TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2016

105

Rafael Tapia Rojo

Modeling biomolecules: interactions, forces and free energies

Departamento

Física de la Materia Condensada

Director/es

Falo Forniés, Fernando

Mazo Torres, Juan José

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

MODELING BIOMOLECULES: INTERACTIONS, FORCES AND FREE ENERGIES

Autor

Rafael Tapia Rojo

Director/es

Falo Forniés, Fernando
Mazo Torres, Juan José

UNIVERSIDAD DE ZARAGOZA

Física de la Materia Condensada

2016

Universidad de Zaragoza

Facultad de Ciencias

Departamento de Física de la Materia Condensada

Modeling Biomolecules: Interactions, Forces and Free Energies

Autor:
Rafael Tapia-Rojo

Directores:
Dr. Fernando Falo Forniés
Dr. Juan José Mazo Torres

La rayuela se juega con una piedrita que hay que empujar con la punta del zapato.
Ingredientes: una acera, una piedrita, un zapato, y un bello dibujo con tiza,
preferentemente de colores. En lo alto está el Cielo, abajo está la Tierra,
es muy difícil llegar con la piedrita al Cielo, casi siempre se calcula mal
y la piedra sale del dibujo. Poco a poco, sin embargo,
se va adquiriendo la habilidad necesaria para salvar las diferentes casillas
(rayuela caracol, rayuela rectangular, rayuela de fantasía, poco usada)
y un día se aprende a salir de la Tierra y remontar la piedrita hasta el Cielo,
hasta entrar en el Cielo, (Et tous nos amours, sollozó Emmanuèle boca abajo),
lo malo es que justamente a esa altura, cuando casi nadie ha aprendido
a remontar la piedrita hasta el Cielo, se acaba de golpe la infancia
y se cae en las novelas, en la angustia al divino cohete,
en la especulación de otro Cielo al que también hay que aprender a llegar.
Y porque se ha salido de la infancia (Je n'oublierai pas le temps des cérises,
pataleó Emmanuèle en el suelo) se olvida que para llegar al Cielo se necesitan,
como ingredientes, una piedrita y la punta de un zapato.

JULIO CORTÁZAR, Rayuela

Contents

List of Figures

List of Tables

Agradecimientos

Me gustaría comenzar dedicando unas líneas a todas las personas que han contribuido, de manera directa o indirecta, a la concepción y realización de esta tesis, no sólo en el plano científico, sino también en el personal.

En primer lugar me gustaría mencionar a mis dos directores de tesis, Dr. Fernando Falo Forniés y Dr. Juan José Mazo Torres. Con ellos me inicié en este mundo cuando aún era un estudiante de física, y gracias a ellos he podido sentarme, años después, a escribir estas líneas.

En el marco científico, también aprovecho para agradecer a los distintos colaboradores que han permitido dotar a este trabajo de la interdisciplinariedad que pienso que requiere cualquier investigación en ciencia a día de hoy. En particular son Dra. María F. Fillat y Dra. Mª Luisa Peleato por el trabajo desarrollado en el capítulo 10, y Dra. Anabel Gracia Lostao, Dr. Carlos Gómez-Moreno y Dr. Carlo Marcuello, por proporcionar los experimentos y discusiones que han ayudado a la concepción y realización del trabajo reflejado en la Parte III de esta Tesis.

Agradezco también a mi familia, en particular a mis padres y hermana, por darme el apoyo y motivación durante toda mi vida para poder llegar a este punto. Agradezco también en este párrafo a María, por formar parte de todo durante estos años.

No puedo olvidarme de todos mis compañeros de carrera, en especial de amigos como Rubén, Marcos, Adrián, Laura, Inés o Mari, por haber contribuido a que los años de Licenciatura fuesen una gran experiencia, y haber mantenido la amistad en los años posteriores.

Mención aparte merecen también mis amigos, en especial los "farreros" (María, Ja, Yerai, Pedro, Adrián, Marc, Dani, María, Arturo, Mario) por tantos momentos memorables y por hacer que todo este tiempo pasase de la mejor manera posible, entre IPA e IPA, a pesar de ello implicase una sucesión innumerable de mañanas de sábado y domingo improductivas.

Resumen

Tradicionalmente, la biología ha sido una ciencia cualitativa. En contraste con la física—dedicada principalmente a encontrar las leyes generales que gobiernan nuestro universo—o la química—preocupada por las propiedades fundamentales de átomos y moléculas—, la biología involucra el estudio de los seres vivos, cuya complejidad intrínseca convierte cualquier enfoque cuantitativo en una tarea complicada. A pesar de ello, la vida tiene lugar dentro del marco de las leyes de la física [1]. Asimismo, las interacciones entre átomos y moléculas son básicas en biología, lo cual dota de relevancia a las leyes de la química. De esta manera, la biología debería poder comprenderse en términos de la física y la química y, por tanto, de una manera cuantitativa.

Este planteamiento es relativamente reciente, y está motivado principalmente por la mejora en la comprensión de los sistemas biológicos y por el desarrollo de novedosas técnicas experimentales. Asimismo, nuevas disciplinas científicas, tales como la física no lineal y de sistemas complejos o la química computacional han contribuido a este enfoque cuantitativo de la biología. Dicho cambio en el *status quo* de la investigación en biología han atraído a científicos de diverso origen a trabajar en problemas comunes, lo que ha creado un nuevo campo intrínsecamente interdisciplinar, conocido como biofísica, o física biológica.

En la biología molecular, la física y la química cobran una relevancia particular. Las moléculas biológicas son más complejas que aquellas de las que se suele ocupar la química. Asimismo, realizan sus funciones de manera individual—operando en el límite de las leyes de la termodinámica—, involucradas en una compleja red de interacciones entre cada uno de los átomos de la propia molécula, así como con el medio que las rodea y otras macromoléculas. Esto plantea una exigencia importante, en la cual la estructura molecular, su comportamiento dinámico y sus interacciones con el medio tienen un papel fundamental a la hora de determinar su función.

Los últimos años se ha producido un progreso muy significativo en esta dirección. Gracias al desarrollo de técnicas sofisticadas de biofísica—desde la cristalografía de rayos X o la Resonancia Magnética Nuclear, a la espectroscopía de fuerzas en moléculas individuales—, ha sido posible la determinación de estructuras con resolución atómica, o incluso la manipulación de moléculas de manera individual, lo que ha permitido incluso investigar directamente su comportamiento *in vivo*.

La complejidad de los sistemas biológicos limita la importancia de las predicciones teóricas. La biofísica computacional ocupa este lugar, y se convierte en un método crucial a la hora de comprender procesos biológicos, predecir nuevos comportamientos o ayudar en la interpretación de resultados experimentales. De manera creciente, las herramientas computacionales cobran una particular relevancia en biología, abarcando desde las simulaciones de dinámica molecular o las redes de interacciones en proteínas, al análisis de bases de datos masivas o la propagación de

epidemias.

La biofísica computacional debe hacer frente a dos problemas fundamentales. El primero es el desarrollo de un modelo riguroso para el sistema a estudiar. Esta elección es el punto de partida de cualquier estudio computacional y debería ser lo suficientemente preciso como para reproducir de manera fiable las propiedades que queremos explorar. De lo contrario no sería un modelo predictivo. Dado que los procesos biológicos ocurren en un rango espacial y temporal muy amplio, el primer paso es elegir el nivel adecuado de modelado. Esta elección depende normalmente de la pregunta que queremos responder. Por ejemplo, algunas enzimas catalizan reacciones red-ox mediante la transferencia de electrones entre ellas. En este caso, los efectos cuánticos no son despreciables, lo que hace necesario tener en cuenta sus contribuciones. Sin embargo, las correcciones cuánticas son en general despreciables en la dinámica de procesos a escala molecular como, por ejemplo, cambios conformacionales en biomoléculas, interacción entre biomoléculas o plegado de proteínas.

A pesar de ello, la potencia computacional disponible establece un cuello de botella que limita tanto el número de partículas como el tiempo durante el que vamos a ser capaces de simularlas. Por ejemplo, el estudio del plegamiento de una proteína pequeña (unos 50 aminoácidos) con resolución atómica require la integración numérica de las ecuaciones de movimiento de unas 10^4 partículas. Con un paso de integración del orden de femtosegundos—tiempo característico de vibración atómica—serían necesarios más de 10^9 pasos de integración por partícula para obtener una trayectoria del orden de microsegundos, tiempo característico en el que pliegan las proteínas más rápidas. Por tanto, a día de hoy es imposible simular sistemas de mayor tamaño o procesos que ocurren a escalas temporales superiores

Para solventar esta limitación, la estrategia habitual es la disminución del nivel de detalle de nuestro modelo, algo a lo que la física está bastante habituada. Del gran número de grados de libertad que tiene un sistema molecular, probablemente sólo unos pocos sean relevantes en el problema de estudio. Los modelos de tipo *coarse-grained* (grano grueso) nacen con esta filosofía. Promediando sobre algunos grados de libertad originales, se mantiene un número más reducido de "súper-átomos" o centros de interacción con los cuales describir nuestro sistema. Esto permite acceder a escalas de longitud mayores, así como a tiempos más largos. La principal dificultad de este enfoque es la elección de los grados de libertad relevantes en nuestro problema. Otro problema es el planteamiento de una parametrización adecuada, que deberá escalar de manera natural con el sistema original de nivel atómico. A un nivel más grueso, existen modelos en la mesoescala o incluso en el continuo, que requieren simplemente identificar las escalas temporales y espaciales características del proceso de interés. Idealmente, debería ser posible progresar de manera continua entre estos niveles de modelado, pasando desde los sistemas cuánticos a la macroescala. No obstante, la posibilidad de realizar transiciones suaves entre cada uno de estos saltos de modelo presenta una gran dificultad (ver Figura 1).

El segundo problema a considerar en la realización de simulaciones computacionales es la transformación de los datos obtenidos en información relevante acerca del sistema. En principio, las simulaciones proporcionan gran cantidad de datos en bruto—por ejemplo una trayectoria larga de cada uno de los grados de libertad considerados—de los cuales no es posible obtener directamente información comprensible acerca del problema que tratamos. Conforme aumenta la potencia computacional, este problema es más relevante. Cada vez más investigadores se

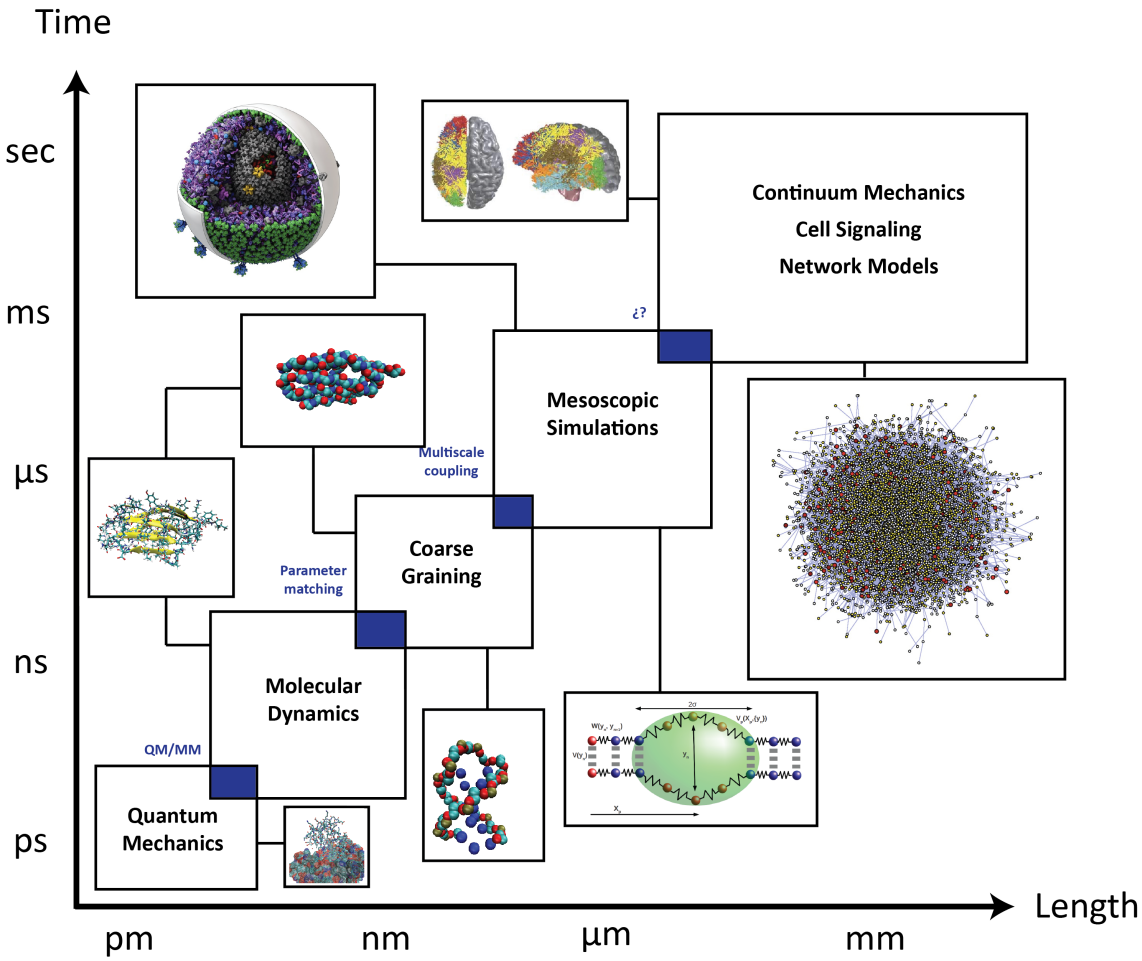


Figure 1: **Diferentes escalas de modelado en biofísica**: La elección de la escala de modelado más apropiada depende de las escalas temporales y espaciales en las cuales ocurren los procesos a modelar. Estos procesos pueden abarcar hasta cinco órdenes de magnitud.

dedican al desarrollo de métodos que permitan la configuración de una descripción físicamente significativa sobre el sistema que estudiamos.

En este contexto, el propósito principal de la presente Tesis Doctoral es el estudio de tres problemas biológicos diferentes, poniendo énfasis en la metodología y el modelado. Tal y como manifiesta el título, las interacciones, las fuerzas y la energía libre son los ingredientes comunes en este trabajo. Primero, las interacciones moleculares no sólo contribuyen a la estabilidad de la estructura molecular, sino también en cada proceso que ocurre en el interior de la célula. De la misma manera, la aplicación de fuerzas a sistemas biológicos ha ganado recientemente una gran relevancia en el campo de la biofísica. Problemas como la adhesión celular, el movimiento molecular o la elasticidad muscular muestran la importancia intrínseca de las fuerzas a nivel molecular. Asimismo, las técnicas de molécula individual han acentuado la importancia de emplear fuerzas mecánicas para inferir la estructura molecular, sus propiedades dinámicas, o incluso su función.

Como último ingrediente, empleamos la energía libre como lengua común en la que explicar los procesos moleculares. La energía libre permite describir la estabilidad de un sistema en particular, ya que proporciona la probabilidad de que un determinado estado se pueble o de que un proceso ocurra. Asimismo, puede describir procesos cinéticos, ya que las transiciones entre estados se pueden comprender como

saltos sobre barreras de energía libre. De esta manera, el conocido como paisaje de energía libre de un sistema molecular proporciona toda la información significativa acerca de un determinado sistema.

Esta Tesis Doctoral está estructurada en cuatro partes, incluyendo las conclusiones finales. La primera, segunda y tercera corresponden en un sentido amplio a tres temas diferentes tratados a lo largo de su desarrollo.

La **Parte I** se centra en el análisis de dinámica de proteínas, enfatizando las distintas descripciones que pueden usarse para comprender su paisaje de energía libre. Comenzamos con el Capítulo 1, que sirve de breve introducción en el problema del plegado de proteína, y de cómo la dinámica molecular es una herramienta potente para explorarlo. El Capítulo 2 presenta los métodos que utilizaremos a lo largo de la Parte I, así como en las siguientes. En particular, presentamos dos metodologías diferentes para describir el paisaje de energía libre de un sistema molecular. El primero plantea su representación por medio de un número pequeño de parámetros de orden. El segundo lo transforma en una red cinética, donde las cuencas de energía libre están definidas como estados con una cierta probabilidad de transición entre ellas. El principal objetivo de esta Parte I es la comparación de ambos enfoques para comprender un sistema molecular.

Para nuestro estudio, escogemos un modelo de proteína *coarse-grained*. Éste es un sistema de relevancia biológica, con un comportamiento dinámico complejo, y por su tamaño es fácil de tratar. El Capítulo 3 revisa los modelos *coarse-grained* de proteína, centrándose en el que usaremos en particular. Finalmente, los Capítulos 4 y 5 muestran los resultados de nuestro estudio, donde la proteína modelo se somete a una fuerza mecánica para forzar la transición de desnaturalización. El Capítulo 4 se centra en diferentes descripciones de baja dimensión, con particular énfasis en la importancia de escoger coordenadas de reacción adecuadas. Asimismo emplearemos métodos de reducción dimensional como herramientas para encontrar parámetros de orden adecuados. En el Capítulo 5 proporcionaremos la descripción del sistema en forma de red cinética, caracterizando su espacio conformacional de manera adecuada, así como revelando sus rutas de desnaturalización.

