Path-metadynamics: A computational study of conformational transitions in proteins

Díaz Leines, G.

Citation for published version (APA):
Chapter 1

Introduction

In this chapter we give an overview of conformational transitions in proteins and the role of computational modelling in understanding such complex processes. First, we introduce the protein folding problem, one of the most famous examples of conformational transitions in proteins. We present the free energy landscape view on protein folding as a funnel mechanism towards the native state, where the folding process takes place through an ensemble of pathways and different intermediate states. We then discuss how the folding funnel idea has been able to explain the Levithal’s paradox. We introduce the two main mechanisms proposed as models to describe the folding events. Next, we present the multidimensional free energy landscape view on proteins, proposed to define the probability distributions of conformational states and the free energy barriers separating them. From this point of view, the function of a protein in general is determined by its dynamical character rooted in the free energy landscape. The hierarchy of timescales and amplitudes of these protein motions, in connection with the energy landscape, is described. We then discuss briefly the recent advances and contributions of molecular dynamics simulations in modelling conformational changes in proteins and other complex transitions. Finally, we describe the view of conformational transitions in proteins as rare events and the importance of developing new computational algorithms to study complex systems. We end with an outline of the thesis.
1-1 Protein structure and function

Proteins are involved in basically all fundamental cell functions and therefore they are considered the machinery of life. They perform a variety of roles within the living organisms such as sensing and signalling, catalysis of metabolic reactions, structural support, transport of molecules from one location to another and replication of DNA [1].

Each protein begins as a polypeptide, translated from a sequence of mRNA into a linear chain of amino acids, also known as its primary structure. They are constituted from a set of 20 amino acids and their sequences are uniquely dictated by the information stored in the nucleotide sequence of their genes, usually of the length of a few hundred aminoacids. After proteins have been synthesized by the ribosome as linear chains, they fold into specific three-dimensional structures in order to become functional. The shape in which they fold determines its physiological role in the cell and it is known as the native conformation [1].

The native conformation possesses different levels of biomolecular structure [1; 2] referred as to: (1) the amino acid chain sequence known as the primary structure, (2) the secondary structure which consists of regular repeating local structures stabilized by hydrogen bonds, notably β-sheets and α- helical conformations, (3) the tertiary structure formed by the packing of secondary structures into three dimensional shapes stabilized by hydrophobic interactions, disulphide bonds, electrostatic interactions, hydrogen bonds and salt bridges and (4) the quaternary structure which consists of the assembly of several folded proteins forming a structure that functions as a single protein complex. Fig. 1.1 shows the different level of structures found in proteins.

The correct functioning of a protein within the cell depends entirely on the ability of the chain of amino acids to fold rapidly and reliably into its native structure. After the discovery of the tertiary structure [3; 4] it was realized that the folded structure makes proteins capable of performing many biological functions and that this three-dimensional shape is related to the chemical properties of the amino acid sequence. A failure to fold correctly, or remain correctly folded, usually produces inactive or dysfunctional proteins that can give rise to different diseases [5; 6; 7; 8; 9; 10]. Some of these diseases, such as cystic fibrosis [6] and some types of cancer [9], are caused by proteins folding incorrectly and not being able to perform their functions. In other cases, proteins with a high probability to misfold can escape and form aggregates within cells or (more commonly) in extracellular space. Several disorders, including Alzheimer’s and Parkinson’s diseases, the spongiform encephalopathies and type II diabetes, are directly associated with the deposition of such aggregates in tissues [7; 8; 11; 12].

In addition to their tertiary structure, proteins can undergo rearrangements in response to changes in their environment and other factors, visiting different conformations to meet their functional roles. These rearrangements are known as conformational transitions and they include processes such as allostery transitions in enzymes, force generation by motor proteins, the opening and closing of ion channels, and the
conformational changes induced by ligand binding to enzymes and receptors. This means that proteins are highly dynamical objects and the detailed molecular description of structural changes (folding and unfolding events) facilitates the understanding of the underlying mechanisms of their functions [2].

