161–R208, 2004.

[95] W. Min, G. Luo, B. J. Cherayil, S. C. Kou, and X. S. Xie. Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Phys. Rev. Lett.*, 94:198302, 2005.

[96] Wei Min, B. P. English, G. Luo, B. J. Cherayil, S. C. Kou, and X. S. Xie. Fluctuating enzymes: Lessons from single-molecule studies. *Acc. Chem. Res.*, 38:923–931, 2005.

[97] E. Montroll and M. F. Schlesinger. *PNAS*, 79:3380, 1982.

[98] E. W. Montroll. Random walks on lattices. iii. calculation of first-passage times with application to exciton trapping on photosynthetic units. *J. Math. Phys.*, 10(4):753–765, 1969.

[99] T. N. Narasimhan. Fourier's heat conduction equation: History, influence, and connections. *Rev. Geophys.*, 37(1):151–172, 1999.

[100] T. Neusius, I. Daidone, I. M. Sokolov, and J. C. Smith. *Phys. Rev. E.*, 83:021902, 2011.

[101] T. Neusius, I. M. Sokolov, and J. C. Smith. *Phys. Rev. E.*, 80:011109, 2009.

[102] Thomas Neusius. *Thermal Fluctuations of Biomolecules. An Approach to Understand the Subdiffusion in the Internal Dynamics of Peptides and Proteins*. PhD thesis, University of Heidelberg, 2009.

[103] G. U. Nienhaus and R. D. Young. Protein dynamics. *Encyclopedia Appl. Phys.*, 15:163–184, 1996.

[104] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. *J. Chem. Phys.*, 126:155102, 2007.

[105] S. Nosé. *J. Chem. Phys.*, 81:511, 1984.

[106] F. Parak, E. N. Frolov, A. A. Kononenko, R. L. Mössbauer, V. I. Goldanski, and A. B. Rubin. Evidence for a correlation between the photoinduced electron transfer and dynamic properties of the chromatophore membranes from rhodospirillum rubrum. *FEBS Lett.*, 117(1):368–372, 1980.

[107] F. Paraka, E. W. Knappa, and D. Kucheida. Protein dynamics: Mössbauer spectroscopy on deoxymyoglobin crystals. *J. Mol. Biol.*, 161(1):177, 1982.

[108] M. Parrinello and A. Rahman. *J. Appl. Phys.*, 52:7182, 1981.

[109] K. Pearson. The problem of the random walk. *Nature*, 72:294, 1905.

[110] J. W. Petrich, J.-C. Lambry, K. Kuczera, M. Karplus, C. Poyart, and J.-L. Martin. Ligand binding and protein relaxation in heme proteins: A room temperature analysis of no geminate recombination. *Biochemistry*, 30:3975–3987, 1993.

[111] J. Philibert. One and a half century of diffusion: Fick, einstein, before and beyond. *Diffusion Fundamentals*, 4:6.1–6.19, 2006.

[112] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. *Bioinformatics*, 29:845, 2013.

[113] A. Rahman. Correlations in the motion of atoms in liquid argon. *Phys. Rev.*, 136(2A):A405–A411, 1964.

[114] B. F. Rasmussen, A. M. Stock, D. Ringe, and G. A. Petsko. Crystalline ribonuclease a loses function below the dynamical transition at 220 k. *Nature*, 357:423–424, 1992.

[115] V. Réat, R. Dunn, M. Ferrand, J. L. Finney, R. M. Daniel, and J. C. Smith. Solvent dependence of dynamic transitions in protein solutions. *Proc. Natl. Acad. Sci. USA.*, 97(18):9961–9966, 2000.

[116] L. E. Reichl. *Modern Course in Statistical Physics, 2nd Ed.* JOHN WILEY & SONS, INC., 1998.

[117] E. Reynaud. Protein misfolding and degenerative diseases. *Nature Education*, 3(9):28, 2010.

[118] H. Scher and E. W. Montroll. Anomalous transit-time dispersion in amorphous solids. *Phys. Rev. B*, 12(6):2455–2477, 1975.

[119] M. Schlesinger. Asymptotic solutions of continuous-time random walks. *J. Stat. Phys.*, 10:421–434, 1974.

[120] J. H. P. Schulz, E. Barkai, and R. Metzler. Aging renewal theory and application to random walks. *Phys. Rev. X*, 4:011028, 2014.