La **Parte II** muestra el estudio de un modelo de DNA al nivel del par de bases. El Capítulo 6 comienza con una breve revisión sobre la molécula de DNA desde la perspectiva biofísica. El Capítulo 7 es una introducción al modelo mesoscópico que emplearemos, el modelo de DNA de Peyrard-Bishop-Dauxois. El Capítulo 8 muestra el primer trabajo en este tema, donde modificamos el modelo original para incluir la interacción con el solvente. En elCapítulo 9 proponemos un modelo de interacción proteína-DNA donde el modelo de Peyrard-Bishop-Dauxois sirve de modelo para la molécula de DNA, y se incluye una partícula que interacciona acoplada a las regiones abiertas en el DNA. En el Capítulo 10 profundizamos en este modelo, analizando promotores de un organismo particular. Nuestro objetivo es la localización de posibles sitios de unión proteína-DNA aplicando un método de análisis detallado sobre simulaciones del modelo.

La **parte III** está dedicada a los experimentos de molécula individual. Presentamos una colaboración experimental, analizando experimentos de molécula individual para la disociación mecánica de complejos biológicos. Nuestro objetivo es proporcionar una visión adecuada del paisaje de energía libre que gobierna el proceso. El Capítulo 11 sirve de breve introducción a este tipo de técnicas, centrándonos en la relevancia de aplicar fuerzas directamente a moléculas individuales. En el Capítulo

12 presentamos algunos métodos significativos que han sido propuestos para obtener energías libres al analizar la respuesta a la fuerza de un determinado sistema. Éstos pueden ser clasificados en dos grupos diferenciados. Unos proponen la obtención de barreras de energía libre, mientras que los otros la obtención de magnitudes de equilibrio a partir del análisis de procesos fuera del equilibrio.

El Capítulo 13 muestra el análisis de experimentos por medio de un protocolo cuidadoso que proponemos para comprender la disociación mecánica de complejos biológicos. Este capítulo incluye la descripción del sistema biológico que estudiamos, el procedimiento que seguimos para analizar los experimentos, así como los resultados obtenidos. Sorprendentemente, nuestros resultados plantean una discrepancia que nos motiva a proponer un nuevo perfil de energía libre para describir este proceso. Esta tarea se lleva a cabo en el Capítulo 14, donde planteamos un modelo físico para este tipo de experimentos, basado en un nuevo perfil de disociación. Realizamos simulaciones en este modelo para reproducir las trayectorias experimentales. Si se aplica el mismo protocolo de análisis, somos capaces de recuperar las características del perfil original. Esto valida nuestro método, y nos permite llegar a conclusiones significativas a cerca del proceso de disociación mecánica de compuestos biológicos.

Summary

Traditionally, biology has been a qualitative science. In contrast to physics—primary devoted to finding general rules that define our universe—or chemistry—concerned with the fundamental properties of atoms and molecules—, biology involves the study of living systems, with an intrinsic complexity which makes any quantitative approach a complicated task. However, life takes place within the confines of the laws of physics [1]. Also, interactions between atoms and molecules are basic processes in biology, giving relevance to chemical laws. It follows then that biology should be understood in terms of physics and chemistry, and so, described in a quantitative way.

This approach is very recent, and it is motivated mainly by advances in our understanding of biological systems and the development of novel experimental techniques. Also, the advent of new scientific disciplines, such as nonlinear and complex system physics or computational chemistry, has contributed to the development of quantitative biology. This recent breakthrough has attracted scientists from different backgrounds to work on common biological problems, giving rise to an intrinsically interdisciplinary field, which is generically known as biophysics or biological physics.

For molecular biology, the importance of physics and chemistry is particularly emphasized. Biological molecules are considerably more complex than those species usually studied in chemistry. Also, they operate at an individual level—in the border of thermodynamic laws—, involved in a complex interplay between each atom within the molecule itself, but also with the environment or other macromolecules. This presents a serious challenge were the structure of the molecules, their dynamical motions, and the complex networks of interactions play central roles in determining their function.

Recent years have witnessed a significant progress in this direction. Boosted by the development of sophisticated biophysical techniques—from X-ray crystallography or Nuclear Magnetic Resonance to single-molecule force spectroscopy—, it has been possible to determine molecular structure at an atomic resolution, or even to probe molecules at an individual level, possibly monitoring directly their *in vivo* function.

The complexity of biological systems gives a rather limited range of action to theoretical predictions. Computational biophysics occupy this role, becoming a crucial method to understand biological process, to predict new phenomenology or to help in the interpretation of experimental data. Computational tools have an increasing role for the study of an large number of problems, spanning from the direct simulation of molecular dynamics, to networks of interacting proteins, analysis of large data bases or epidemic spreading.

Computational biophysics face typically two different problems. The first one is the development of an accurate model for the system we want to simulate. This is

the input of our simulations, and must be able to reproduce faithfully its behavior in order to provide a reliable source of information. Biological processes span over a huge range of length and time scales. The first step is the choice of the appropriate level of modeling, which usually hinges upon the particular question we want to answer. For example, some enzymes catalyze redox reactions which involve the transfer of electrons. Quantum effects are non-negligible here, so a proper description of such process should account for these contributions. Nevertheless, quantum corrections are usually meaningless when describing biological macromolecules, like for example in conformational changes, ligand:receptor interaction or protein folding.

Nevertheless, the available computational power sets up a natural bottleneck which limits how many particles and for how long can we simulate them. For example, if we want to learn how does an small protein of ~ 50 amino acids reaches its folded conformation with atomic detail, we must integrate numerically the equations of motion of $\sim 10^4$ particles. With a time step in the order of fs—characteristic time of atomic vibration—we should integrate over 10^9 times the equations of motion per atom to achieve a trajectory of µs, characteristic time to observe a single folding event for fast folding proteins. If we wish to study a larger system such the interaction of some drug with a lipid bilayer, or processes which occur on longer time scales, an atomic resolution is hopeless.

The strategy upon this limitation is to decrease the level of detail of our model, something to which physics is quite used to. Out of the thousands of degrees of freedom involved in a molecular system, theres is likely a few significant ones for the particular question we want to explore. Coarse-grained models born with this philosophy, as they integrate out some degrees of freedom to leave out a small number of "super-atoms" or interaction centers, with which the system is described. This reduction allows us to access to larger length or longer time scales. Obviously, the main difficulty is the choice of the relevant degrees of freedom of our problem and the selection of an appropriate parametrization, which should scale properly with the lower scale model. Above coarse-graining, we can choose models in the mesoscale or even in the continuum simply by identifying which are the length and time scales in which the properties of interest manifest. Ideally, it should be possible to move continuously through the different modeling levels, from the very tiny atomic quantum systems, to the macroscale, although a smooth transition between each jump presents an additional challenging issue (see Fig. 2).

A second concern when dealing with computer simulation is to transform the data into valuable knowledge about our system. In principle, our simulation renders a large amount or "raw" data—like for example a very long time series of all the involved degrees of freedom—from which it is not straightforward to meet a meaningful answer for the question we are asking. As the computational power increases, this issue becomes more and more relevant, and more researchers are devoted to the development of methods which can yield a physically relevant picture of the system subject to study.

In this context, the principal purpose of this Thesis is to tackle three different biological problems, emphasizing the modeling and the methodical steps. As the title reads, interactions, forces and free energies are the common ingredients of this work. First, molecular interactions contribute, not just to the stability of the structure of molecules, but also to every function inside the cell. Forces -meaning external forces- have gained recently a lot of popularity in the field of biophysics. It

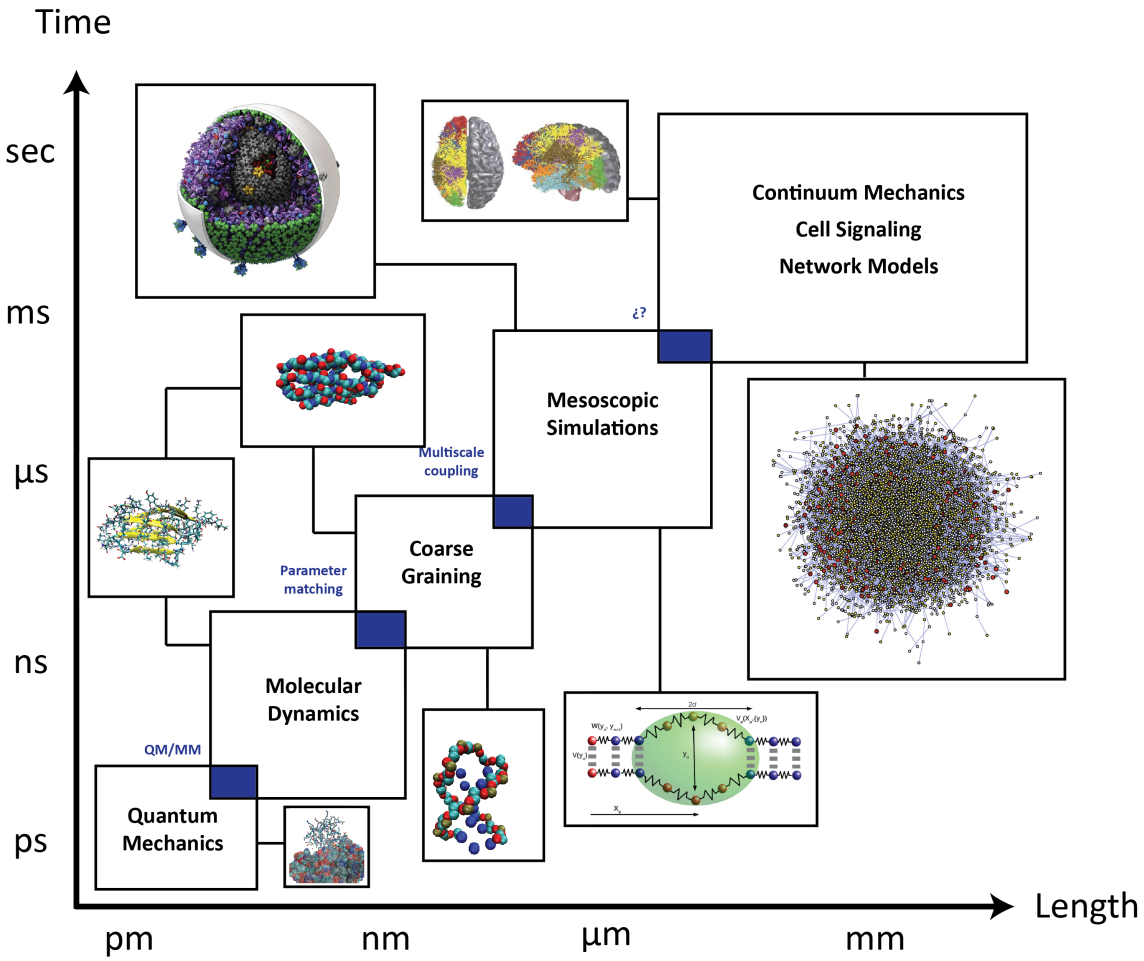


Figure 2: **Different modeling scales in biophysics**: The choice of the most appropriate modeling scales hinges upon the length and time scales at which the desired processes occur. These processes span over five orders of magnitude.

is known that forces are ubiquitous in biology, participating in processes as diverse as cellular adhesion, molecular motion or muscle elasticity. Nevertheless, the birth of single-molecule techniques have stressed the importance of force as a probe to inspect molecular structure, dynamics and function. As a last ingredient, free energy is a common language for explaining molecular processes. Free energy differences allow to describe the stability of a particular system, as it gives the probability of a determinate state or process to occur. Also, it can be used to describe kinetic processes by understanding transitions between states as hopping events over free energy barriers. In this sense, the so called free energy landscape of a molecular system provides all relevant information about the system and thus constitutes the ultimate information we want to acquire.

This Thesis is structured in four parts, including the present introduction and the final conclusions. The first, second and third parts correspond broadly to the three different topics which have been treated within the last years.

Part I is devoted to the analysis of protein dynamics, focusing on different descriptions that can be used to understand their free energy landscape. We start with Chapter 1, which serves as a brief introduction into the problem of protein folding, and how molecular simulations are a useful tool for understanding it. Chapter 2 presents the methods to be used through Part I and other parts of the Thesis. In

particular, we present two different approaches for understanding the free energy landscape of a molecular system. One focusses on its representation through a small number of order parameters. The other transforms it into a kinetic network, where the free energy basins are defined as states with a certain probability of transition between them. The main objective of Part I is to compare the performance of both approaches on a particular molecular system.

We choose a non-native coarse-grained protein model for such purpose. This system has biological relevance, with a complex dynamic behavior, but it is easy to simulate due to its small size. Chapter 3 focuses on coarse-grained protein models, explaining in particular the one we employ. Finally, Chapters 4 and 5 show the results of our study, where the model protein is subject to a mechanical force to enhance the unfolding transition. We wonder which is the best way to characterize this process, understanding which mechanism does the protein follows to transit from the native to the unfolded structure. Chapter 4 focusses on different low dimensional descriptions, emphasizing the importance of a correct choice of the reaction coordinates. Also, we employ a systematic method for dimension reduction as a useful tool to define meaningful order parameters. Chapter 5 gives the network description of the system, where the conformational space is correctly described, and the unfolding pathways unveiled.

Part II focuses on the study of the DNA model at the base-pair level. Chapter 6 starts as a brief overview of the DNA molecule from the perspective of a biophysicist. Chapter 7 is an introduction to the mesoscopic model we use, the Peyrard-Bishop-Dauxois DNA model. Chapter 8 is the first work on this topic, where the original model is modified in order to include the interaction with the solvent when the double-helix is disrupted and the base-pairs exposed to the solvent. In chapter 9 we propose a protein-DNA interaction model where the DNA molecule is modeled with Peyrard-Bishop-Dauxois model, including an interacting particle which couples to open regions of the DNA. Chapter 10 makes a further study on this model, which we use to analyze promoters from a particular organism in order to locate protein-binding sites by applying a suitable analysis method on our model.

Part III is dedicated to probably the most popular experimental techniques in biophysics, single-molecule experiments. Here we present a work in collaboration with an experimental group, where we analyze single molecule experiments for mechanical unbinding of biological complexes, in order to provide a correct vision of the free energy landscape governing such process. Chapter 11 is a short introduction on this sort of techniques, focusing on the relevance of applying forces to individual molecules. Chapter 12 presents some reliable methods which have been proposed for obtaining free energies by analyzing the force response of a particular system. They can be classified in two different sets, ones devoted to the recovery of free energy barriers, and the other obtaining equilibrium magnitudes by analyzing a non-equilibrium process.

Chapter 13 presents the analysis of the experiments by means of a careful protocol we propose to understand mechanical unbinding of biological complexes. This chapter includes a description of the biological system, the procedure we follow to analyze the experiments and the results we obtain. Surprisingly, our findings lay out a discrepancy which motivates us to propose a new free energy profile to describe this process. This task is undertaken in Chapter 14, as we suggest a model for this sort of experiments, based on the new unbinding profile. We perform simulations on

the model to mimic the experimental trajectories. Applying the same analysis protocol, we can recover the characteristics of the target profile, validating our method and allowing us to arrive to meaningful conclusions about the process of mechanical unbinding of biological complexes.

Part I

Understanding Molecular Simulations: Of Low Dimensional Representations and Markov State Models

It is going to be necessary that everything that happens
in a finite volume of space and time would have to be
analyzable with a finite number of logical operations.
The present theory of physics is not that way, apparently.
It allows space to go down into infinitesimal distances,
wavelengths to get infinitely great, terms to be summed in infinite order,
and so forth; and therefore,
if this proposition [that physics is computer-simulatable] is right,
physical law is wrong.

RICHARD P. FEYNMAN

Chapter 1

Of Proteins, Protein Folding and Free Energy Landscapes: A Brief Overview

This section intends to be a general introduction into the biochemistry of proteins, and the problem of protein folding. We give a quick overview on some tools and theories biophysicists have proposed to deal with this problem, stressing the current importance of molecular simulations. Finally, motivated by single molecule experiments, we introduce briefly mechanical unfolding of proteins.

1.1 Proteins, a Brief Introduction into the Machinery of Life

Proteins are Linear Chains of Amino Acids

Proteins are macromolecules which play a central role in biology. They perform a vast array of functions, from catalysis or molecular recognition, to transport or DNA replication. From the biochemical point of view, they are linear chains of amino acid residues. Nearly every known protein is built from just 20 different amino acids, bound together through covalent bonds. Yet, they present a large variability in size, function and complexity. Proteins arrange tridimensionally into functional structures determined by their amino acid composition. However, they frequently suffer post-transcriptional modification that may alter this final sequence, depending on the environmental conditions. Additionally, some proteins have non-peptidic groups attached to them -the so called cofactors, such as metal ions, NADP or some vitamins. Even more, they usually form stable associations -protein complexes- in order to perform a particular function.

In a deeper sight, amino acids are the building blocks of proteins. They share common structural features. Particularly, the 20 standard amino acids which form proteins are α-amino acids. They are organic compounds which have a carbon atom (α-carbon) with a carboxylic acid (-COOH) and amine group (-NH_2) attached to it. Additionally, a side-chain specific to each amino acid determines the particular properties of the molecule, such as size, solubility or electric charge. Amino acids can be classified according to their lateral chains or R-groups into five groups (see Fig. 1.1).