Even though nowadays we know that under ambient conditions most proteins fold into three-dimensional structures (their native conformation) and that the dynamical transitions of proteins between different conformations are essential to perform specific tasks, still there are open questions about the detailed mechanism or the pathway in which a conformational transition takes place. These questions remain important because having answers would allow us to construct predictive models of proteins that help in understanding, for instance, the origin of the different neurodegenerative disorders, or allow us to design ligands that can modulate the equilibria of conformational transitions and their rates.

This thesis addresses the development and application of computational methods to study the mechanism of conformational changes in proteins at the atomistic level. These methods aim to provide efficient and plausible models for protein folding that help in understanding their functions. We use molecular dynamics simulation techniques to study conformational changes in proteins and prove that the exploration of transition pathways in complex systems is now becoming accessible to computer simulations.

![Figure 1.1: Hierarchical structure of proteins. (a) The primary structure: A linear chain of amino-acids. (b) The secondary structure: the $\alpha$-helix and $\beta$-sheet structures are shown in all-atom representation with the hydrogen bond network [13], together with a cartoon representation of the structures. (c) The tertiary structure of the protein is shown as a mix of secondary structures: $\alpha$-helix (purple) and $\beta$-sheets (yellow). (d) The quaternary structure is a complex formed by proteins in their tertiary structure.](image)
CHAPTER 1. INTRODUCTION

1-2 The mechanism of protein folding

One of the most fundamental and universal examples of conformational transitions is the folding of proteins into their compact three-dimensional structures. Understanding the mechanism of protein folding provides unique insight in the way in which evolutionary selection has influenced the properties of a molecular system for functional advantage [10]. This example is interesting by itself and its study became one of the main motivations for this thesis. In this section we focus on the description of the protein folding problem as an example of conformational transitions.

1-2.1 Transition pathways

Understanding the molecular process of protein folding has presented a challenge for both experimentation and theory. Anfinsen et al. [14] showed that not only local structure but also the global three-dimensional structure of proteins can be reached reliably by the protein molecule through purely chemical processes without any special help of biological machinery and using only the information of the protein amino acid sequence. This was confirmed when X-ray diffraction revealed that proteins are apparently not simple repetitive structures but are compact objects with complex folds whose three-dimensional structure was hard to predict a priori [3; 4]. These findings inspired an intensive investigation of the mysterious mechanism by which a polypeptide chain folds to a specific three-dimensional protein structure.

However, a key question on how the correct fold emerged from the information of the amino acid sequence remained unclear: how, from an enormous number of accessible configurations possible for a random coil to be in, the native conformation is found within a timescale that ranges from microseconds up to minutes? That is, how is the energy landscape unique to a specific protein and defined by the sequence so that the protein reaches the tertiary structure on these timescales? This question and its answer is well known as the Levithal’s paradox [15]. Levithal calculated that the time needed to sample the available configurations of a relatively small protein of 100 amino acids would be in the order of 10^{10} years. Based on this finding, he concluded that there must be a bias towards the native state, a folding pathway, otherwise no proteins could fold within the lifetime of the universe [16; 15].

In general, the complex folding behaviour was hard to relate to any theoretical understanding of protein structure and energetics. An initial proposal was the idea that there must be a pathway for folding with a well-defined sequence of events which follow one another so as to carry the protein from the unfolded random coil to a uniquely folded metastable state [17; 18; 19]. In the late eighties and early nineties another formulation of the energy landscape from an statistical physics point of view was proposed to study protein folding [20; 21; 15; 22; 23]. In this view, often referred as the ‘new view’ of the protein folding, the folding process is described as a stochastic process of conformational changes by which the polypeptide chain immersed in water folds into its functional tertiary structure, reaching thermodynamic equilibrium from
1-2. THE MECHANISM OF PROTEIN FOLDING

a random coil and after visiting a variety of micro states. This stochastic process involves the molecular chain performing brownian motion where the water medium exerts random forces such as hydrogen bonds, thermal fluctuations in covalent bonds, bond angles, etc.