[121] R. K. Scopes. The steady-state kinetics of yeast phosphoglycerate kinase anomalous kinetic plots and the effects of salts on activity. *Euro. J. Biochem.*, 85:503–516, 1978.

[122] D. E. Shaw, R. O. Dror, J. K. Salmon, J.P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC09)*, Portland, Oregon, November 14, 2009.

[123] T. Y. Shen, K. Tai, R. H. Henchmann, and J. A. McCammon. *Acc. Chem. Res.*, 35, 332.

[124] T. Y. Shen, K. Tai, and J. A. McCammon. *Phys. Rev. E*, 63:041902, 2002.

[125] Z. Siwy and A. Fulinski. *Phys. Rev. Lett.*, 89(15):158101, 2002.

[126] R. G. Smock and L. M. Gierasch. Sending signals dynamically. *Science*, 324:198–203, 2009.

[127] I. M. Sokolov and J. Klafter. From diffusion to anomalous diffusion: A century after einstein's brownian motion. *Chaos*, 15:026103, 2005.

[128] C. Song, L. K. Gallos, S. Havlin, and H. Makse. How to calculate the fractal dimension of a complex network: the box covering algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 207(3):P03006, 2007.

[129] C. Song, S. Havlin, and H. A. Makse. Self-similarity of complex networks. *Nature*, 433:392–395, 2005.

[130] R. F. Tilton, J. C. Dewan, and G. A. Petsko. *Biochemistry*, 31:2469, 1992.

[131] B. H. Toyama, J. N. Savas, S. K. Park, M. S. Harris, N. T. Ingolia, J. R. Yates III., and M. W. Hetzer. Identification of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell*, 154:971, 2013.

[132] A. P. Valente, C. A. Miyamoto, and F. C. L. Almeida. Implications of protein conformational diversity for binding and development of new biological active compounds. *Curr. Med. Chem.*, 13:3697–3703, 2006.

[133] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159:98, 1967.

[134] M. von Smoluchowski. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. *Annalen der Physik*, 326(14):756–780, 1906.

[135] D. J. Wales. *Energy Landscapes*. Cambridge University Press, Cambridge, 2003.

[136] Y. Wang and H. P. Lu. *J. Phys. Chem. B*, 114:6669, 2010.

[137] H.C. Watson, N. P. Walker, P. J. Shaw, T. N. Bryant, P. L. Wendell, L. A. Fothergill, R. E. Perkins, S. C. Conroy, M. J. Dobson, and M. F. Tuite. *EMBO. J.*, 1:1635, 1982.

[138] G. H. Weiss and R. J. Rubin. Random walks: Theory and selected applications. *Adv. Chem. Phys.*, 52:363, 1983.

[139] M. B. Weissman. *Rev. Mod. Phys.*, 60(2):537, 1988.

[140] P. Welch. *IEEE Trans. Audio Electroacoust.*, 15:70, 1967.

[141] X. S. Xie. Single-molecule approach to dispersed kinetics and dynamic disorder: Probing conformational fluctuation and enzymatic dynamics. *J. Chem. Phys.*, 117(24):11024–11032, 2002.

[142] Q. Xue and E. S. Yeung. Differences in the chemical reactivity of individual molecules of an enzyme. *Nature*, 373:681–683, 1995.

[143] H. Yang, G. Luo, P. Karnchanaphanurach, T.-M. Louie, I. Rech, S. Cova, L. Xun, and X. S. Xie. Protein conformational dynamics probed by single-molecule electron transfer. *Science*, 302:262–266, 2003.

[144] X. Zhuang, H. Kim, M. J. B. Pereira, H. P. Babcock, N. G. Walter, and S. Chu. Correlating structural dynamics and function in single ribozyme molecules. *Science*, 296:1473–1476, 2002.

[145] R. Zwanzig. Rate processes with dynamical disorder. *Acc. Chem. Res.*, 23:148–152, 1990.

[146] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.*, 89:20–22, 1992.