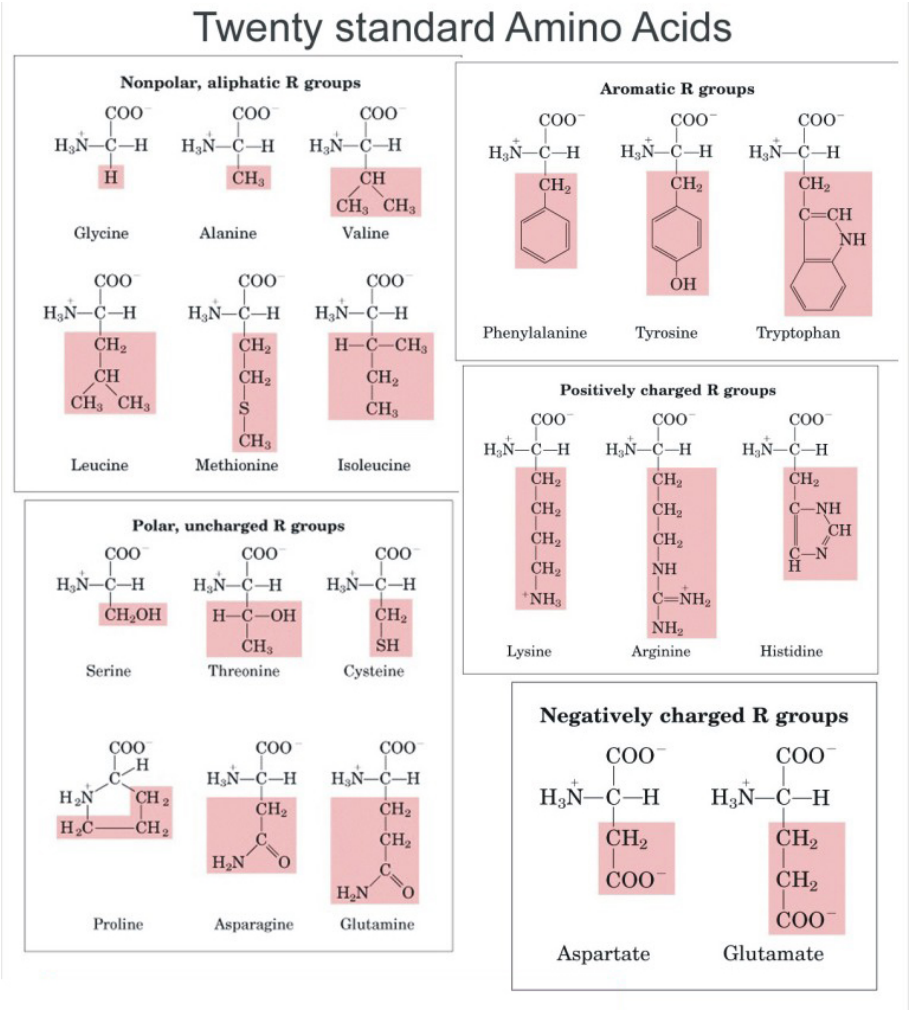


Figure 1.1: **The 20 amino acids found in proteins**: Amino acids are ordered according to their side chain (picture taken from [2]).

1. **Hydrophobic side chain**: These aminoacids have apolar side chains which tend to aggregate due to hydrophobic interactions. They are Alanine (Ala, A), Glycine (Gly, G), Isoleucine (Ile, I), Leucine (Leu, L), Methionine (Met, M) and Valine (Val, V).

2. **Aromatic side chain**: They have aromatic groups in their lateral chains and are also hydrophobic. They are Phenylalanine (Phe, F), Tryptophan (Trp, W) and Tyrosine (Tyr, Y).

3. **Polar Neutral side chain**: These amino acids do not have an electric charge but they have a net dipolar moment. In this regard they are hydrophilic and highly soluble. They are Proline (Pro, P), Asparagine (Asn, N), Cysteine (Cys, C), Glutamine (Gln, Q), Serine (Ser, S) and Threonine (Thr, T).

4. **Negatively charged side chain (acidic)**: These amino acids have a net negative charge at neutral pH. They are Aspartic acid (Asp, D) and Glutamic acid (Glu, E).

5. **Positively charged side chain (basic)**: These amino acids have a net positive charge at neutral pH. They are Arginine (Arg, R), Histidine (His, H)

and Lysine (Lys, K).

The Peptide Bond

The covalent bond between two amino acids is known as the peptide bond. The formation of this bond is a condensation processes between the amino group of one residue and the carboxyl group of another, yielding a water molecule and the dipeptide molecule, see Fig. 1.2 (A). The peptide bond forms a rather restrained molecule, where six atoms are constrained to lie in a plane, as nitrogen and carbon atoms in the NH-CO unit are sp^2 hybridized. Additionally, there is a certain freedom of rotation about the $C_\alpha - CO$, $C_\alpha - NH$, hindered only by the steric interaction between the nonbonded atoms.

These angles are usually labeled as ψ and ϕ respectively, see Fig. 1.2 (B). The angle ω for the CO-NH bond has an almost fixed value of $\omega = 180^o$ (planarity of the bond). This is the *trans* configuration. The *cis* configuration $\omega = 0^o$ is energetically unfavorable because of the steric clash between the side chains of the amino acids. For a *trans* peptide bond the distance $C_\alpha - C_\alpha$ is of $\approx 3.8\,\text{Å}$.

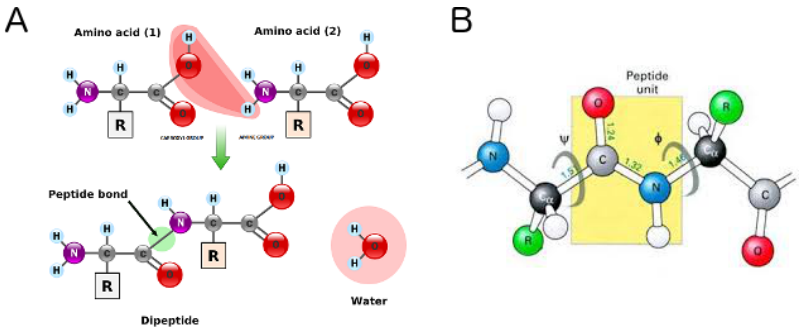


Figure 1.2: **Peptide bond formation and geometry of the peptide bond**: Panel (A) shows the condensation relation that can form the peptide bond between two amino acids. Panel (B) shows the geometry of the peptide bond, with atoms α-C, β-C and N laying on the same plane. Dihedral angles ψ and ϕ are defined. (Picture modified from [2])

Disulfide bonds are an additional kind of covalent bonds which appear in proteins, forming between oxidized sulfur atoms of cysteine residues. Disulfide bridges have crucial effects on the flexibility of proteins and on the stabilization of quaternary structures.

A linear chain made up of several amino acids bonded by peptide bonds is a polypeptidic chain. Amino acids belonging to a polypeptidic chain are usually referred to as *residues*.

Elements of Protein Structure

Proteins arrange tridimensionally in a particular structure, which determines its function. The structure of a protein is described in four different levels, the primary, secondary, tertiary and quaternary structure:

1. **Primary structure**: The primary structure of a protein is the linear sequence of amino acids in the polypeptide chain. Due to the asymmetry of the peptide bond, the polypeptide chain is directional. According to the free amine and carboxylic acid groups, the two ends are referred to as the N-terminus and

C-terminus, respectively. The unfolded chain has a net dipolar moment, as the NH and CO groups act as hydrogen-bond donors and acceptors.

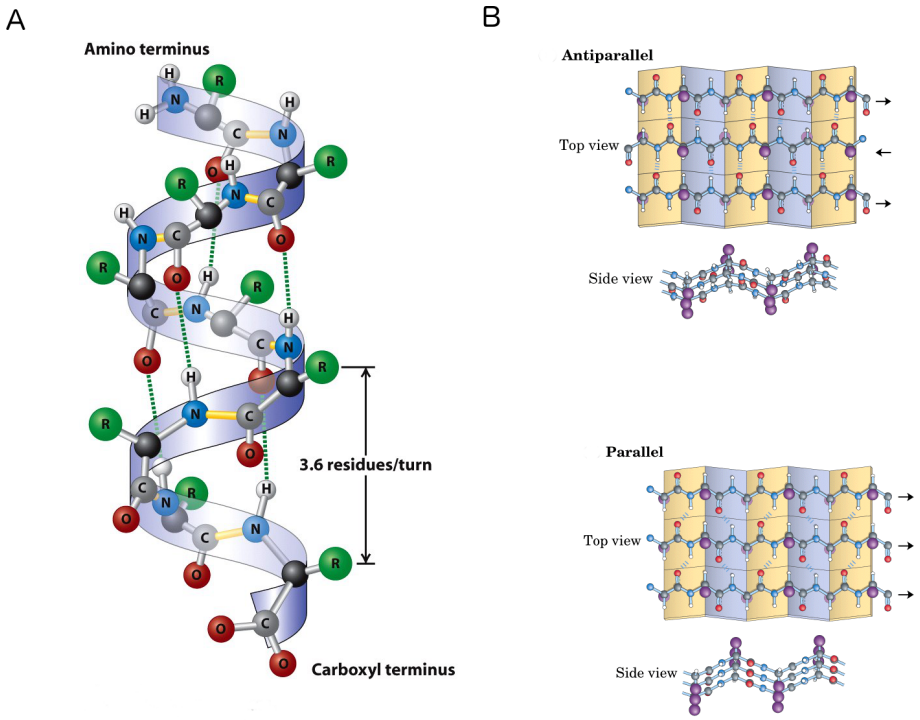


Figure 1.3: **Elements of protein secondary structure, α helix and β sheets**: Panel (A) shows the basic structure of the α helix, with 3.6 residues per turn. Panel (B) shows the structure of β sheets on their two possible arrangements, parallel and anti parallel. (Modified from [2]).

2. **Secondary structure**: The secondary structure is the local stable substructures which are recurrent in proteins. There are two main elements of secondary structure, namely the α-helix and the β-sheets.

   (a) **α-helix**: The α-helix was predicted in 1951 by L. Pauling based upon geometric considerations and the analysis of peptide bonds in small molecules, being subsequently supported by X-ray diffraction patterns [3]. This structure is formed by linking of NH hydrogen-bond donors and CO hydrogen-bond acceptors, separated by regular elements in the amino acid sequence, typically four residues ahead.

   Most of the helices occurring in proteins are right-handed α-helices, with 3.6 residues per turn. This correspond sto a pitch of 5.4 Å or 1.5 Å per residue (see Fig. 1.3 (A)). α-helices have backbone dihedral angles around $(\phi, \psi) = (-60°, -45°)$. In some occasions, other helices might form. Left handed helices are quite rare, while the 3_{10} helix (H-bonds every three residues) and π-helix (H-bond every 5 residues) are occasionally found.

   (b) **β-sheet**: β-sheets are the other common element of secondary structure, also based upon hydrogen bonding between the NH and CO groups. Nevertheless, the H-bonds in these structures are less local, being the donor and acceptor groups chemically separated by large distances along the polypeptidic chain. β-sheets consist of β-strands—locally extended

chains—connected laterally by H-bonds. The dihedral bonds are around $(\phi, \psi) = (-140^o, 135^o)$

There are two distinct ways in which β-strands can align themselves, the parallel and antiparallel β-sheets. In an antiparallel β-sheet the successive β-strands alternate directions so that the N-terminus of one strand is adjacent to the C-terminus of the other. This arrangement has high inter-strand stability, as the NH group of one peptide unit form a H-bond with the CO of the other in a planar way, forming rings of ten atoms (see Fig. 1.3 (B)). The parallel β-sheet have all β-strands with the N-terminus oriented in the same direction. The H-bonding between strands induces a slight non-planarity in the pattern, reducing the stability. The dihedral bonds are about $(\phi, \psi) = (-120^o, 115^o)$.

Most of the proteins can be described as series of α-helices and β-sheets connected by loop regions, which can have different sizes and shapes. These loops are often partially stabilized by polar interactions between residues. The quite restricted variety of observed configurations in proteins leads to a rather small range of allowed values for the dihedral angles (ϕ, ψ). A useful representation is the so called Ramachandran plot (see Fig. 1.4), where the dihedral angles are plotted against each other, finding a set of regions which represent the "allowed" conformations. Additionally, Ramachandran plots are a useful and intuitive way to understand the overall structure of a certain protein, given that the secondary structure elements have well defined values on it.

At a structurally higher level, it is often recurrent to find certain combinations of elements of secondary structure, which are called *motifs* or supersecondary structure. Examples are helix-turn-helix motifs, or β-β hairpins, with two antiparallel strands separated by a loop. β-α-β, Zinc fingers or *Greek keys* are more complex motifs, often associated with particular functions.

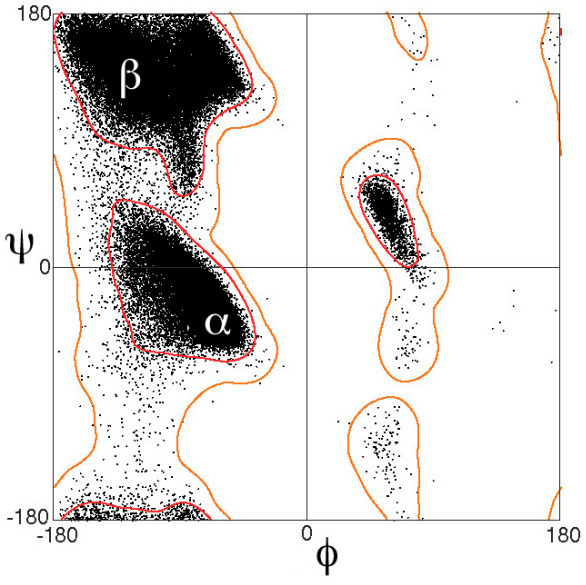


Figure 1.4: **General Ramachandran plot for proteins**: Dihedral angles plotted for 100000 general amino acids types. (Lovell et al. 2003 Proteins 50:437)

3. **Tertiary structure**: The next level is the tertiary structure, the three-dimensional arrangement of the whole protein. The tertiary structure is defined by the spatial coordinates from each of the atoms integrating the protein. Some proteins have more or less unstructured elements, which are hard to label or have dynamical structures. The extreme case is that of *intrinsically disordered proteins*, which lack of any fixed or ordered three-dimensional structure.

4. **Quaternary structure**: The highest level of protein structure is the quaternary, which refers to the arrangement of multiple polypeptidic units onto a multi-subunit complex.

Determining Protein Structure

The determination of protein structure is one of the cornerstones of structural biology, and of many related disciplines. Unveiling the position of every atom of a protein in the native state helps in the further comprehension of its function, always intimately ligated to the particular arrangement of the residues along the 3D structure. Furthermore, a coordinate file is the seed for any molecular dynamics simulation. Thus, protein structure determination is of capital importance for computational biophysics.

Currently, there are several methods for predicting the structure of proteins and most of them imply the combination of several experimental, statistical and computational techniques. Nevertheless, the two principal ones are X-ray diffraction and Nuclear Magnetic Resonance (NMR), although single molecule fluorescence combined with high-resolution microscopy are gaining importance in the last few years.

The concept of X-ray diffraction for protein structure determination is its use in solid state physics, although some technicalities hinder the process. The first problem is to produce the crystal of the subject protein, which is a craft task. Highly hydrophobic proteins, such as membrane proteins, raise particularly hard challenges, given that water cannot be used as a solvent. The second difficulty comes with the interpretation of the diffraction patterns, far more complex than in solid state physics, given the number of atoms present the unit cell. Fitting the diffraction patterns to biologically meaningful structures requires often important computational efforts.

NMR is an additional tool for determining protein structure, which has gained a lot of popularity recently, specially with small proteins. NMR allows resolving structures in solution, while X-ray diffraction produces a frozen picture of the structure, rather than the likely actual ensemble of native configurations. Study of resonance from particular protons renders a high resolution map which can be deconvoluted to yield the structure. This resolution decreases with large proteins due to the increase of the coupling distance.

From a practical point of view, the most useful resource regarding protein structure is the Protein Data Bank (PBD, `http://www.rcsb.org/`), which is used as a repository for resolved structures, becoming a crucial tool for scientists working in the field.

1.2 Proteins Fold: from Sequence to Structure and from Structure to Function

Proteins perform a large variety of biological functions. Once the polypeptidic chain has been synthesized in the ribosomes, the protein must fold or assemble into its biologically functional state, the *native state*. The particular tridimensional arrangement of the residues into the functional three-dimensional structure determines the function of the protein. Departures from the native structure, *misfolding* can imply severe problems.

The problem of how does a protein find its unique native structure is a long debated one. It started with the set of experiments Anfinsen and coworkers performed in the 60s [4]. They ansewered partially this question proving that, at least for small globular proteins, the native structure is determined only by the amino acid sequence. In other words, the primary structure determines the secondary and tertiary ones. This postulate of molecular biology implies also that the native state is unique, stable and has kinetic accessibility.

This picture has gained complexity, specially in larger proteins or some particular cases. Some proteins need assistance of other ones, called chaperones, to fold properly, and they fail to reach this structure in *in vitro* experiments. Furthermore, proteins such as prions are an exception of Anfinsen's dogma, as they remain on stable conformations which differ from the native folding state. This *misfolded* structures are cause of diseases such as the bovine spongiform encephalopathy or Alzheimer disease, due to amyloid aggregation.

Levinthal's Paradox and Folding Funnels

Anfinsen's dogma [4] is associated with another concept pointed out by Levinthal in 1969, Levinthal's paradox [5]. He reflected on how did proteins explored its conformational space to find the unique native structure in biological relevant times. If a protein has r residues, and each of them can adopt n stable conformations, then the system has n^r local energy minima. Considering that the system sampled the minima in the fastest possible way, for example the typical vibrational time scale $10^{-12}s$, then the time needed to explore the whole conformational space is huge compared to experimental folding times. Even for modest values such as $r = 100$ and $n = 2$, one would need $\sim 10^{18}s$ for a proper sampling. In a more detailed way, one can say that the number of local minima on a potential energy surface scales exponentially with the system size [6, 7].

The conclusion of this paradox is that proteins cannot sample randomly their conformational space in order to reach the native conformation. Proteins speed up their folding mechanism by the formation of local interactions which determine folding mechanism of the peptide. These local structures serve as nucleation points like, for example, the formation of stable secondary structure elements which guide the folding pathways. Indeed, this kind of protein folding intermediates or partially structured transition states have been detected experimentally. This folding mechanism has been often referred to as *funnel-like energy landscapes* [8, 9].