It is now clear that understanding protein folding involves a stochastic search of many conformations accessible to a polypeptide chain, rather than the evolution in a single pathway consisting of mandatory steps between specific partly folded states. According to this point of view, the energy landscape of a protein is a funnel described as a high dimensional rugged landscape with a complicated topological form containing traps where the protein stays for some time. There are multiple pathways towards the native structure, which is a global energy minimum in the energy landscape. Moreover, the energy landscape should have a funnel shape, so that only a small number of all possible conformations are sampled by any given protein molecule during its transition to the native conformation. It is believed that the landscape is encoded by the amino acid sequence and it is natural selection which has enabled proteins to evolve so that they are able to fold rapidly and efficiently [24; 25; 26].

Fig. 1.2 shows an schematic representation of the multiple folding pathways towards the native structure of a protein. After proteins are synthesized on the ribosomes, they fold into their three dimensional structures visiting different partially folded intermediates. However, there are cases in which proteins with high propensity to fold incorrectly, follow different routes through partial folded intermediates and misfolded states which can lead to the formation of protein aggregates. As we have mentioned previously, an increasing number of disorders including Alzheimer’s and Parkinson’s diseases are directly associated to potentially toxic aggregates such as as amyloid fibrils (see also Fig. 1.3).

1-2.2 The free energy landscape view

From the point of view of the stochastic nature of protein folding, each protein conformation has a free energy that determines its probability to be sampled at a certain temperature $T$ [27]. The free energy comprises the sum of enthalpic and entropic effects expressed as $F(Q) = E(Q) - TS_{\text{conf}}(Q)$ where $E$ is the internal energy of the protein conformation and the solvent (dispersion forces, electrostatic interactions, van der Waals potentials and hydrogen bonding), $Q$ is the conformational reaction coordinate and $S_{\text{conf}}$ is the conformational entropy described by hydrophobic interactions. The functional dependence of $E$ in all the degrees of freedom of the protein is called the energy landscape, which is thought to be rugged, containing a lot of minima with a global energy minimum [24; 1]. For low temperatures, only the energy landscape is relevant and the protein stays in a global minimum in its tertiary compact structure. When temperature increases, the conformational entropy term $S_{\text{conf}}$ will increase the chance of the protein to adopt more extended conformation. Statistical fluctuations give then rise to the ensemble of pathways towards the folded state.

The theoretical framework of the free energy landscape and the funnel concept has
CHAPTER 1. INTRODUCTION

Figure 1.2: Multiple states and pathways for folding and aggregation. The polypeptide chain can follow transition pathways through intermediate states (or through misfolded states) towards the native state or towards aggregation states. Figure adapted from Ref. [28].

Figure 1.3: A schematic free energy landscape for protein folding and aggregation. The landscape shows the different conformations and stable states ‘funneling’ towards the native state (via intramolecular contacts) or towards the aggregation state (via intermolecular contacts). The intra- and inter-molecular contacts represent the entropic effect of the system. Figure adapted from Ref. [29].
helped to understand the mechanism of protein folding. The protein does not have to follow a specific path through the configuration space, but instead it can travel through different pathways, sampling various partially folded structures on the free energy landscape. Moreover, the free energy surface funnels the multitude of denatured conformations to the unique native structure. Fig. 1.3 shows a schematic representation of the funnel landscape. The width of the funnel is a measure of the entropy or the number of micro-states that can be visited. Towards the native state the entropy decreases and the shape of the landscape reduces the search through configuration space and enables proteins to fold in a reasonable time. Larger polypeptides (more than 100 residues) normally have more rugged free energy landscapes that include on and off pathways, partially folded intermediate states that sometimes require the assistance of chaperones to fold into the native conformations. Above a critical protein concentration, the free energy landscape becomes even more complicated as polypeptides can interact to form aggregates, oligomers and even amyloid fibrils (see darker region of the free energy landscape in Fig. 1.3) [29; 10].

Protein stability depends on the free energy difference between the folded and unfolded states, expressed in terms of energy and entropy differences as $\Delta F = \Delta E - T\Delta S_{conf}$. As the binding energy term $\Delta E$ increases or the entropy difference between the folded and unfolded states decreases, the folded protein becomes more stable. Substantial perturbations of the folded conformation would then require a significant increase in free energy.