Appendices

Appendix A

Relationship between the MSD and ACF

A.1 Relationship between ACF and MSD in case of stationary time series

In this section, we use $\langle A \rangle_t$ to denote the time average of the quantity A over the time length t. TA-MSD of a stochastic time series $R(t)$ can be rewritten as the following:

$$\overline{\delta^2(\Delta)} = \left\langle [R(t+\Delta) - R(t)]^2 \right\rangle_{t_{\max}-\Delta} \tag{A.1}$$

$$= \left\langle R^2(t+\Delta) + R^2(t) - 2R(t+\Delta)R(t) \right\rangle_{t_{\max}-\Delta} \tag{A.2}$$

$$= \left\langle R^2(t+\Delta) \right\rangle_{t_{\max}-\Delta} + \left\langle R^2(t) \right\rangle_{t_{\max}-\Delta} - 2\left\langle R(t+\Delta)R(t) \right\rangle_{t_{\max}-\Delta} \tag{A.3}$$

and the unnormalized ACF $C'(\Delta)$ as the following:

$$C'(\Delta) = \left\langle [R(t+\Delta) - \langle R \rangle]\,[R(t) - \langle R \rangle] \right\rangle_{t_{\max}-\Delta} \tag{A.4}$$

$$= \left\langle R(t+\Delta)R(t) - R(t+\Delta)\langle R \rangle - R(t)\langle R \rangle + \langle R \rangle^2 \right\rangle_{t_{\max}-\Delta} \tag{A.5}$$

$$= \left\langle R(t+\Delta)R(t) \right\rangle_{t_{\max}-\Delta} + \langle R \rangle^2 - \langle R \rangle \left\langle R(t+\Delta) + R(t) \right\rangle_{t_{\max}-\Delta} \tag{A.6}$$

$$= \left\langle R(t+\Delta)R(t) \right\rangle_{t_{\max}-\Delta} + \langle R \rangle^2 - \langle R \rangle \left(\left\langle R(t+\Delta) \right\rangle_{t_{\max}-\Delta} + \left\langle R(t) \right\rangle_{t_{\max}-\Delta} \right). \tag{A.7}$$

The *stationarity* of $R(t)$ guarantees

$$\left\langle R^2(t+\Delta) \right\rangle_{t_{\max}-\Delta} = \left\langle R^2(t) \right\rangle_{t_{\max}-\Delta} = \left\langle R^2 \right\rangle \tag{A.8}$$

$$\left\langle R(t+\Delta) \right\rangle_{t_{\max}-\Delta} = \left\langle R(t) \right\rangle_{t_{\max}-\Delta} = \langle R \rangle, \tag{A.9}$$

therefore, using the relation A.8 and A.9, the MSD(Δ) and $C'(\Delta)$ can be rewritten to

$$\overline{\delta^2(\Delta)} = 2\left(\left\langle R^2\right\rangle - \langle R(t+\Delta)R(t)\rangle_{t_{\max}-\Delta}\right). \tag{A.10}$$

and

$$C'(\Delta) = \langle R(t+\Delta)R(t)\rangle_{t_{\max}-\Delta} - \langle R\rangle^2, \tag{A.11}$$

therefore

$$\Leftrightarrow \quad \langle R(t+\Delta)R(t)\rangle_{t_{\max}-\Delta} = C'(\Delta) + \langle R\rangle^2. \tag{A.12}$$

Insert A.12 into A.10, we get

$$\overline{\delta^2(\Delta)} = 2\left(\left\langle R^2\right\rangle - \langle R\rangle^2 - C'(\Delta)\right) = 2\left[\left\langle dR^2\right\rangle - C'(\Delta)\right]. \tag{A.13}$$

Since $C'(\Delta) = C'(0)C(\Delta)$ the normalization factor $C'(0)$ is

$$C'(0) = \langle [R(t+\Delta) - \langle R\rangle]\,[R(t) - \langle R\rangle]\rangle_{t_{\max}-\Delta}\,|_{\Delta=0} \tag{A.14}$$

$$= \left\langle R^2(t)\right\rangle_{t_{\max}} - \langle R\rangle^2 \tag{A.15}$$

$$\equiv \left\langle R^2\right\rangle - \langle R\rangle^2 = \left\langle dR^2\right\rangle \tag{A.16}$$

Therefore, the Eq. A.13 becomes

$$\overline{\delta^2(\Delta)} = 2\left\langle dR^2\right\rangle [1 - C(\Delta)] \tag{A.17}$$

where $\langle dR^2\rangle$ is the variance of $R(t)$.