The folding funnel hypothesis proposes a particular shape for the energy landscape of a protein. Here, the native state corresponds to the terminus of a collection of convergent folding pathways that reach the target structure by decreasing system-

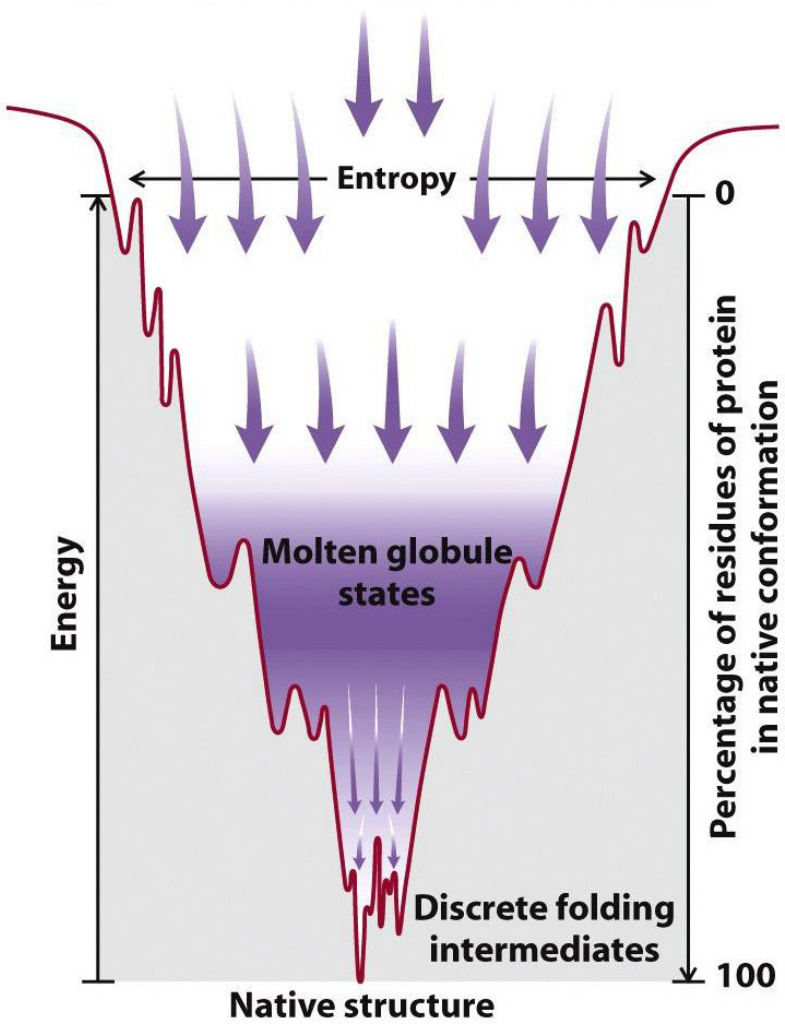


Figure 1.5: **Folding funnel hyphotesis**: The folding process is represented as a search process in a funneled landscape, where the depth represents the energetic stability and the width the entropy. Starting from a random coiled conformation, the system moves to a molten globule state with some secondary structure, to finally find the unique native minimum (taken from [2]).

atically the free energy. Folding funnels (Fig. 1.5) usually represent the energetic stabilization of the structures as the depth of the well and the conformational entropy as its width. The system starts as a random coiled structure, which is just the polypeptidic chain as a random polymer. These states have large conformational entropy but are completely unstructured. The system evolves to the so-called molten globule, which is a collapsed state with some native-like secondary structure content, decreasing the conformational entropy and increasing the energetic stabilization. This region can contain some minima that act as kinetic traps, slowing down the relaxation towards the native state. These partially folded structures have been found in several studies [10]. Finally, the system optimizes the tertiary structure interactions and reaches the unique energy minimum. This folding mechanism is intimately related with the so-called *principle of minimum frustration* [11], which states that evolution has selected proteins in order to avoid mis-folded traps, and non-native contacts, in order to optimize the searching process of the native structure.

Thermodynamics of Protein Folding

Protein structure is highly influenced by the environmental conditions. For example, globular proteins, which exist in solution inside the cell, often contain a hydrophobic core. Nonpolar side chains hide in this inner core, while polar or charged residues locate on the surface, exposed to water molecules. Transmembrane proteins span the lipid bilayer, and have a different structure. Nonpolar residues locate in the region in contact with the lipid hydrophobic tails, while the "head" and "tail" of the proteins, which are in contact with the inner and outer part of the cell, keep hydrophilic residues.

In a simple thermodynamic scenario, protein folding is an equilibrium reaction where the laws of equilibrium thermodynamics govern. This is a reversible transition between two states $F \leftrightarrow U$. The folded protein should be stabilized by the free energy ΔG^0:

$$\Delta G^0 = G^N - G^U = \Delta H - T\Delta S, \tag{1.1}$$

where $\Delta G^0 < 0$. The protein stability is an interplay between the enthalpic and entropic contributions. Enthalpic contributions come mainly through intermolecular bonds—like the H-bonds, salt-bridges or the S-S bonds—and H-bonds with surrounding water molecules. The entropic contribution comes from two main sources, the conformational freedom of the protein structure, and the hydrophobic interaction with the hydration water molecules. This interplay is very subtle and the native structure of a protein *marginally stable*, as ΔG^0 is of few kT at room temperature.

In few words, one can illustrate the balance between entropy and enthalphy in the following way. The unfolded state has a high degree of conformational entropy, being a free flexible polymer which can adopt a high number of conformations, whereas the native state has minimal conformational entropy. The enthalpy is larger in the native state, as intermolecular interactions (contacts) form between residues (for example the hydrogen bonds within a α-helix). The contribution of the bonds with the water molecules in the solvent, usually cancels out, being approximately the same in both states. The key contribution is the entropy of the hydration of water. In the random coiled conformation the hydrophobic side chains are exposed to the solvent, leading to a large penalty in entropy.

In this regard, the thermodynamic force which drives protein folding is the hydrophobic effect. This is the so-called *hydrophobic collapse hypothesis*. Water is a relatively structured liquid, forming a local network of hydrogen bonds in an ice-like fashion. When introducing a hydrophobic solute, water molecules respond by further ordering around it, giving rise to a large negative entropy change. As the hydrophobic solute cannot form hydrogen bonds with the first solvation shell of water molecules, there are fewer favorable orientations for molecules in this shell than in bulk. The burial of nonpolar residues increases the entropy of the solvent, paying the conformational entropy loss of the protein, and leading to an overall stable structure.

1.3 Molecular Dynamics Simulations as a Tool for Understanding Protein Folding

Unveiling the particular mechanism of protein folding is a hard problem to study experimentally. Protein folding is known to be a sensitive process where a single point mutation might alter the whole process. Furthermore, studying protein folding with experimental techniques represents a serious challenge, given the heterogeneous nature of an ensemble of proteins folding in an experiment. This provides a coarse view, which unfortunately, lacks of any atomic resolution.

Therefore, simulating protein folding in a computer appears as a great opportunity to gain insight into such problem. Simulations can shed light onto the precise mechanism of protein folding with atomic detail, suggesting new hypothesis and new experiments or novel interpretations.

1.3.1 Molecular Dynamics: Atomic Simulations and Force Fields

Molecular dynamics imply the numerical integration of the equations of motion of some particular N-body system, providing molecular *trajectories*.

Molecular dynamics have become an standard tool in molecular biophysics, and in the particular problem of protein folding. The intrinsic complexity of the system, involving several coupled degrees of freedom, make most theoretical predictions worthless, and numerical approaches are necessary. Molecular dynamics usually face three problems which are related to the three:

1. **The model**: Prior to simulating, one must define the equations of motion to integrate, and thus the Hamiltonian which governs the system, the physical model. It must appropriately give rise to predictions which agree with the observed phenomenology.

2. **The simulation**: Molecular systems are often made of a huge number of particles, and producing long-enough trajectories is challenging, even with the current computational time and sampling tools.

3. **The analysis**: One must be able to obtain valuable information from the large amount of data current molecular simulations produce.

Most common models for molecular simulations are atomic-detailed, where each atom is taken as a classic interaction center, and the different interactions are parametrized in order to produce reliable molecular dynamics trajectories. Atomic simulations define a“box” where the molecule is located with the solvent, represented as explicit water molecules or with some implicit model. Trajectories are produced by integrating the Newton equations of motion for the N involved atoms.

This is a challenging problem from perspective of the computational power, as many particles are involved, and the time step for the integration is that of atomic vibrations, in order to provide a reliable scale. Brute force strategies, run extensive simulations on supercomputers or dedicated systems, which currently allow to reach ms scale, enough for describing the folding process of small fast-folding proteins [13] (see Fig. 1.6). “Smarter” approaches rely on sampling effectively the huge

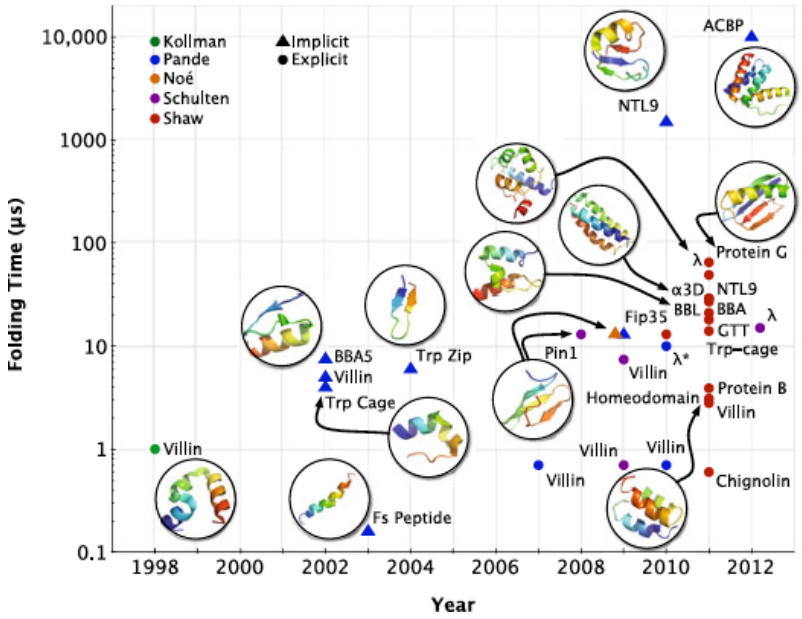


Figure 1.6: **Protein folding mechanism characterized by molecular simulations along the last few years**: The accesible folding times by simulation have increased over the last few years, allowing to simulate the folding mechanism of proteins with folding times up to few μs. Implicit solvent simulations increase greatly the simulation performance. (Picture taken from [12]).

conformational space in order to provide a vision of the equilibrium ensemble without running explicitly the kinetics (see Chapter 2). Additionally, implicit solvent models and coarse-grained ones allow a great reduction of the number of particles, extending the scales (see Chapter 3).

Force Fields and Molecular Dynamics Simulation Packages

Molecular dynamics simulations must render reliable trajectories to reproduce the actual behavior of the studied molecule. The critical point is the definition of appropriate models for the interactions between the atoms in the molecules, the so-called force field problem. Force fields provide potential shapes and parameters for every atom in a system, and should be general and exportable to any system.

Usually, force fields distinguish between two kinds of interaction, the bonded and unbonded terms. Bonded terms involve covalent bonds, which cannot be broken in a classical molecular dynamics interaction. Nonbonded terms describe long-range electrostatic interactions and van der Waals forces. The specific content of each term depends on the particular force field, but usually the total energy is a sum of different contributions which resemble the actual interactions within a molecule and with the solvent.

Bonded interactions are usually a sum of three terms, the bond, a three-body bending angle potential and a four-body dihedral interaction. The nonbonded contribution is the sum of the electrostatic term and the van der Waals interaction. In addition to the particular shape of the previous potentials, force fields must define a set of parameters for every kind of atom, chemical bond, dihedral angle ... These parameter sets are usually determined by empirical arguments, or in some cases from first principles.

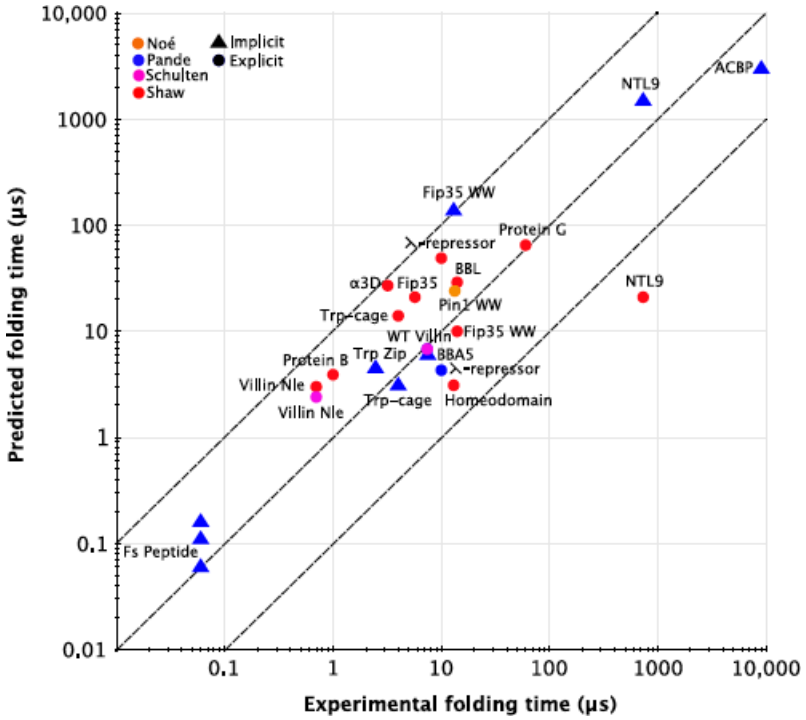


Figure 1.7: **Comparison of simulated and experimentally measured folding times**: Central line means total agreement, while outside lines the range of an order of magnitude agreement. Predictions from molecular dynamics simulations are reasonable, given that experimental folding times depend critically on several conditions, such as temperature, pH or salt content, (taken from [12]).

Water molecules require a special remark, as they form the large majority of the system atoms, having key consequences on the behavior of a biological macromolecule. Although water molecules are formed by three atoms, they can be modeled in different ways in order to account properly for the polarization effects. In this sense, there are models with increasing number of interaction sites (from two sites up to five or six), creating virtual sites to account for charge distributions. Three sites models such as TIP3P are often used for molecular simulations as they meet the compromise between accuracy in the description and complexity. Other models such as TIP4P, OPC or TIP5P can be applyed depending on the needs.

Currently, a large number of force fields have been developed, and they are largely available in the most usual molecular simulation packages, such as `GROMACS` [14], `CHARMM` [15] or `NAMD` [16]. It is now widely accepted that current force field are sufficiently accurate for the quantitative prediction of a large number of biomolecular problems, with some known limitations [17]. Figure 1.7 shows a comparison between experimentally measured times for protein folding and the molecular dynamics prediction. The central line is perfect agreement, while the upper and lower disagreement within an order of magnitude. Up-to-date possibilities agree in a satisfactory way with the experimental phenomenology. It must be taken into account that experimental folding times depend critically on conditions such as temperature, pH or salt content.

Open Challenges in Molecular Simulations

In the particular problem of protein folding, the combination of suitable models, computational power and adequate analysis methods, has made possible to simulate directly protein folding, at least for proteins with folding times between $10 - 100\mu s$. Nevertheless, this has been possible only in the very last few years, opening an exciting field which proposes an important number of challenges to be explored.

1. **Longer Timescales**: While up to date we are able to produce—with special supercomputers such ANTON [13]—molecular trajectories up to few ms, many biological processes occur in longer timescales. Particularly, most proteins fold in times which are three orders of magnitude longer. This leaves obviously a lot of space to explore. One strategy might be continuing developing supercomputers or highly parallelizable devices such as GPUs. Nevertheless, brute force approaches might increase slightly current numbers, but not several orders of magnitude. Effective sampling techniques, or multiscale models are new strategies which rely less in computational power.

2. **Larger systems**: The folding mechanism of some proteins of less than 100 amino acids has been resolved by molecular dynamics. Nevertheless, this is a rather limited size. Studying larger systems demands more computational power or more effective sampling techniques. Mixed models or better coarse-grained ones can also conduct promising advances.

3. **Better analysis techniques**: Extensively long simulations lead an extensively large amount of data, which must be processed in order to extract valuable information. A lot of work is now being devoted on proposing rigorous techniques to produce unbiased information about a large molecular trajectory or set of trajectories. Constructing faithfully the free energy landscape of a molecular system from the kinetic trajectories remains still as an open issue.

Particularly, in the field of protein folding, there are several unanswered questions which wait to be answered. When unveiling the unfolding mechanism of a protein, the first question is if this mechanism is unique, or the system might follow several pathways to reach the unique native state. Describing folding through several pathways increase largely the complexity of the process, but also when folding occurs catalyzed by chaperones. This question is related with that of finding metastable states in the unfolding mechanism. There is a common paradigm in protein folding which is that proteins fold in a two-state manner, by hopping between two energetic basins separated by a barrier. Another way to rephrase this problem is asking ourselves about the involved time scales. Experiments measure with large time windows, compared to atomic resolution. Thus, two-state folding might be an apparent process, and looking at smaller time scales might reveal existence of short-lived metastable states which play a relevant role in defining the folding pathways. Molecular simulations are optimal tools to explore this problem, given the intrinsic atomic resolution, and a large number of analysis methods which allow looking into this feature, as we will explore on Chapter 2.

1.4 Mechanical Unfolding of Proteins: From Single-Molecule Experiments to Steered Molecular Dynamics

The study of proteins or molecules subject to mechanical external forces has recently become a hot topic in protein science [18]. This is due to different reasons. First, the advent of single molecule experiments (see Part III) has represented a revolution in the area of experimental biophysics. It is possible to apply directly a force to an individual molecule, in order to induce some sort of conformational change. These techniques changed the experimental paradigm, moving away from typical biochemical assays, which rely on ensemble average, and to sample directly the distribution of some molecular property.