Two generic mechanisms for protein folding have been proposed to explain the free energy barriers that appear between different intermediate states: The so called nucleation-condensation (NC) model and the diffusion-collision (DC) model. Fig. 1.4 shows a schematic representation of both mechanisms and the free energy barriers associated to them. In the NC model the free energy barrier is associated to the formation of a folding nucleus needed to reach the native state [30]. According to this model, the number of contacts or non local interactions in the protein form a folding nucleus from which the condensation of the protein structure follows towards the native state. On the other hand, the DC model proposes that secondary structure elements are formed fast and are metastable [19]. In this model the polypeptides folds through a folding intermediate state where secondary structure is formed, followed by a stochastic search through the conformational space towards the native state. According to this model, the formation of secondary structure reduces considerably the number of conformations that the protein visits allowing the efficient and fast folding. It has been proposed that a combination of both mechanism models could act during the folding process of even small proteins [31].
1-3 Conformational changes in proteins

Although the idea of conformational transitions rooted in the free energy landscape is most familiar in the context of protein folding (the funnel hypothesis), in the last decades it has been proposed that the function of proteins in general is governed by their dynamical character, which can be described by a multidimensional free energy landscape [33]. From this point of view, proteins can sample a large ensemble of conformations around the average structure as a result of thermal energy. This means that proteins are marginally stable and even in the most favourable equilibrium state they can fluctuate from their average conformation. In a macroscopic conformation with a number of atoms around $10^{23}$ these fluctuations are negligible and not observable. But for proteins of a few thousands of atoms, these fluctuations can be pronounced. Their marginal stability is the reason why proteins can function, for instance, in signalling or regulation. A small stimulus like absorption of a photon by a chromophore,
electron transfer or the binding of a ligand can cause large conformational changes. Therefore, a description of proteins requires the definition of a free energy landscape that defines the probability distribution of conformational states (thermodynamics) and the free energy barriers between them (the kinetics).

The idea of the free energy landscape of folded proteins was introduced by Frauenfelder et al. in 1975 [34]. They performed experiments to measure the kinetics of carbon monoxide and oxygen rebinding to myoglobin and developed a free energy landscape model based on the observation of multiple free energy barriers and non exponential kinetics below a temperature of 230 K. They were able to successfully connect the free energy landscape to the myoglobin function and characterize the features of the free energy landscape, calculating the heights of barriers between stable states and localizing multiple conformational states. Since these initial proposals, several protein landscapes have been characterized with different experimental and computational techniques.

The free energy landscape of a protein, which has many atoms, is multidimensional and is influenced by temperature, pressure and solvent conditions. A change in these conditions can influence the relative population of the states and the kinetics between them. As mentioned before, the fluctuations observed in proteins near and far from equilibrium govern their biological functions. For this reason, it has become important to focus on the study of protein motion rather than only on their static structures. Protein dynamics includes any time-dependent change in atomic coordinates of the system. The dynamical character of a protein rooted in the free energy landscape can be characterized in terms of the different timescales (kinetics) and the amplitude and direction (structural changes) of the fluctuations. The timescales of the fluctuations can be divided in two categories: slow timescales and fast timescales. In the slow timescales, fluctuations between different states on the free energy landscape are separated by large free energy barriers of several $kT$ (with $k$ the Boltzmann constant and $T$ the temperature), corresponding to timescales of microsecond, milliseconds and even seconds. Usually these slow timescales are associated to large amplitude collective motions between few stable states. In the fast timescales, the fluctuations are defined within the free energy well where a large ensemble of structurally similar states is separated by small free energy barriers, in the order of less than 1 kT. These fluctuations are typically associated to small amplitude conformational changes on the timescale of picoseconds or nanoseconds [33].