A.2 General relationship between ACF and MSD

If the stationary conditions A.8 and A.9 are not satisfied, we can re-write MSD and ACF using A.3 and A.7, respectively, as

$$\langle R(t+\Delta)R(t)\rangle_{t_{\max}-\Delta} = \frac{\langle R^2(t+\Delta)\rangle_{t_{\max}-\Delta} + \langle R^2(t)\rangle_{t_{\max}-\Delta} - \overline{\delta^2(\Delta)}}{2}, \tag{A.18}$$

and

$$\langle R(t+\Delta)R(t)\rangle_{t_{\max}-\Delta} = C'(\Delta) - \langle R\rangle^2 + \langle R\rangle\left(\langle R(t+\Delta)\rangle_{t_{\max}-\Delta} + \langle R(t)\rangle_{t_{\max}-\Delta}\right) \tag{A.19}$$

$$= \left\langle dR^2\right\rangle \cdot C(\Delta) - \langle R\rangle^2 + \langle R\rangle\left(\langle R(t+\Delta)\rangle_{t_{\max}-\Delta} + \langle R(t)\rangle_{t_{\max}-\Delta}\right). \tag{A.20}$$

By setting right-hand sides of Eqs. A.18 and A.20 to equal, we get

$$\overline{\delta^2(\Delta)} = A - 2\left\langle dR^2 \right\rangle \cdot C(\Delta) \tag{A.21}$$

with

$$A = 2\left\langle R \right\rangle^2 - 2\left\langle R \right\rangle \left(\left\langle R(t+\Delta) \right\rangle_{t_{\max}-\Delta} + \left\langle R(t) \right\rangle_{t_{\max}-\Delta} \right) + \left\langle R^2(t+\Delta) \right\rangle_{t_{\max}-\Delta} + \left\langle R^2(t) \right\rangle_{t_{\max}-\Delta}. \tag{A.22}$$

It can be easily verified that, $A = 2(\langle R^2 \rangle - \langle R \rangle^2) = 2\langle dR^2 \rangle$, if conditions A.8 and A.9 are satisfied. In this case, the stationary relationship (A.17) is recovered. In general, the quantity A can depend on the lag time Δ and the maxervation time length $t_{\max}$, i.e. $A = A(\Delta, t_{\max})$ and variance can be dependent on $t_{\max}$, i.e. $\langle dR^2 \rangle = \langle dR^2(t_{\max}) \rangle$, therefore

$$\overline{\delta^2(\Delta)} = A(\Delta, t_{\max}) - 2\left\langle dR^2(t_{\max}) \right\rangle \cdot C(\Delta). \tag{A.23}$$

Appendix B

TA-MSDs and ACFs data for the structural dynamics of PGK, K-Ras and ePepN

B.1 Inter-domain dynamics of PGK

Table B.1: KWW fit parameters obtained from the fit of the ACFs of the PGK inter-domain distance time series using Eq. 4.12.

Observation time t	KWW exponent β	KWW parameter τ (ps)
10 ns	0.028	18.30
500 ns	0.0039	24.47
17 μs	0.0022	24.55