Inspired by these techniques, steered molecular dynamics forces a system to evolve away from its initial equilibrium condition by fixing some group of atoms and applying a mechanical force to some other group. This is interesting for different reasons. One might wish just to study the behavior of a system when subject to a mechanical force, maybe to compare with predictions from single-molecule techniques. Also, forces help in gaining knowledge from the *unbiased* system. They lower barriers, and the transitions between different states are accelerated. Removing the effect of the force in the dynamics, we achieve an effective sampling of the conformational space of the system, allowing the system to surmount high barriers, which involve slow time scales. This is the inspiration for popular methods such as Umbrella Sampling [19].

Finally, the effect of forces to biomolecules is interesting for pure biological reasons. Forces are ubiquitous in biology, and many systems are subject to mechanical forces on their *in vivo* functions. Perhaps the most popular example is protein titin [20]. Titin is a giant modular protein, that functions as a molecular spring and is responsible for the passive elasticity of muscles (see Fig. 1.8). Titin is particularly resistant to external forces, and it has been proved experimentally to unfold under force and refold back when relaxed [21].

In this sense, the study of molecular unfolding of proteins is an interesting problem, which yields valuable knowledge about the mechanism by which proteins fold or unfold. Nevertheless, the exact connection between the behavior of a protein under a mechanical bias and its folding process in thermal equilibrium conditions, remains still as an open problem. Mechanical (un)folding of proteins is a transition between two states with low conformational entropy, so different from regular folding. Force narrows the energy landscape in the pulling direction, forcing the system to react in this direction. We will deal with this topic in a more extensive way in part III.

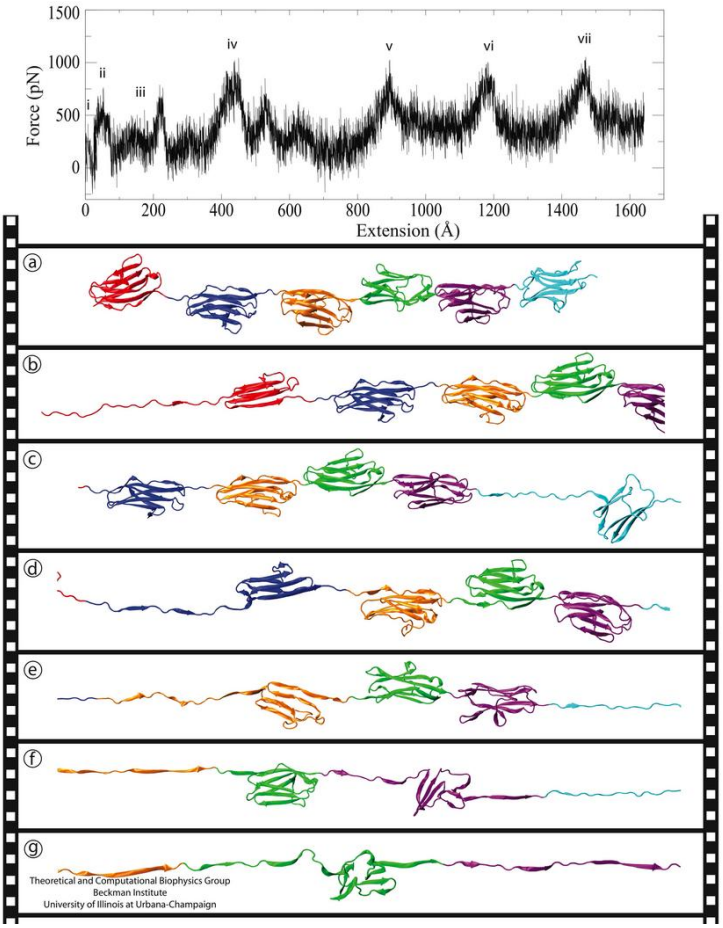


Figure 1.8: **Steered-molecular-dynamics simulation fully extending a six-titin polyprotein, with the individual domains unraveling one by one.**: As in the experiments, each domain unravels individually, showing the classic sawtooth pattern (taken from [20]).

Chapter 2

What Do We Do with all this Data? Free Energy Methods for Understanding Molecular Simulations

This chapter offers a practical review on some analysis methods we employ through next chapters. We present the free energy landscape formalism for understanding molecular simulations and discuss some available approaches to represent it. We start discussing the role of order parameters and reaction coordinates for low-dimensional representations. We address also methods for reducing the dimensionality of molecular systems, namely Principal Component Analysis. Next, we review the topic of Markov state models, discussing their approach for understanding free energy landscapes and offering a practical vision for building them. Finally, we emphasize the particular protocol we will follow.

2.1 Understanding Molecular Simulations

Molecular simulation is a valuable method for studying a large variety of molecular processes, such as protein folding, ligand binding or enzyme-catalysis [22]. Applying simulation to such problems appeals for many reasons. Unlike experiments, simulations provide an atomic-resolution picture, crucial for understanding processes at that scale. Furthermore, such problems are intrinsically complex, involving a large number of coupled degrees of freedom, so theoretical models fail in describing them with generality. This spatial and temporal high resolution allows to compute a large number of observables, and thus to establish a direct connection with experiments.

Nevertheless, there are three main problems molecular simulations face:

1. **Development of accurate models:** Independently on wether we perform a quantum-detailed, all-atom or coarse grained simulation, we must provide a model based on physical assumptions, which will hopefully describe the actual behavior of our system. In the case of atomic simulations, the interactions between atoms are described in terms of empirical potentials which should be correctly parametrized in a general way (force field). In coarse-grained simulations, the modeling step is more critical, as they rely on *ad hoc* physical

assumptions to decrease the complexity of the system (see Chapter 3). In principle, at least for the problem of protein folding, current forcefields are capable of reproducing results in agreement with experiments. Nevertheless, more work remains to be done in the development of more accurate models.

2. **The sampling problem:** Molecular processes span over a large range of time scales. For example, while some conformational changes might occur in few nanoseconds, proteins fold from few μs to s. This requires trajectories of at least few ms in order to provide reliable statistics of such process. In all atom simulations, we integrate the classical mechanics equations with femtoseconds timesteps of (10^{-15}), requiring around $\sim 10^{12}$ timesteps for reaching such scale. Given that a typical system size for a explicit solvent simulation can reaches several thousands of atoms, the computational demand becomes tremendous.

   In the last few years, different strategies have been developed to deal with this limitation. The most straightforward one is brute force. The recent development of effective software and hardware tools has increased dramatically the range of accesible problems. Dedicated supercomputers such as ANTON have allowed to unveil the folding process of several fast folding proteins [13], reaching single trajectories of few milliseconds. The advent of GPUs, which allow a highly effective parallelization, have boosted also the accessible simulation times. In combination with methods such as Markov state models [12] (see Section 2.5), it is now possible to estimate the equilibrium ensemble of a system by simulating several short trajectories and combining them in an adequate way [10].

   Given that human mind goes faster than Moore's law, several efficient sampling techniques have been developed in the last years in order to simulate equilibrium properties without reaching directly the involved timescales. The underlying difficulty of the sampling problem is the existence of large free energy barriers which must be overcome in order to sample effectively the configurational space of the system. This determines a rare event and motivates the development of different techniques to boost simulations or to sample efficiently the conformational space. For example, umbrella sampling [19] or metadynamics [23] allow calculating free energy profiles that cannot be sampled directly. Other methods such as Transition Path Sampling [24] estimate the ensemble of transition paths between two pre-defined states.

3. **Robust data analysis:** Even if we are able to gather enough statistics to sample effectively the equilibrium ensemble of the problem system, we still need to transform all this (likely) huge amount of data into valuable scientific knowledge. Molecular trajectories are extensive high-dimensional time series, and they present a "Big Data Challenge". Providing effective and efficient analysis techniques might present a new limiting factor for the years to come [12].

In this chapter, we focus on this last problem, presenting some of the advances that have been developed recently in order to extract valuable information from molecular simulations. Most of these techniques root in the idea of free energy landscapes, a common framework in this area [25–27]. Nevertheless, there is no agreement in which is the most adequate way to represent them. This gives rise to

a large number of different proposals, usually hinging upon the particular question we ask ourselves.

Some methods rely on low-dimensional representations of the free energy landscape. Processes like the unfolded-folded transition of a protein are ideally understood as a one-dimensional phenomenon, were the system transits from one state to the other. This simple vision requires identifying the appropriate coordinate, and to integrate out all other "meaningless" degrees of freedom. Several techniques have been developed to identify meaningful order parameters which capture the essence of our system, allowing for a systematic dimensional reduction [28]. For instance, Principal Component Analysis comes from the statistical analysis world, and it has been very useful for understanding molecular simulations [29, 30].

Other methods do not rely on any—arbitrary or not—choice of order parameters or collective variables and represent the free energy landscape as a kinetic network of states with certain populations, and transition rates between them. This is usually known as Markov state [12, 31, 32]. This vision provides a very intuitive way of picturing the free energy landscape of our system. Nevertheless, it suffers from the challenge of finding unbiased and systematic ways of defining such states.

2.2 Free Energy Landscapes

One of the most challenging problems current research in physics, chemistry or biology, providing an unbiased understanding of the static and dynamic behavior of complex systems such as spin glasses, atomic clusters, or biological macromolecules. The main difficulty lays on the large number of coupled degrees of freedom these systems contain, and on the wide range of involved time scales at which they behave. Such systems involve typically several metastable states, responsible of very slow dynamics. Thus, describing the relevant configurations, and specially their kinetic behavior presents a serious challenge from the theoretical, experimental and computational perspective.

In this context, free energy landscapes have become a common term [25–27, 33]. Mathematically, if $\mathcal{X}$ is the conformational space of our system[1]—where typically $\mathcal{X} = \mathbb{R}^n$, with n is the number of degrees of freedom of our system, being $n = 3N$ if we have N atoms in coordinate space—then the free energy landscape is a continuous function $F : \mathcal{X} \to \mathbb{R}$, which associates every physical conformation $X \in \mathcal{X}$ a free energy value. The free energy landscape can be then understood as a hypersurface in $\mathbb{R}^{n+1}$, where hills and valleys represent maxima and minima of F. These minima describe the metastable states, separated by free energy maxima, or free energy barriers. If we were to know completely the free energy landscape of a particular system, we would have all relevant information about it. Obviously, this approach poses a number of problems. The first one is how to represent it, as normally n is very large, so the hypersurface in $\mathbb{R}^{n+1}$ does not constitute the best representation if we wish to gain direct understanding. The second problem is how to sample this free energy landscape. As mentioned, complex systems have large n and slow timescales, so it might be unfeasible to tackle the problem with a direct sampling, as some minima might act as kinetic traps which slow down the dynamics.

[1] In general, only the conformational part of the whole state space Ω is relevant for free energy calculations.

A remark should be done on the difference between free energy landscapes and energy or potential free energy landscapes [6]. Mathematically, they are similar representations of the conformational space of the system, where every physical state has an assigned number, although it is different in each case. The potential energy landscape is in encoded in the Hamiltonian of the system, while the free energy landscape depends on the ensemble, and thus is temperature dependent. Thinking of protein folding, the folded an unfolded structure represent both potential local minima. Nevertheless, in a free energy landscape representation the temperature would determine which is the stable structure. At high temperatures, the denatured state constitutes the deepest free energy minima, while at low temperatures, the native state is the stable configuration.

2.2.1 Of Low Dimensional Representations of the Free Energy Landscape of the System

We start with a brief reminder of statistical mechanics. In the canonical ensemble, the state of a system of N particles at temperature T is defined by the canonical partition function

$$\mathcal{Z} = \frac{1}{h^{3N}} \iint e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/kT} \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}, \tag{2.1}$$

where $\mathbf{q}$ are the $3N$ generalized coordinates of the system and $\mathbf{p}$ the conjugated momenta. $\mathcal{H}(\mathbf{p},\mathbf{q})$ is the Hamiltonian of the system and h a normalization constant. The probability $\pi(\mathbf{p},\mathbf{q})$ of finding the system in a particular configuration of the state space Ω, $\mathbf{x}(\mathbf{p},\mathbf{q}) \in \Omega$ is given by

$$\pi(\mathbf{p},\mathbf{q}) = \frac{e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/kT}}{\iint e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/kT} \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}}. \tag{2.2}$$

This allows us for example to compute the ensemble average of a particular observable $\langle f \rangle$ which is defined on state space Ω as

$$\langle f \rangle = \frac{\iint f\mathbf{p},\mathbf{q}) e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/kT} \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}}{\iint e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/kT} \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}} = \iint f(\mathbf{p},\mathbf{q})\pi(\mathbf{p},\mathbf{q})\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}. \tag{2.3}$$

For example, the average energy $\langle E \rangle$ can be expressed as,

$$\langle E \rangle = \iint E(\mathbf{p},\mathbf{q})\pi(\mathbf{p},\mathbf{q})\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q} = \frac{1}{\mathcal{Z}} \int E_X e^{-\beta E_X} dX = -\frac{1}{\mathcal{Z}} \frac{\partial}{\partial \beta} \mathcal{Z} = -\frac{\partial \log \mathcal{Z}}{\partial \beta}, \tag{2.4}$$

where $\beta = (k_B T)^{-1}$. Now, the free energy of the system is tightly related with the partition function, as:

$$F = -k_B T \log \mathcal{Z}. \tag{2.5}$$

This equation forms the fundamental connection between thermodynamics and statistical mechanics in the canonical ensemble, meaning that knowing F is equivalent to knowing $\mathcal{Z}$, and thus, with F we have all the available information about our system.

Computing explicitly the partition function is unfeasible for most interesting systems. Usually, it is enough to compute free energy differences, which is a tractable problem, as it requires just to compute ensemble averages as in Eq. (2.3), which are easier to evaluate. There are two different different problems when it comes to calculate free energy differences. The first one is to determine the change in free energy of a system as a function of a model parameter λ. In this case, the Hamiltonian of the system depends on some external parameter λ which we can control to induce transitions between two values $\lambda = A$ and $\lambda = B$. Then, we have two partition different functions $\mathcal{Z}_A$ and $\mathcal{Z}_B$, and the free energy difference is

$$\Delta F_{AB} = -k_B T \log \mathcal{Z}_B / \mathcal{Z}_A. \tag{2.6}$$

The second option is to compute differences of free energy between metastable states within the *same system*. This is equivalent to know the relative value of the free energy for a subset of particles in the state space Ω. It is convenient to express this free energy as a function of a (or set) of coordinates R. This approach provides an intuitive vision of the free energy of the system, subject to a proper choice of R.

We define a collective variable $R(X)$ which is function of the conformations X of the system [2]

$$R(X) \equiv R(\mathbf{q}) = R(q_1, q_2, \cdots, q_{3N}). \tag{2.7}$$

This variable is a function of the positions of the system, for example the distance between two groups of atoms, some torsional angle, or combinations of such kind of quantities. All up to the imagination of the researcher. This coordinate restricts the system to a hypersurface $R(\mathbf{q})$ in phase space, so that the free energy F_R, the partition function $\mathcal{Z}_R$ and the collective-variable probability p_R, on this restricted hypersurface are the magnitudes of interest.

The phase-space probability $\pi_R(\mathbf{p}, \mathbf{q}; R')$ of finding the system at a particular value of R' of the collective variable R is

$$\pi(\mathbf{p}, \mathbf{q}; R') = \pi(\mathbf{p}, \mathbf{q}) \delta(R' - R(\mathbf{q})). \tag{2.8}$$

Then, the probability $p_R(R')$ restricted to the R—which is the probability computed *along* the collective variable R'—can be obtained by averaging out all remaining degrees of freedom,

$$p_R(R') = \frac{\iint e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/k_B T} \delta(R' - R(\mathbf{q})) \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}}{\int \int e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/k_B T} \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}}, \tag{2.9}$$

and the restricted partition function,

$$\mathcal{Z}_R(R') = \frac{1}{h^{3N}} \iint e^{-\mathcal{H}(\mathbf{p},\mathbf{q})/k_B T} \delta(R' - R(\mathbf{q})) \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q}, \tag{2.10}$$

so we can write

$$p_R(R') = \frac{\mathcal{Z}_R(R')}{\mathcal{Z}}. \tag{2.11}$$

The free energy along the collective variable R can be now calculated simply as:

[2] Recall that we define R in the conformational space $\mathcal{X}$, not the full state space Ω. Usually collective variables are defined over the conformations the system adopts, ruling out velocities.

$$F_R(R') = -k_B T \log \mathcal{Z}_R(R') = -k_B T \log p_R(R') - k_B T \log \mathcal{Z}. \tag{2.12}$$

This quantity $F_R(R')$ is a very relevant one in molecular simulations, and Eq. (2.12) gives a practical form of calculate it. If R is a coordinate by which we can properly describe the system[3], $F_R(R')$ is the magnitude by which we can understand it. Through this projection we have free energy wells which represent stable configurations, and free energy barriers separating them. In practice, if we have a perfect sampling of the configurational space $\mathcal{X}$, computing $F_R(R')$ is straightforward. We just have to calculate the value of R for each configuration $R_1(X_1)$, $R_2(X_2)$, $\cdots$, and calculate the probability of each value R along the simulation, obtaining $p_R(R')$. $F_R(R')$ is recovered from expression (2.12) up to an insignificant constant. In most cases, perfect sampling is not possible. A number of different techniques have been proposed to estimate $p_R(R')$, for example, biasing the system, which allows to decrease barriers and help the system jump over them in reasonable computational times [19, 23, 33, 34].

In general, R does not have to be a single collective variable $R \in \mathbb{R}$. For example, it is often to compute two-dimensional representations of the free energy landscape, along two different collective variables R and S. Then we have $F_{R,S}(R', S')$, and all the derivation above is easily applicable to the case.