Fig. 1.5a shows an example of a free energy landscape that defines the amplitude and timescale of the protein motions based on the description of Frauenfelder et al. [34]. The states are defined as the minima in the free energy landscape and the transition states are the maxima. The free energy barriers between the states determines the rate of the transitions. A change in the conditions of the system will alter the free energy landscape (from the dark blue to the light blue profile in Fig. 1.5 a). The lower tiers describe fast fluctuations between a large number of similar related substates of the wells (tier-1 and tier-2). These fast motions usually
Figure 1.5: Schematic representation of the free energy landscape of proteins and the amplitude and timescale associated to these protein motions: (a) Free energy landscape of a protein showing the hierarchy of the protein dynamics and the timescales associated to the different free energy barriers. (b) Timescales of the process occurring in protein dynamics. The change from a light blue to a dark blue free energy profile represents the effect of changing the conditions of the protein (temperature, pressure, solvent conditions, etc.). Figure adapted from Ref. [33].
include groups of atoms fluctuating on timescales that range from picoseconds to nanoseconds, such as loop motions and side chain rotations, and even lower tiers with faster motions exist, such as femtosecond bond vibrations (see Fig. 1.5 b). On the other hand, transitions between tier-0 states are rare due to the low probability of the conformations that allows the transition (transition state). These transitions usually involve larger domain motions that fluctuate on timescales that range from microseconds to seconds (see Fig. 1.5). Dynamics on these timescales has recently received a lot of attention because many biological processes, such as enzyme catalysis, signal transductions, protein folding and protein-protein interactions occur on these timescales.

1-4 Computational modelling of complex transitions

1-4.1 Conformational changes in proteins are rare events

Thanks to advances in experimental studies using different crystallography and spectroscopy techniques, the structural data of the intermediate protein conformations in the pathways of conformational changes can be resolved. Time resolved crystallography and spectroscopy techniques have also provided dynamical information about conformational changes along the pathways. Nevertheless, up to now experimental techniques still remain limited to the characterization and identification of different intermediate conformations along the transition pathways, requiring significant populations and sufficiently long life times to detect them. Moreover, the full understanding of a conformational transitions requires complementary information of the molecular mechanism between the conformations along the pathways.

Faced with these experimental challenges, it is desirable to develop general computational techniques that allow us to explore in atomistic detail the structural changes occurring during molecular transitions in proteins. During the past decades Molecular Dynamics (MD) simulations have grown into powerful tools for the theoretical studies of complex systems that consist of millions of particles and that are interesting for a variety of fields, such as physics, material science, chemistry, and biology. MD simulations have provided insight in the stability and the dynamics of many complex molecular systems. By numerically integrating the equations of motion under the action of all intra and intermolecular forces, MD samples the trajectories of different configurations providing a detailed picture of transition pathways in molecules [35; 36]. For this reason, MD simulations appear ideally suited to study the protein conformational dynamics since the method performs a practical statistical mechanical sampling of the phase space [37].

However, despite the success and the advances in these algorithms and the tremendous increase in computer power, most of the molecular transition processes, such as protein folding, occur on time and length scales inaccessible to standard all-atom MD
simulations. While the conformational transitions in proteins occur too fast for high-resolution experimental methods, they become too slow to occur in fully atomistic MD simulations within a reasonable time of computation. These long timescales are normally associated to large relatively large free energy barriers that separate the stable states, hampering the efficient sampling of pathways with conventional MD. The crossing of such a barrier becomes a rare event for the fundamental dynamical time step of molecular motions (usually femtoseconds). Even with all the computational power available nowadays, the simulations can only access up to a millisecond time scale \cite{38}. Therefore, the study of rare events requires the development and application of special computational methods that allow the crossing of large free energy barriers for the study of complex transitions.

Rare event methods can be employed in a variety of complex processes that goes beyond those of protein folding transitions or conformational transitions in proteins. Examples of rare events include nucleation in first order phase transitions, chemical reactions, transport phenomena in solids and liquids, biomolecular isomerizations or even transitions of comets between different orbits of the solar system. Fig. 1.6 gives a general idea about the time and length scales that MD simulations can access and the regions in which rare event methods can be required. In this thesis we focus on the computational study of conformational transitions in proteins that occur on time scales that go from microseconds to milliseconds.
1-4.2 The development of new rare event algorithms to study conformational changes in proteins

Special rare event methods have been developed over the years and applied to the study of conformational transitions in proteins and to the study of many other transitions in complex systems. A first idea or strategy to overcome the difficulties associated to the wide variety of lengths and timescales consists of coarse graining the description of the systems by eliminating unimportant variables or coordinates and maintaining only the just the relevant degrees of freedom (or collective variables) that describe the transitions. For example, the association of two protein monomers to form a dimer could be described by the number of contacts or by the contact distances between the monomers instead of all the degrees of freedom coming from a full atomistic description. Such simplified models can be simulated efficiently on very long time scales. However the development of such models requires considerable insight in the problem to reduce the description to a few degrees of freedom. In most complex systems, the a priori knowledge of the relevant degrees of freedom to describe transitions becomes very difficult to obtain and therefore quite inefficient.