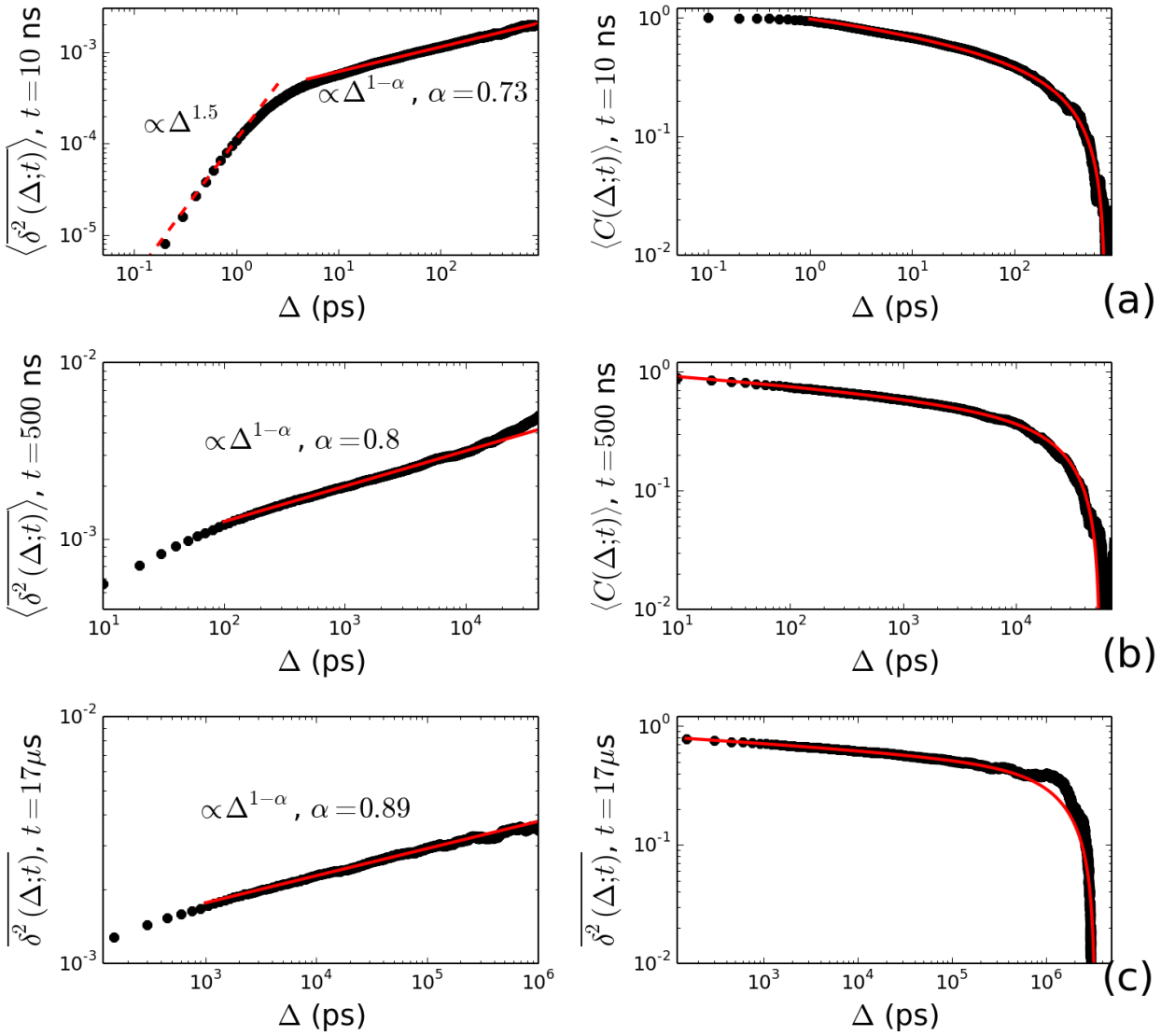

Figure B.1: Fits of the TA-MSD (left panel) and ACF (right panel) of PGK inter-domain distance trajectories at different observation times t using a power-law and Eq. (4.12), respectively. (a) $t = 100$ ps, (b) $t = 10$ ns and (c) $t = 17\mu$s. The KWW-parameters in Eq. 4.12 obtained from the fit are shown in Tab. B.1.

B.2 Inter-segment dynamics of K-Ras

Table B.2: KWW fit parameters obtained from the fitting of the ACFs of the time series of the K-Ras inter-segment distance between segment 1 and 2 using Eq. 4.12.

Observation time t	KWW exponent β	KWW parameter τ (ps)
10 ns	0.031	8.49
500 ns	0.22	88667.2