Of Free Energy Profiles and Potentials of Mean Force

A plot of $F_R(R')$ as a function of the collective variable R' is usually termed as a free energy profile, meaning the projection of the whole free energy landscape onto a single coordinate. Nevertheless, the term *Potential of Mean Force* (PMF) is often mentioned in these contexts, and both terms are used as synonyms, although they do not mean strictly the same thing.

The PMF was introduce by Kirkwood in 1935 [35] and literally, is the average force one should perform to constraint our system at a particular value of R. Equation (2.12) can be rewritten as:

$$F_R(R') = F_R(\infty) + \int_{\infty}^{R'} \frac{\mathrm{d}F_R(R'')}{\mathrm{d}R''} \mathrm{d}R'', \tag{2.13}$$

where $-\mathrm{d}F_R(R'')/\mathrm{d}R''$ is the mean force needed to keep the system at R''. In this sense, both term are in most cases interchangeable. In practice, the estimation of $F_R(R')$ from a molecular simulation is called the free energy profile along coordinate R.

Some methods bias the system to force it to sample different regions along R which are highly inaccessible by direct dynamics. For example, an external force can be used to constrain the system at a particular value of R, such as in Umbrella Sampling [19]. In such cases, the recovered profile is usually called the PMF rather than the free energy profile. Nevertheless, the particular relation between the PMF and the free energy profile is subtle and rather technical. The underlying idea is that by constraining our system with a force we are altering the actual the phase

[3] Typically, R determines a slow degree of freedom, representing motions of low frequency. In this sense, all high frequencies are integrated out as they do not contain relevant information about the system—they can correspond to simple atomic vibrations, or thermal fluctuations—and the lower ones define the meaningful coordinates for a particular process.

space mainly due to the contribution of the velocities. More details on this point are specified on [36, 37].

Of Reaction Coordinates and Order Parameters

So far, we have carefully referred to R as a *collective variable*. In many cases, R is called an *order parameter* or *reaction coordinate*, in a rather loose way. These two concepts are not the same, although they are used frequently as synonyms through the literature.

An order parameter indicates the degree of order in the system, or more generically, it is a variable chosen to describe changes in a system. In this context, order parameters are collective variables used to describe transformations from the initial state to a target one. The representation $F_R(R)$ along some order parameter R describes transitions the system shows along the trajectory between some states, which are represented as free energy minima in the profile $F_R(R)$.

An order parameter may, although does not have to, correspond to the path along which the transformation takes place in nature. Then, we can call R to be a *reaction coordinate*. In a simple two-state picture, a free energy profile along a reaction coordinate is characterized by two free energy wells, representing the reactive and product states, separated by a free energy barrier. If R is a good reaction coordinate, this representation gives the real free energy barrier—the one from which the rates can be computed—and its maximum coincides with the transition state of the system. More details on this point are given in Section 2.3.

Finding proper reaction coordinates or order parameters is a difficult task, yet a central one in free energy calculations [33]. The choice of order parameters might have key consequences on the efficiency and accuracy of our free energy calculations. Nevertheless, there is not a definite answer for the general question of how to make the best (or at least an appropriate one) choice of an order parameter. In many cases, the answer lies in our intuition about the system.

2.3 Reaction Coordinates in Molecular Dynamics Simulations

Reaction Coordinates play a central role in understanding molecular processes, as most of them can be described in terms of an effective reaction between an ensemble of "reactant" states to "product" states. A clear example is protein folding, where the reactant ensemble is the denatured state and the product the native one.

Finding an adequate reaction coordinate to describe the reaction dynamics appears as a major challenge. Reaction coordinates provide an intuitive understanding about the overall behavior of the system. Also, they allow identifying properly the initial and final states (reactants and products) of the reaction, helping in the identification of possible intermediate states. Finally, as they monitor the reaction pathway, the transition state and the free energy barrier are properly defined, and so the transition rates, which are observables directly measurable in experiments.

The concept of transition state is an important one in this context. Transition states are the ensemble of configurations between reactants and products. Considering overdamped diffusion along the reaction coordinate, the transition state is the

maximum of the barrier, where the system has equal probability of falling to one or the other side.

Intuitively, we expect from a good reaction coordinate a meaningful description of the progress of a reaction. No significant detail of our system should be lost, given that all all the dynamics is projected onto this single coordinate. Poor choices lead to non-Markovian, long-time memory effects, whereas good reaction coordinates are essentially Markovian [38, 39]. This implies that, knowing the value of the reaction coordinate at a certain configuration, it is possible to predict the fate of the trajectory initiated from there. If our bad projection is collapsing two dynamically different states onto a same value, the Markovian condition is be fulfilled, as the trajectory fate depends on at which state we start.

2.3.1 How Good is my Reaction Coordinate?

The identification of proper reaction coordinates is crucial to validate our understanding of the molecular system. A first strategy to "rate" our reaction coordinate q is simple inspection, given than it should be able to distinguish properly the states visited along the dynamics.

Nevertheless, there are systematic test which help us in rating a reaction coordinate giving quantitative measures on the quality of q. We discuss here the Bayesian relation test, as shown in [38, 39]. Nevertheless, additional tests have been proposed [40].

Bayesian Test for Reaction Coordinates

Let us consider a molecular system with deterministic Newtonian or stochastic dynamics in the configurational space. First, we define transition paths as those trajectory segments that exit from the reactant region A and reach the product region B without crossing back to A and vice versa. For example, in the case of protein folding, they correspond to those fragments of trajectory starting from the unfolded ensemble and folding back without unfolding again, and the other way around.

Choosing the reaction coordinate q, we construct the probability distributions $p_{eq}(q)$—which gives simply the probability for the system of being on a particular value of q—and $p(q|TP)$— the distribution of probability of the transition paths on q. The first distribution is calculated over the equilibrium ensemble, while the latter over the transition path ensemble. These two probability distributions are related to each other through a Bayesian expression for conditional probabilities [38, 39],

$$p(q|TP)p(TP) = p(TP|q)p_{eq}(q), \tag{2.14}$$

where $p(TP)$ is the fraction fraction of time the system spends on transition paths and the new quantity $p(TP|q)$ is the probability for the system of being on a transition path (TP), given that the system is in q.

For a good reaction coordinate, $p(TP|q)$ should have a single and sharp peak, where all the transition states are collapsed into a single value of q. In the diffusive limit, this peak is equal to 0.5, the probability of, once in the transition state, go to the product state, or back to the reactant one.

In practice, this gives an easy way to test reaction coordinates. The equilibrium distribution p_{eq} can be computed directly from long equilibrium distributions

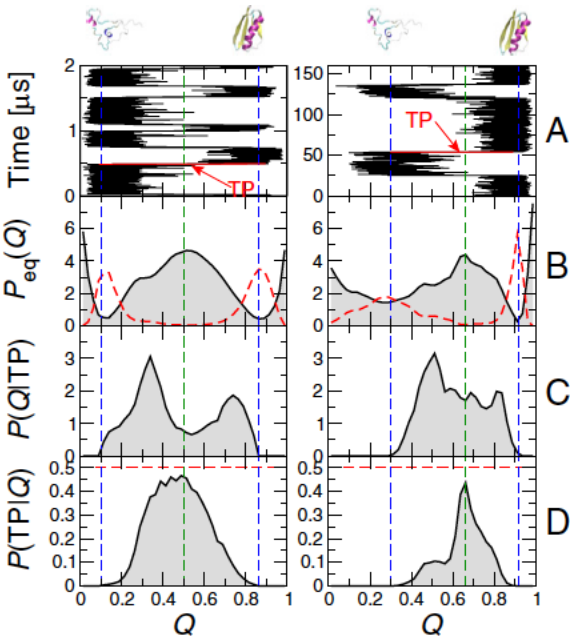


Figure 2.1: **Bayesian test on the fraction of native contacts**: The results for a Go model (left) and an all-atom model (right) are compared. Q is a proper reaction coordinate given the unimodal distribution shape of $p(TP|Q)$, where the peak approaches to the diffusive limit, (taken from [41]).

with multiple transitions, or estimated by some enhanced sampling method such as umbrella sampling [19]. From this distribution, it is straightforward to identify the reactants and produces states, and the transition paths, in order to compute $p(q|TP)$. Also, one might apply transition-path sampling [24], if its not possible to sample enough the TP ensemble.

2.3.2 Popular Reaction Coordinates in Molecular Simulations

In many cases, intuition is the best guide to choose a proper reaction coordinate. We review in this Section some of the most popular choices to describe molecular systems.

Root Mean Square Displacement (RMSD)

The RMSD is one a very popular reaction coordinate in molecular dynamics [22, 42]. It is a measure of the average distance between the atoms (usually just the heavy or the backbone atoms) of two superimposed proteins. In molecular simulation, this usually means to use the native structure as the reference one, an calculating the $RMSD$ of every frame X with respect to it. Mathematically,

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_{ij}^2}, \tag{2.15}$$

where r_{ij} is the distance between each of the N considered pairs, like the backbone or α-carbon atoms. The structures we compare must be aligned and with the center of mass subtracted, in order for the $RMSD$ to be properly calculated.

In protein dynamics, it is very useful for identifying conformational changes, or to distinguish between the native and unfolded state. Nevertheless, big conformational changes such as the latter, usually yield problems in the alignment step, making the RMSD hard to be trusted. Also, it has often been claimed that different conformational states can fall onto very similar RMSD values, for example when a hinge motion takes place. Many criticism exists also on the RMSD to describe nucleic acids, such as RNA [43].

Fraction of Native Contacts (Q)

Native contacts are a popular concept especially in protein simulation, although it is also used for nucleic acids and other biomolecules. Starting with a protein in its native structure, two atoms are said to be *native contacts* if they are closer than a certain cutoff distance, usually around 7.0 Å. In practice, usually backbone or heavy atoms are the only ones taken into account.

Native contacts have been often claimed to play a key role in the folding mechanism of a protein [11, 41, 44, 45], in relation with the"principle of minimal frustration"[4]. The evolution of native contacts Q along the dynamics is a common reaction coordinate, specially in protein folding. In Gō-models—coarse grained protein models whose Hamiltonian is defined upon the native contacts map, see Chapter 3—Q is the natural reaction coordinate. Nevertheless, it has been reported that Q is the collective variable able to capture the transition states in all-atom protein folding studies (see Fig. 2.1 and [41]).

In practice, there are several ways to define Q. After choosing which atoms would be accounted for (for example the α-carbons or the backbone atoms), one defines the native contact map matrix Δ_0, which has element $\delta_{ij} = 1$, if atoms i and j are closer than a certain threshold distance—usually 7.0 Å if we keep only the α-carbons—and $|i - j| > 3$. Either other case is $\delta_{ij} = 0$. The number of native contacts is $Q_0 = \frac{1}{2}\sum_{ij}\delta_{ij}$.

One option is to apply the same criterion to every frame of the simulation and define Q as the fraction of native contacts which survives. This is, calculate matrix Δ for every frame and compare it to Δ_0, checking which fraction of native contacts appear in both. An additional definition is suggested in [41],

$$Q(X) = \frac{1}{N}\sum_{(i,j)}\frac{1}{1+\exp[\beta(r_{ij}(X) - \lambda r_{ij}^0)]}, \tag{2.16}$$

where the sum runs over the N pairs of native contacts (i, j), $r_{ij}(X)$ is the distance between pairs i and j in frame X, r_{ij}^0 is the distance between i and j in the native configuration, λ is parameter which accounts for fluctuations and β is an smoothing parameter. This definition is a Fermi function and it allows a smoother calculation that the previous proposal, which is equivalent to applying an step function. Suitable parameters are $\beta = 5.0\,\text{Å}^{-1}$ and $\lambda = 1.2$.

[4] See Chapters 1 and 3 for further detail on this point and how it has motivated the development of coarse-grained models.

Geometrical Coordinates, the Radius of Gyration R_g and the End-to-End Distance

The radius of gyration R_g is a popular magnitude in polymer physics, used to describe the size of a polymer chain [46]. Thus, it can be used as a measure of a protein size over a molecular simulation, and to describe conformational changes which are related with significant changes in size, like it occurs in protein folding. Mathematically, R_g is defined as:

$$R_g^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \langle\mathbf{r}\rangle)^2 = \frac{1}{2N^2}\sum_{ij}(\mathbf{r}_i - \mathbf{r}_j)^2, \tag{2.17}$$

where N is the number of atoms, or monomers, $\mathbf{r}_i$ the vector position of the i-th atom, and $\langle\mathbf{r}\rangle$ their average positions. The first equation represents as the average displacement of the atoms from the chain average position, and second as the average of the square distances between pairs of atoms.

Polymer physics yields a number of analytical results on R_g which might be useful in protein dynamics, specially at high temperatures, when the interactions are ruled out, and the polypeptidic chain behaves as a free polymer. For example, for an entropically governed polymer chain, $R_g = 6^{-1/2}\sqrt{N}l$, where N the number of monomers and l the distance between monomers. This expression is useful to locate the denatured state.

In some occasions, the distance between two atoms within a molecule is an appropriate reaction coordinate. For example, chemical reactions depend on the transference of an electron, and thus its position describes the progress of the reaction. ligand:receptor binding or unbinding can also be monitored by studying the distance between the center of mass of both molecules (see Part III).

For the particular case of molecules under an external force, the direction of the pulling force becomes the natural reaction coordinate of the system. This is the case of many single molecule techniques (see Part III), where one fixes one end of the molecule and applies a force to the other. The end-to-end distance is defined as:

$$\xi = |\mathbf{r}_1 - \mathbf{r}_N|, \tag{2.18}$$

where $\mathbf{r}_1$ and $\mathbf{r}_N$ are the vectors positions of the first and last "particles" this is, the entities to which we apply the force. In the so called steered molecular simulations, this procedure is also a common one, allowing to gain insight about a molecule behavior, as the force tilts its free energy landscape allowing the system to sample it more effectively. Then we can subtract the effect of the force, and obtain information about the unforced system (see Chapter 1).

Optimized Reaction Coordinates

Most popular reaction coordinates are based on our "intuition" about the system. When this intuition fails, it is advisable to have some systematic technique to propose proper reaction coordinates. Most of these approaches rely on "optimizing" some definition of a reaction coordinate, in order to improve its quality based on some criterion or test, such as the Bayesian test.

For example one might describe each configuration from a molecular simulation through the contact matrix Δ_{ij}, whose elements are 1 if r_{ij} is smaller than a certain

cutoff distance and zero otherwise[5]. A reaction coordinate q can be the projection of Δ_{ij} onto some matrix with arbitrary weights $\tilde{w}$, $q = \sum_{ij} w_{ij}\Delta_{ij}/2$. The strategy is, starting from a matrix with arbitrary weights, find the set of weights which optimizes the "quality" of q, for example by applying the Bayesian criterion and looking for the best $p(TP|q)$ possible [38].

Along with the previous example of, several other techniques have been developed [47, 48], employing reaction coordinates which have their obvious advantages, but lack of a clear physical interpretation.

2.4 Reducing the Dimension of the System

Reducing the number of degrees of freedom from a large molecular system to a bunch of meaningful order parameters is an appealing approach for a number or reasons. Order parameters allow describing some transition or process in a simulation which is not evident by simple observation of the trajectories or of some observable we calculate. Also, they might be used to define an smaller but more meaningful configurational space for future construction of a Markov state model, this is define a proper metric for the definition of the states (see Section 2.5).

In this section, we review Principal Component Analysis (PCA), a widely employed method in the molecular dynamics community [29, 49].

2.4.1 Principal Component Analysis (PCA)

PCA is a statistical method that uses an orthogonal transformation to convert a set of coordinates (possibly correlated) into a new set (called the principal components, PCs) where the instantaneous correlations vanish. This transformation is defined such that the first PC has the largest possible variance, and each succeeding component has the highest possible variance, constraint to be orthogonal to the preceding ones. The resulting vectors form an uncorrelated orthogonal basis set.

It is a popular method in statistics and has also been frequently used in the Molecular Dynamics community to identify the linear subspaces where the largest-amplitude motions occur, hoping to relate such motions with relevant transitions in the system [29].

Let $\mathbf{x}$ be the vector of the N order parameters we used, for example the Cartesian position of the atoms of our molecule. Then, the covariance matrix $\tilde{C}$ for $\mathbf{x}$ is defined by the elements

$$c_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle. \tag{2.19}$$

The elements of the covariance matrix are the covariances between elements i and j if $i \neq j$ and the autocovariances if $i = j$.

By diagonalizing the covariance matrix,

$$\tilde{C}\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i, \tag{2.20}$$

we obtain the set of eigenvectors $\mathbf{v}_i$ and eigenvalues λ_i which coincide with the autocovariances of the PCs $\lambda_i = \sigma^2$. The eigenvectors are usually sorted according

[5]Recall that this definition is different from the fraction of native contacts as we do the configuration with the native structure.

to the magnitude of λ_i. If $\tilde{V}$ is the eigenvector matrix $\tilde{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_N]$ and the matrix of variances is $\tilde{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_N)$, then the PCs are defined by projecting $\mathbf{x}$ onto the principal eigenspaces:

$$\mathbf{q}^T = \mathbf{x}^T \tilde{V}. \tag{2.21}$$

As the PCs are sorted according to their autocovariance λ_i, some threshold is selected and those PCs with smaller autocovariance ignored. Thus, one keeps a smaller range matrix $\tilde{V}' \in \mathbb{R}^{N \times M}$, where $M < N$ is the number of PCs we keep. Used in this way, PCA is a way for dimension reduction, and thus the M dominant PCs can be used as new order parameters.