Another class of rare event methodologies aim to explore the free energy landscapes of the complex systems spanned by few descriptors (collective variables), by locating minima, saddle points, transition states or possible transition pathways between the stable states. Once the transition states on the energy surface or the free energy surface have been located with these methods, the kinetics can be studied with the fundamentals of transition state theory (this topic will be explained more in detail in chapter 2).

Methods like metadynamics, umbrella sampling, parallel tempering, replica exchange molecular dynamics (REMD), etc. belong to the class of algorithms that aim to reconstruct the free energy landscape by accelerating the dynamics of the system using an artificial bias potential or making use of higher temperatures, to enhance the sampling of regions on the free energy that otherwise would not be visited within the time scale of MD [40; 41; 42; 36]. On the other hand, methods like nudged elastic band method (NEB), action-based methods, the string method (SM), transition path sampling (TPS) explore the free energy landscape by determining pathways between two a priori known stable states by, for instance, minimizing the potential of energy (or free energy) to obtain stationary points along the path, minimizing the action to obtain likely mechanisms, or by sampling the ensemble of dynamical trajectories connecting the stable states by making use of Monte Carlo methods (MC) in the space of trajectories [43; 44; 45; 46; 47; 48; 49]. Some of these methods employed along this thesis are covered in detail in chapter 2.

Despite the success of these recent developments in rare event methodologies, most of these approaches still suffer from disadvantages coming either from the number of degrees of freedom that they can handle before becoming computationally inefficient, or the strong dependency in the initial conditions selected (such as the definition of the stable states or the initial pathways connecting two minima) or the difficulty
to analyse and visualize the ensemble of dynamical trajectories to extract relevant variables that describe a complex transition and that unfortunately, often remain hidden to the eye.

The free energy landscapes of proteins are rough, with many saddle points, ensembles of transition states and large free energy barriers separating the minima, normally described by high dimensional spaces from where their complexity arises. It appears important to develop new rare event methods and algorithms that allow the identification of transition mechanisms and reaction rates to take an step further in answering the question on how proteins are able to generate unique folding states and perform their functions.

1-5 Outline of this thesis

The aim of this thesis is to address the development and the application of a new rare event method, path-metadynamics, to unravel the transition mechanisms of complex systems. Particularly, we are interested in the study of folding and unfolding events that allow us to explore the free energy landscape of a protein by finding transition pathways and estimating free energy barriers between stable states in high-dimensional spaces of relevant collective variables. In Chapter 2 we give an overview of the theoretical background of molecular dynamics (MD), transition path theory (TPT) and rare events methods that provided the fundaments and key concepts for the development of the path-metadynamics method. This thesis is divided in two main parts: (1) Chapters 3 and 4 address the development of path-metadynamics, which consists of a path finding method for the exploration of high dimensional free energy landscapes. (2) Chapters 5 and 6 show the application of path-metadynamics in studying conformational transitions in proteins. More in detail, Chapter 3 presents the theoretical and computational fundaments behind the method to find the average transition pathways between two stable states in a high dimensional space. Next, in Chapter 4, we use the well known alanine dipeptide model system for conformational transitions to review and extend the theoretical concepts of path-collective variables, average transition paths and the minimum free energy paths and to assess the performance of path-metadynamics. In Chapter 5 we apply path-metadynamics to study the millisecond partial unfolding of a biological relevant signalling protein, the Photoactive yellow protein. We propose an strategy to locate relevant collective variables and estimate accurate folding barriers using path-metadynamics and subsequent analysis of free energy profiles. We complete the thesis in Chapter 6 by applying the path-metadynamics method in the prediction of the unfolding-dissociation mechanism of one of the most studied coiled coils in globular proteins, the leucine zipper protein domain of the yeast transcription factor GCN4. This coiled coil has motivated many studies working into the fundamental relation between the amino acid sequence and protein folding. We end with a summary and outlook of the thesis.
1-6 References


CHAPTER 1. INTRODUCTION