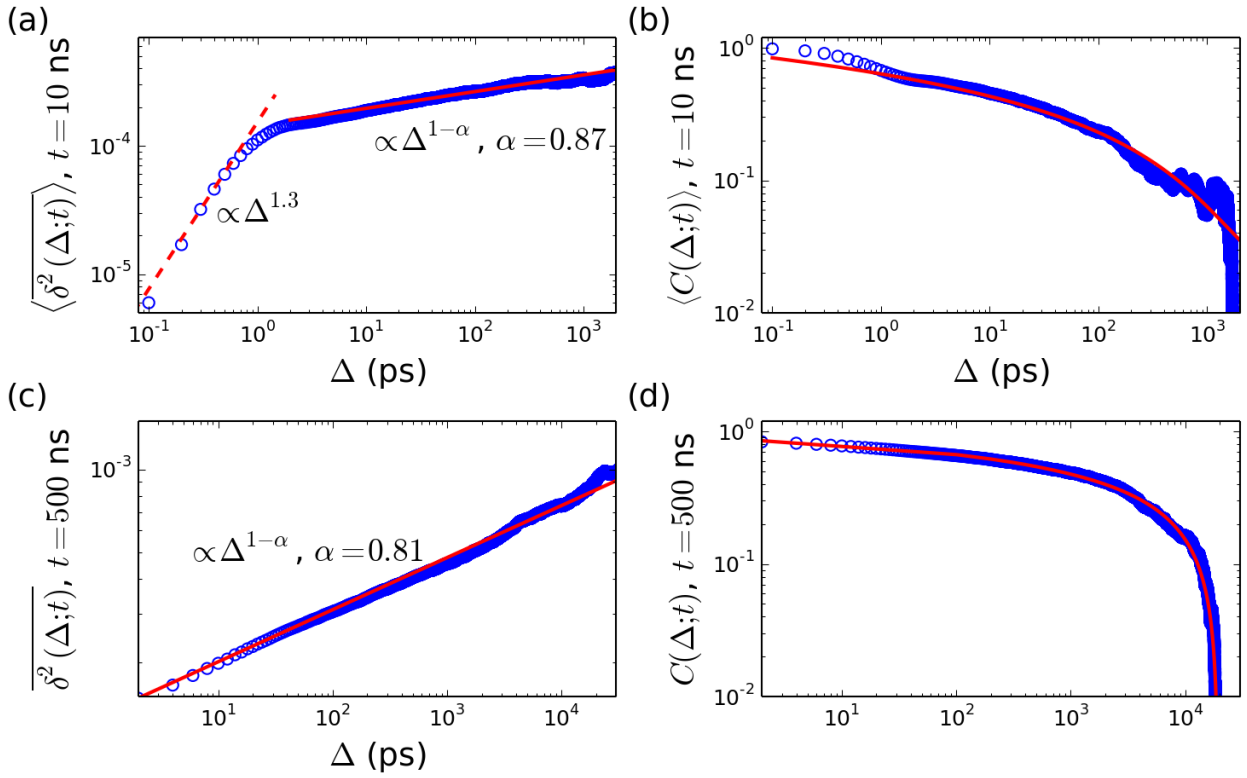

Figure B.2: Inter-segment distance dynamics of K-Ras. Segments as defined in figure caption of Fig. 4.1. **(a)** TA-MSD and ACF (Eq. 4.3, and Eq. 4.1, if only a single time series is available) of the domain motion. A power law is used to fit the TA-MSD and the ACF is fitted using the noisy CTRW model (Eq. 4.12) with a total observation time $t = 10$ ns. **(b)** TA-MSD and ACF with $t = 500$ ns. The KWW-parameters in Eq. 4.12 obtained from the fit are shown in Tab. B.2.

B.3 Inter-domain dynamics of ePepN

Table B.3: KWW fit parameters obtained from the fitting of the ACFs of the time series of the ePepN inter-domain distance between domains I and II using Eq. 4.12.

Observation time t	KWW exponent β	KWW parameter τ (ps)
10 ns	0.047	8.73
800 ns	0.0012	23.01

Table B.4: KWW fit parameters obtained from the fitting of the ACFs of the time series of the ePepN inter-domain distance between domains II and III using Eq. 4.12.

Observation time t	KWW exponent β	KWW parameter τ (ps)
10 ns	0.062	0.018
800 ns	0.0017	23.91

Table B.5: KWW fit parameters obtained from the fitting of the ACFs of the time series of the ePepN inter-domain distance between domains II and IV using Eq. 4.12.

Observation time t	KWW exponent β	KWW parameter τ (ps)
10 ns	0.036	11.23
800 ns	0.39	23702.2

Table B.6: KWW fit parameters obtained from the fitting of the ACFs of the time series of the ePepN inter-domain distance between domains IV and rest of the protein atoms (domains 1-3) using Eq. 4.12.