The PCs themselves contain valuable information about the system. Each PC is associated with characteristic "motions" of the system, which can be indentified by analyzing carefully the eigenvectors. The PCs $q_i(t)$ can be used as to calculate low-dimensional free energy surfaces along them. The first ones are associated with large-amplitude motions, and would show multimodal distributions (with several free energy wells) revealing states of the system and containing relevant information about the dynamics. The last ones have unimodal distributions and are associated with small-amplitude motions, likely the thermal fluctuations, so they do not provide any relevant information about the system dynamics.

We use the PCs in these two ways, as a tool for finding order parameters which contain meaningful information about the system and as a dimension reduction tool, yielding a subspace which might be employed for building Markov state models.

2.5 Markov State Models: a Network Description of the Configurational Space of our System.

Up to here, we have focused on understanding the free energy landscape through low dimensional projections. this approach needs finding proper order parameters which capture the essence of the system subject to study. This Section reviews Markov state models. This methods represents the free energy landscape as a kinetic network where the states are related to free energy basins. A rate matrix gives the probability to jump between states, and thus is related to the free energy barriers.tes, likely related to actual free energy basins, connected through a rate matrix.

2.5.1 What are Markov State Models?

Markov state models have become a popular and powerful way for analyzing and understanding molecular simulations [10, 12, 31, 32, 50, 51]. A Markov model is a network of conformational states and a transition probability matrix which describes the probabilities of jumping from one state to another within some time interval. Importantly, this network must be markovian for such time interval (from now one the lag time). This means that the probability of the system to evolve to a new state depends only on the present state and not on the past history. Also, detailed balance must hold, in order to represent the microscopic reversibility, or the equilibrium dynamics.

This picture allows to gain significant insight onto a molecule's properties as it gives a rather intuitive picture of their conformational space (the states and their

stability), and the rates connecting them. They provide also a direct connection with experimental measurements, being possible to project the dynamics onto some observable which can be directly compared with a signal obtained from a measure [10, 12, 52]. Additionally, they can be used as starting point for adaptive sampling methods [12], helping in running efficient simulations.

Nonetheless, the critical step is the definition of the states themselves [31]. Ideally, one should be able to arrive to a meaningful network of states which is easy to understand, and where the states correspond to free energy basins, where the boundaries between them are correctly identified. In Markov state models, these states are defined by kinetic criteria rather than purely geometric ones. This allows to fulfill the previous condition.

Geometric criteria are easier to meet, yet usually inadequate, as there are no physical reasons for which the free energy barriers would correlate with the geometric features of our system. For example, one could choose the fraction of native contacts Q or the RMSD, and define the states according to some fine partitioning on them. Likely, conformations with similar values of Q or the RMSD could be clustered onto the same free energy basin, even though they are not alike (similar values of the RMSD might answer to rather different conformations, like the pivoting of some hinge). In turn, conformations very separated in this space might transit between one another in fast times, and then should belong to the same basin.

The most common approach for building Markov state models is a two step process, where one uses first geometric, then kinetic criteria. First, small volume elements are defined according to some fine geometric criterion. They represent the *microstates* of the system. This first network is typically be very large and hardly understandable. Also, some groups of microstates show fast kinetic transitions between them (fast relaxing processes), while others follow slow kinetic relaxations. This scale separation is relevant to lump the microstate network into kinetic relevant clusters, associated to the free energy basins. They define the macrostates of the system. and the boundaries between them correspond to the free energy barriers.

In the following sections, we review the basic aspects of Markov state modeling. We start with a brief statement of Markov state model theory. We follow with a practical summary regarding construction of Markov state models for molecular systems, stressing the particular protocol we use in the present work.

2.5.2 Markov State Model Theory

We start by establishing the theoretical framework of Markov state models. The basic ingredient is the transition probability matrix, which has a particular physical interpretation, related with the dynamical model we use to understand our system [12, 51].

Continuous Molecular Dynamics

Consider state space Ω, which contains all dynamical variables needed to describe the instantaneous state of the system. For example, for a molecular simulation, Ω contains the positions and velocities of all the atoms of the molecule and of the surrounding bath particles. The state of the system at time t is characterized by $\mathbf{x}(t) \in \Omega$. We consider a dynamical process which is continuous in space and

can either be continuous (for theoretical treatments) or discrete (for computational purposes) in time. The dynamical process $\mathbf{x}(t)$ satisfies some properties:

1. $\mathbf{x}(t)$ is a Markov process in Ω, this is, the instantaneous change of the system is calculated based on $\mathbf{x}(t)$, without requiring previous history. This condition is true in general for most practical purposes. For example, in an all-atom simulation, as we are integrating the classical equations of motion, the trajectories are Markovian by definition. The same applies to Langevin dynamics. This might change if we look just on a subset of Ω (*e.g.* we take only the coordinates of the molecule of interest, or part of it).

2. $\mathbf{x}(t)$ is ergodic, this is, the dynamical process is aperiodic and Ω has not disconnected subsets that cannot be reached within one trajectory. So in $t \to \infty$, each point of Ω can be infinitely visited. This implies that any running average of an observable $f : \Omega \to \mathbb{R}^d$ is given by a unique stationary distribution $\pi(\mathbf{x})$, so that for every initial state $\mathbf{x}$,

$$\lim_{T\to+\infty} \frac{1}{T} \int_0^T \mathrm{d}t f(\mathbf{x}(t)) = \int_\Omega \mathrm{d}\mathbf{x} f(\mathbf{x})\pi(\mathbf{x}), \tag{2.22}$$

   so the fraction of time the system spends in any of its states during an infinitely long trajectory is given by the stationary density $\pi(\mathbf{x}) : \Omega \to \mathbb{R}_{0+}$, with $\int_\Omega \mathrm{d}\mathbf{x}\pi(\mathbf{x}) = 1$. This stationary density means that, if $P(\mathbf{x}, \mathbf{y}; \tau)$ is the transition probability density between two points $\mathbf{x}, \mathbf{y} \in \Omega$ within interval time τ, then,

$$\pi(\mathbf{y}) = \int_\Omega \mathrm{d}\mathbf{x} P(\mathbf{x}, \mathbf{y}; \tau)\pi(\mathbf{x}). \tag{2.23}$$

   This stationary density is unique, and in most relevant cases corresponds to the associated thermodynamic ensemble. In molecular simulations at constant temperature T, we have

$$\pi(\mathbf{x}) = \mathcal{Z}^{-1} e^{-\mathcal{H}(\mathbf{x})/k_B T}, \tag{2.24}$$

   with $\mathcal{H}(\mathbf{x})$ the Hamiltonian and $\mathcal{Z} = \int d\mathbf{x} e^{-\mathcal{H}/k_B T}$ the partition function.

3. $\mathbf{x}(t)$ is reversible, this is, the transition probability $P(\mathbf{x}, \mathbf{y}; \tau)$ fulfulls the detailed balance condition

$$\pi(\mathbf{x})P(\mathbf{x}, \mathbf{y}; \tau) = \pi(\mathbf{y})P(\mathbf{y}, \mathbf{x}; \tau). \tag{2.25}$$

   Physically, this means that in equilibrium, the fraction of systems going from $\mathbf{x}$ to $\mathbf{y}$ per unit time is the same as the fraction of systems from $\mathbf{y}$ to $\mathbf{x}$.

These conditions do not place too demanding restrictions on our dynamics. In practice, most stochastic thermostats are consistent with them.

Now, instead on focusing on the long time evolution of individual trajectories, we focus on the evolution of an ensemble density. We start with an ensemble of molecular systems at time t distributed in Ω with some probability density $p_t(\mathbf{x})$ which is different from $\pi(\mathbf{x})$. In a time interval of τ this density changes with the

action of the transition probability density $P(\mathbf{x}, \mathbf{y}; \tau)$. This change from $p_t(\mathbf{x})$ to $p_{t+\tau}(\mathbf{x})$ can be described by the action of a continuous operator, the propagator $\mathcal{P}(\tau)$, defined as

$$p_{t+\tau}(\mathbf{y}) = \mathcal{P}(\tau) \circ p_t(\mathbf{y}) = \int_\Omega \mathrm{d}\mathbf{x} P(\mathbf{x}, \mathbf{y}; \tau) p_t(\mathbf{x}). \tag{2.26}$$

As we apply the propagator $\mathcal{P}(\tau)$ to the density $p_t(\mathbf{x})$ we obtain a modified probability density, each time more similar to $\pi(\mathbf{x})$. In infinite time, any initial probability density becomes the $\pi(\mathbf{x})$. An alternative but equivalent description can be done in terms of the transfer operator $\mathcal{T}(\tau)$, which has some properties that will be useful later on [12, 51]. The difference is that $\mathcal{T}(\tau)$ does not propagate probability densities but functions $u_t(\mathbf{x})$, which differ from the probability densities by a factor of the stationary density, this is $p_t(\mathbf{x}) = \pi(\mathbf{x}) u_t(\mathbf{x})$. Thus

$$u_{t+\tau}(\mathbf{y}) = \mathcal{T}(\tau) \circ u_t(\mathbf{y}) = \frac{1}{\pi(\mathbf{y})} \int_\Omega \mathrm{d}\mathbf{x} P(\mathbf{x}, \mathbf{y}; \tau) \pi(\mathbf{x}) u_t(\mathbf{x}). \tag{2.27}$$

These two operators have some important properties

1. Both $\mathcal{P}(\tau)$ and $\mathcal{T}(\tau)$ fulfill the Chapman-Kolmogorov equation

$$\begin{aligned} p_{t+n\tau}(\mathbf{x}) &= [\mathcal{P}(\tau)]^n \circ p_t(\mathbf{x}), & (2.28) \\ u_{t+n\tau}(\mathbf{x}) &= [\mathcal{T}(\tau)]^n \circ u_t(\mathbf{x}), & (2.29) \end{aligned}$$

   which means that they can be used to propagate to arbitrarily long times $t + n\tau$.

2. $\mathcal{P}(\tau)$ has eigenfunctions $\phi_i(\mathbf{x})$ and eigenvalues λ_i,

$$\mathcal{P}(\tau) \circ \phi_i(\mathbf{x}) = \lambda_i \phi_i(\mathbf{x}), \tag{2.30}$$

   while $\mathcal{T}(\tau)$ has eigenfunctions $\psi_i(\mathbf{x})$ with the same eigenvalues λ_i

$$\mathcal{T}(\tau) \circ \psi_i(\mathbf{x}) = \lambda_i \psi_i(\mathbf{x}). \tag{2.31}$$

   If the dynamics are reversible, all λ_i are real and lie in the interval $-1 \leq \lambda_i \geq 1$. Also, the two eigenfunctions are related by a factor of the stationary density $\pi(\mathbf{x})$

$$\phi_i(\mathbf{x}) = \pi(\mathbf{x}) \psi_i(\mathbf{x}). \tag{2.32}$$

3. The eigenfunction with the largest eigenvalue $\lambda_1 = 1$ corresponds to the stationary distribution $\pi(\mathbf{x})$

$$\mathcal{P}(\tau) \circ \pi(\mathbf{x}) = \pi(\mathbf{x}) = \phi_1(\mathbf{x}), \tag{2.33}$$

   then the eigenfunction $\psi_1(\mathbf{x})$ is constant on all state space Ω as $\phi_1(\mathbf{x}) = \pi(\mathbf{x}) \psi_1(\mathbf{x}) = \pi(\mathbf{x})$.

This calculation of the eigenspectrum of the operators $\mathcal{T}$ and $\mathcal{P}$ becomes an important issue, as it allows us to decompose the dynamics into m slow dynamical processes and the remaining fast ones. In this sense:

$$\begin{aligned} u_{t+n\tau}(\mathbf{x}) &= \mathcal{T}_{\text{slow}}(n\tau) \circ u_t(\mathbf{x}) + \mathcal{T}_{\text{fast}}(n\tau) \circ u_t(\mathbf{x}) \\ &= \sum_{i=1}^{m} \lambda_i^n \langle u_t, \psi_i \rangle_\pi \psi_i(\mathbf{x}) + \mathcal{T}_{\text{fast}}(n\tau) \circ u_t(\mathbf{x}), \end{aligned} \tag{2.34}$$

where $\langle a, b \rangle$ is the scalar product of function b onto a. This decomposition is physically relevant as $\mathcal{T}_{\text{slow}}$ contains the dominant process, while $\mathcal{T}_{\text{fast}}$ all fast processes which are usually of little interest. The slow dynamics are a superposition of dynamical processes with an associated eigenfunction ψ_i or ϕ_i and an eigenvalue λ_i. These processes decay with time, until just the first term $\lambda_1 = 1$ remains, which gives the stationary density $\phi_1(\mathbf{x}) = \pi(\mathbf{x})$. Other eigenfunctions correspond to processes which decay with time, and are dynamical rearrangements which occur while the ensemble relaxes to the equilibrium distribution. We can associate a physical (measurable) timescale for each of these processes,

$$t_i = -\frac{\tau}{\log \lambda_i}, \tag{2.35}$$

so we write

$$u_{t+n\tau}(\mathbf{x}) \approx 1 + \sum_{i=2}^{m} e^{-n\tau/t_i} \langle u_t, \psi \rangle_\pi \psi_i(\mathbf{x}), \tag{2.36}$$

where we neglect all fast processes.

Discretization of State Space

Molecular simulations take place in a full continuous state space Ω, although with discrete timestep. Nevertheless, by construction, Markov state models require a discretization of the state space in order to obtain a tractable description of the dynamics. As already mentioned, Markov state models partition the state space into discrete states and compute the transition matrix which models the jump processes observed in the dynamics. In practice, this is not done by discretizing the propagator, bur rather by discretizing the state space and estimating the corresponding transfer operator from the simulation data we have.

We consider a discretization of the state space Ω into N sets S_i. This process can be a simple partition with sharp boundaries of the considered degrees of freedom, or of some reduced amount, for example by applying a dimension reduction technique. The stationary probability π_i to be in set i is given by the full density,

$$\pi_i = \int_{\mathbf{x} \in S_i} d\mathbf{x} \pi(\mathbf{x}), \tag{2.37}$$

where S_i is the i-th partition of the state space so that $S = \{S_1, \cdots, S_N\} : \bigcup_{i=1}^{N} S_i = \Omega$. The Markov state model is defined by a transition probability matrix $\tilde{T}(\tau) \in \mathbb{R}^{N \times N}$, which is the discrete approximation of the transfer operator $\mathcal{T}$. Physically, every

element T_{ij} is the time-stationary probability to find the system in set j at time $t+\tau$ given that at time t it was in set i.

Now, all derived theory an discussions done for the continuous case are equivalent here with the current definitions. For example, if we have a column vector $\mathbf{p}(t) \in \mathbb{R}^N$ giving the population of our sets S_i at time t, we can compute the probabilities after time τ as

$$p_j(t+\tau) = \sum_{i=1}^{N} p_i(t) T_{ij}(\tau), \tag{2.38}$$

and the stationary probabilities of the discrete states π_i

$$\pi^T = \pi^T \tilde{T}(\tau). \tag{2.39}$$

In the same way, we can make an eigenvalue decomposition of matrix $\tilde{T}(\tau)$ to find the N associated dynamical processes, and describe the system with just the m slow ones (according to some threshold criterion).

2.5.3 Practical Guide to Building Markov State Models

We give here a brief overview on some practical aspects regarding the building of Markov state models from molecular trajectories, emphasizing the different steps needed, and the requirements at each stage.

Defining the Microstates

The first step is to map the molecular trajectories onto a complex network, or graph, the microstate network. We classify the conformations the system visits according to some geometric criterium, in order to discretize effectively the conformational space. This geometric partition (or distance metric) should be kinetically relevant, this is, only conformations the system can jump between rapidly should be grouped together. In principle, no optimal general choice exists, and several choices can be made, depending on the particular problem [12, 53, 54].

The distance metric defines the conformational space to be discretized into individual bins. This metric should distinguish between rapidly interconverting conformations. Typically, it is useful to look for some meaningful order parameter or collective variable as a direct partitioning of the coordinate space $\{\mathbf{r}\}$ would yield to massive partitions. For example, the RMSD is often a reasonable choice for studying conformational changes in proteins [10, 31]. Other option is to resort to techniques for reducing the number of coordinates, such as PCA, and perform a fine partitioning directly on them [49, 55].

The partitioning of the conformational space might be done directly (i.e. into bins of equal volume) [32, 56] or by means of some clustering algorithm, such as k-Centers Clustering or k-Medoids clustering (see ref. [12] for review on some methods). At this point, the dynamical trajectory is translated into a sequence of discrete bins. The microstate network is defined by the transition count matrix C_{ij}, which counts the observed jumps between bins i and j, and the occupation vector π_i, which gives the weight of node i.

In this sense, we represent a dynamical trajectory $\{\mathbf{r}(t)\}_{i=1}^{N}$ as a weighted and directed complex network, where the nodes stand for conformational microstates, and the links observed transitions between such microstates.

Estimating the Transition Matrix

The microstate network is defined by the transition probability matrix, which is simply obtained from the transition count matrix C_{ij}. We define the transition probability matrix $\tilde{T}$ as,

$$T_{ij}(\tau) = \frac{C_{ij}}{\sum_k C_{ik}}, \tag{2.40}$$

where τ is the lag time of the model. T_{ij} gives the probability of observing a transition from state i to j in the unit time τ. This way of estimating the transition matrix is the most straightforward one, and works well when we have a large amount of data, ideally infinite. In practice, the estimation of the transition matrix might suffer some problems, mainly due to finite sampling or imperfections in the definition of the microstates.

The estimation of the transition matrix, requires choosing the way to count the transitions, given the lag time. There are two main ways to do this. The direct one is to look at independent transitions at the lag time τ. Assuming that the the conformations are sampled at some regular interval Δ, where $\tau = n\Delta$, for some $n \geq 1$, one can count transitions as $\sigma(0) \to \sigma(\tau)$, $\sigma(\tau) \to \sigma(2\tau)$, and so on. This approach is equivalent to considering τ the new sampling interval, instead of Δ, and "throwing out" the rest of the data.