Observation time t	KWW exponent β	KWW parameter τ (ps)
10 ns	0.017	19.61
800 ns	0.31	107802

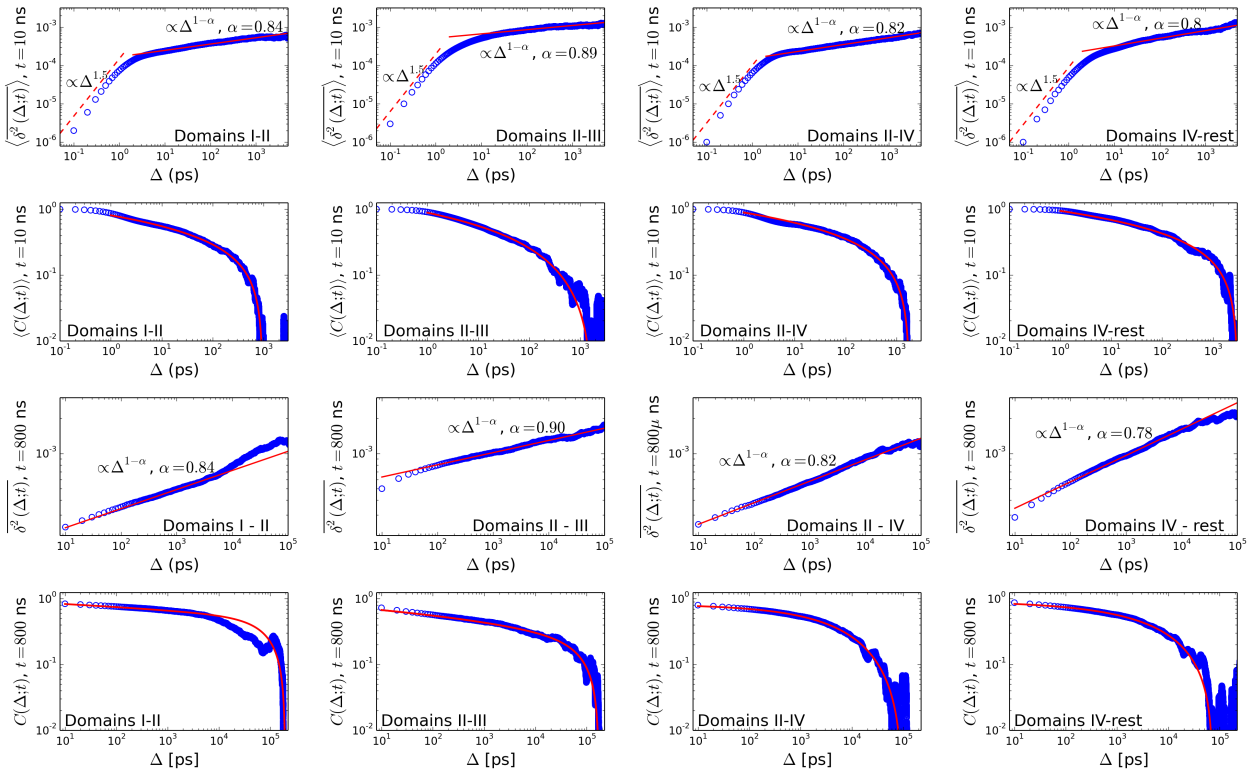

Figure B.3: Dynamics of the inter-domain distance trajectories of ePepN. First and third rows: TA-MSDs for different domain pairs with $t = 10$ ns and 800 ns, respectively. Second and fourth rows: ACFs for diffferent domain pairs with $t = 10$ ns and 800 ns, respectively. Each column contains the TA-MSD and ACF data of the inter-domain distance time series of a specific pair of domains at both observation timescales of 10 ns and 800 ns; column 1: domains I–II, column 2: domains II–III, column 3: domains II-IV, column 4L domains IV and rest of the protein atoms (*i.e.* domains I-III). All TA-MSDs are fitted by power law and all ACFs are fitted by the noisy CTRW model (Eq. 4.12). The results of the fit parameters β and τ of Eq. 4.12 in the are given in Tables B.3-B.6

Vitae

Xiaohu Hu studied physics at the Ruprecht-Karls-Universität Heidelberg (Heidelberg University), in Heidelberg, Germany and graduated with the degree Diplom Physiker (the former German M.Sc. equivalent degree in Physics) and with a minor in Biology in 2008. In 2009, he moved to Knoxville, Tennessee, USA to join the Ph.D. program Genome Science and Technology at the University of Tennessee, Knoxville with a major in Life Sciences. During his Ph.D., he performed research at the Center for Molecular Biophysics at the Oak Ridge National Laboratory under the supervision of Dr. Jeremy C. Smith. During this time, he published three research papers as the first author and co-authored a fourth paper. His research interests include protein dynamics and collective phenomena in complex systems.