The sliding window approach seems more appropriate, as it avoids some imprecisions the previous scheme might lead to. Here, one counts as $\sigma(0) \to \sigma(\tau)$, $\sigma(\Delta) \to \sigma(\Delta + \tau)$, $\sigma(2\Delta) \to \sigma(2\Delta + \tau)$, and so on.

Markovianity of the Model

Markov state models are expected to be Markovian in the chosen lag time interval. In principle, the continuous simulated dynamics, is Markovian as any configuration can be determined from previous one. Nevertheless, when discretizing the state space, we are coarsening the it, so the model might be Markovian only at longer time scales. We might have for example, long internal barriers into our states (if they have been defined poorly) which would violate the Markov assumption.

In this regard it is useful to test this condition in order to choose an appropriate lag time interval. Most tests lay on the validity of the Chapman-Kolmogorov equation

$$\tilde{T}(n\tau) = \tilde{T}(\tau)^n, \tag{2.41}$$

being n the number of steps of length τ. This equation implies that taking n steps in a model with lag time τ is equivalent as taking one step in model with lag time of $n\tau$.

A way to check this is to study the relaxation time scales of the system, which are related with the eigenvalues of the transfer matrix, as discussed previously,

$$t_i = -\frac{\tau}{\log \lambda_i}, \tag{2.42}$$

where t_i is the relaxation time and λ_i the i-th eigenvalue. Now, the relaxation times for a Markov model with a lag time of $n\tau$ should be the same as those for a Markov model with lag time τ,

$$t_i = -\frac{n\tau}{\log \lambda_{i,T(n\tau)}} = -\frac{n\tau}{\log \lambda^n_{i,T(\tau)}} = -\frac{-n\tau}{n \log \lambda_{i,T(\tau)}} = -\frac{\tau}{\log \lambda_{i,T(\tau)}}, \tag{2.43}$$

where $\lambda_{i,T(\tau)}$ is an eigenvalue of $T(\tau)$. Examining a plot of the relaxation timescales as a function of the lag time (see Fig. 2.3), their stabilization means that the models starts to satisfy the Markov assumption. The appropriate lag time to choose, or Markov time, is simply the smallest lag time that gives a Markov behavior.

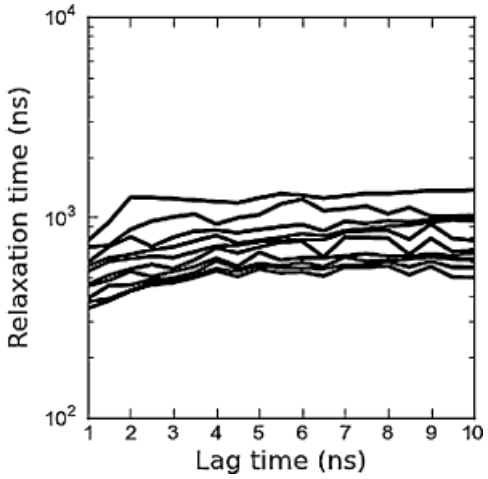


Figure 2.2: **Example of relaxation timescales as a function of the lag time**: The proper lag time is the minimum at which the system behaves markovian, in this case $\approx 2ns$ (picture taken from ref. [12]).

Detailed Balance

Detail balance, or microscopic reversibility, must hold for any meaningful Markov model. This implies that every time there is a transition from i to j, there should be a compensating transition from j to i. Not fulfilling this condition would lead to existence of sink or source states, that would avoid from representing the actual behavior of the system in long time scales.

There are different reasons why a model could not satisfy detail balance. Limited sampling is an obvious one is. Also it is not having true equilibrium sampling, or a poor definition of the microstates. For example, if they have been visited just once along the trajectory, they become source or sink states, and should be trimmed off the model. Detailed balance can be enforced by symmetrizing the count matrix,

$$\hat{C}_{ij}(\tau) = \frac{C_{ij} + C_{ji}}{2}, \tag{2.44}$$

where $\hat{C}_{ij}$ is the estimate of reversible counts, while C_{ij} of the actual transitions.

Ergodicity

A valid Markov state model must be ergodic, which implies that the network must be connected. Physically, it means that every state can be reached from any other arbitrary state. Disconnected components can arise when the model is built from various simulations starting at different conditions. When sampling is not enough, they would mix, or overlap, leading to a lack ergodicity. This is a clear feature of having poor sampling, and implies the necessity of collecting more data, or discarding the disconnected components, keeping just the largest one.

Defining the Macrostates

There are a number of reasons for coarse-graining the microstate model into a model made of larger macrostates. This final mesoscale model would be as quantitative predictive as the original one, and far more compact. Typical microstates networks are made of thousands of nodes, so it may be difficult to gain physical insight from the system by inspecting directly this model. The microstates can have little physical significance, so might have the transitions between them. Likely, there would be fast time scales involved in transitions within some groups of nodes, while slower timescales involved in more significant transitions between other groups. The first group of nodes should be lumped together, as the slow time scales define the metastable states of the free energy landscape.

When addressing this coarse-graining, two major problems should be faced. First we have to determine which microstates should be merged together. Second, it must be determined how many macrostates must be built. Several methods have been developed to answer these questions several methods have been developed. A popular one is the Perron Cluster Cluster Analysis (PCCA), which uses the eigenspectrum of a transition probability matrix to build the coarse-grained model [12, 57]. This method is based on identifying the slow timescales as the dominant eigenvalues in the spectrum of the transition probability matrix. This requires a clear gap between fast and slow scales in the eigenvalues. Additional algorithms are based more or less on this idea, such as PCCA+ -which improves the error propagation the simple PCCA suffers-, SHC, BACE, and many others (see [12] for review). Through this work we usea slightly different algorithm, proposed in [32, 56] and widely used by us since then [55, 58, 59]. We define it in the following lines.

The Stochastic Steepest Descent (SSD) Algorithm

This algorithm por detecting basins of attraction is inspired in the deterministic *steepest descent* algorithm for finding minima on a potential energy surface. The *Stochastic Steepest Descent* (SSD) algorithm was designed for detecting basins of attraction over a discretized free energy surface, as the microstate Markov state model network is [32].

Intuitively, this algorithm clusters the nodes according the probability flux relaxation, and thus it is based on kinetic criteria according to the system. Starting from a random state i, we concentrate the initial probability ($\pi_i(0) = \delta a, i, \quad \text{for} \quad i = 1, \cdots N$), and allow the probability distribution to evolve in time, letting the Markov chain to relax as $\pi(\tau) = \tilde{T}\pi(0)$. Starting from node a, we relax through the maximum probability flux, moving to some node b, where we can concentrate again all

the probability $P_i(1) = \delta_{b,i}$. This process is iterated until we end up on a node to which the probability flux leads. This is a *minimum* in the free energy landscape. By repeating this process over the whole network, we end up with a set of pathways, which drive the system to different minima. All nodes belonging to pathways leading to the same free energy minimum are defined to be in the same basin of attraction, and clustered within the same macrostate. In this way, we split the network onto a set of macrostates or basins of attraction.

We can define the procedure in a more formal way. Let us start by defining the auxiliary vector $\mathbf{\Omega} = \{\omega_i\}$ which labels the nodes, such that $i = 1, \cdots, N$.

1. We start with $\omega_i = 0, \cdots \forall i$.

2. We select a random node l such that $\omega_l = 0$ (not previously labelled), and place it as the first node in an auxiliary list L.

3. We chose among the neighboring nodes the one to which the maximum transition probability leads, $T_{lm} = \max\{T_{lj}, \forall j \neq l\}$, and we check that this node m satisfies one of the following conditions:

   (a) If $T_{lm} > T_{ml}$ and $\omega_m = 0$, then m was not previously visited and the step $l \to m$ is a descending free energy pathway. Node m is added to the list L and we go back to step (3), using m instead of l.

   (b) If $T_{lm} > T_{ml}$ and $\omega_m \neq 0$, the the step $l \to m$ is descending but m was already visited by the algorithm. It is allocated to all nodes in list L, $\omega_j = \omega_m, \quad \forall j \in L$, and we go back (2).

   (c) If $T_{lm} \leq T_{ml}$, then the transition $l \to m$ is not descending and we remove it from the network until the algorithm is over. We go back to step (3) unless $2D$ links for node l have been removed (where D is the dimension of the state space that was used for building the network), l is labeled as the local minimum of the net, and $\omega_j = l$ for every node in the list $j \in L$, so we go back to step (2). This restriction is associated with the dimensionality D and prevents transitions from a local minimum to any node on the same basin or to a node with less energy belonging to a different basin.

The process is over when every node in the network has been labelled, meaning that the algorithm has went through the whole network, splitting it into the individual basins of attraction (the labels we have been setting). The maximum descending flux pathways have been defined, relating every node with some other node labelled as local mimum. All nodes with this same label are kinetically related in the free energy landscape of the system.

One of the most important features of the SSD algorithm is that it scales with $N \log N$, unlike most algorithms which scale with a power of the system size [32, 56].

Coarse-Graining the Microstate Network

After we have applied the SSD algorithm—or any other one—we can redefine a new network, where the nodes and links will be built according to the coarse-grained

metastable macrostates the algorithm has found. Given π_i the population of microstate i, the population π_α of macrostate α is simply,

$$\pi_\alpha = \sum_{i \in \alpha} \pi_i, \tag{2.45}$$

while the transition probability from microstate i to j, the transition probability from macrostates α to β is

$$T_{\alpha\beta} = \frac{\sum_{j\in\beta}\sum_{i\in\alpha} T_{ij}\pi_j}{\sum_{j\in\alpha}\pi_j}. \tag{2.46}$$

This definitions ensure a proper normalization and the satisfaction of the detailed balance condition.

Given this final network, several thermodynamic and kinetic properties can be computed in a rather straightforward way:

1. The difference of free energy from state α and β is simply $\Delta F_\alpha = -k_B T \log \pi_\alpha/\pi_\beta$.

2. The entropy of the macrostates can be computed from the distribution of probability of the microstates belonging to it. Simply, $S_\alpha = -k_B \sum_{i\in\alpha} \pi_i \log \pi_i$.

3. The rate constant for the transition from basin α to β—assuming local equilibrium—is simply $k_{\alpha\beta} = T_{\alpha\beta}/\tau$.

4. The average escape time of basin α to any other one is $t_\alpha = \tau/(1 - T_{\alpha\alpha})$.

Building Free Energy Disconnectiviy Graphs

Given the intrinsic multidimensionality of the molecular systems, disconnectivity graphs are an appealing way of picturing the free energy landscape as represented by the Markov state model network [6, 60]. They constitute a hierarchical representation of the relative free energy as derived from the transition probability matrix in the network.

We start by defining a control parameter, namely the adimensional free energy F/k_BT, where $F_i/k_BT = \log \pi_w - \log \pi_i$ is the adimensional free energy of state i relative to the weightiest one w. This parameter will be used as a threshold for value for ranking the nodes in the network. Starting with its minimal value ($i = w$), we increase it, letting new nodes, together with their links, to appear. These nodes might appear linked to any of those which are already in the network, or as disconnected components which would get connected at some value of this threshold. This "top-down" procedure, allows as to produce a simple hierarchical organization of the states in the network, helping in visualizing its relation.

2.5.4 Analysis Protocol to be Used

In this Section we describe the analysis protocol we employ in this Thesis. Obviously it is not the only one, probably nor the optimal, but it has worked in a pretty solid and robust way so far [55, 58, 59].

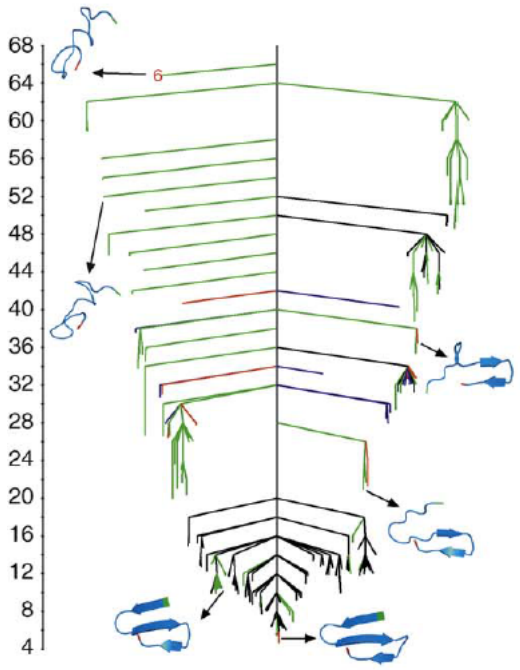


Figure 2.3: **Free energy dendrogram for protein pinWW**: A funnel folding structure is observed thanks to the disconnectivity graph representation, where no order parameter is needed to project the dynamics. (Picture taken from [60]).

1. If Ω is the state space of the our whole system (for example the biomolecule and surrounding bath particles if present), then $\mathbf{x}(t)$ denotes the state of the system at time t. First step is to decide the subset of Ω which is of interest to us. For example, if we are studying a coarse-grained protein model in implicit solvent, probably it would be the whole system. For an all atom simulation, just the molecule, or likely part of it, like the backbone or α-carbon coordinates.

2. Apply some method to reduce the number of coordinates, namely PCA. Just the first few coordinates are kept. In order to decide the threshold, inspect the accumulated sum of the eigenvalues of the PCA, and decide some cutoff, usually to keep $70 - 80\%$ of the cumulant autocovariance. These small set of coordinates become configurational space over which the Markov state model will be built.

3. Build the microstate Markov network by discretizing the configurational space into discrete bins. Then, build the count matrix by looking at the trajectory and counting jumps from bin to bin. The residence probability of each bin can be also calculated from the trajectory.

4. Check if the microstate model is meaningful, detail balance holds and it is Markovian, given the chosen lag time.

5. Apply the SSD algorithm to the microstate network in order to define the basins of attraction of the system, or free energy minima.

6. Check the significance of the coarse-grained model, and use it at convenience to analyze and understand the system subject to study.

2.5.5 Analysis of Markov State Models. Transition Pathways

A direct "look-and-see" approach to a Markov state model provides already a relevant amount of information about our system. For example, it is possible to identify directly the most relevant states in our system, as the occupation of the macrostates itself reveals their stability. Also, rates between states might be identified just by looking at the transition matrix, or we can compute some observable for the model and compare with experimental results.

Particularly, the problem of finding the set of pathways connecting two subsets of a network is very appealing, as it goes directly to the problem of protein folding. Markov state models can help in gaining insight to questions such as: how does an ensemble of denatured proteins find the unique native conformation? Is there a hierarchical folding? Is there a unique folding pathway? How do the tertiary and secondary structures form?

Formally, the problem is stated at follows. Let A and B be two subsets in the state space, for example the denatured ensemble and the native ensemble, respectively. All remaining states are labelled generically as intermediates I. Then, what is the probability distribution of the trajectories leaving A and moving to B? Or, what is the typical sequence of I states followed to transit from A to B? In the case of protein folding, this is just the determination of the folding routes the protein follows to reach the native state.

This generic question might be answered when a Markov state model is available and we apply Transition Path Theroy (TPT) [10, 61, 62].

Discrete Transition Path Theory

In order to describe TPT it is necessary to introduce the essential ingredient, the commitor probability q_i^+. The commitor probability is the probability that, when being at state i, the system would reach set B without passing back to set A. If we think of protein folding, it is simply the probability of folding. All states in A have $q_i^+ = 0$ and all states in B have $q_i^+ = 1$, by definition. For the states in I the commitor probability increases gradually as states are "kinetically closer" to B. The commitor probability is computed by solving the following system of equations:

$$-q_i^+ + \sum_{k \in I} T_{ik} q_k^+ = -\sum_{k \in B} T_{ik}, \qquad \text{for} \quad i \in I. \tag{2.47}$$

The backward commitor probability q_i^- (probability of, being in state i going to set A rather than B), for dynamics obeying detailed balance is simply

$$q_i^- = 1 - q_i^+. \tag{2.48}$$

Now, given two states i and j, the probability flux between them is given by $\pi_i T_{ij}$, which is the absolute probability of finding the system at this transition. We are now interested just in those transitions which move from A to B without

crossing back to A. The part of the flux belonging to those trajectories is obtained by multiplying the flux by the probability to come from A and move to B,

$$f_{ij} = \pi_i q_i^- T_{ij} q_i^+. \tag{2.49}$$

This is not actually the quantity that interests us, as we want to remove any contribution from recrossings or detours. For example, the system might jump multiple times between i and j increasing the flux, and we want a single transition per pathway. Thus, the net flux is:

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\}. \tag{2.50}$$

f_{ij}^+ defines the net flux and is a network of fluxes leaving states A and entering states B. This is a flux-conserving network, as the input flux equals the output flux for an intermediate state i, but for the source A and sink B. The total flux of the transition $A \to B$ is simply,

$$F = \sum_{i \in A} \sum_{j \notin A} \pi_i T_{ij} q_j^+ = \sum_{i \notin B} \sum_{j \in B} \pi_i T_{ij} (1 - q_i^+), \tag{2.51}$$

which gives the expected number of transitions from A to B per unit time τ. From this magnitude, the rate constant k_{AB} can be calculated as,

$$k_{AB} = F / \left(\tau \sum_{i=1}^{N} \pi_i q_i^- \right), \tag{2.52}$$

where m is the number of states. Obviously TPT is general for any network, so it might be not just applied on the microstate network, but also on the macrostate one, providing likely more convenient info.

Pathway Decomposition

The flux network obtained by TPT can be decomposed into individual pathways from A to B. If detail balance holds, the flux can be completely decomposed into pathways, with no cycles. A pathway decomposition consists merely in choosing a pathway P_1 and removing its flux $f(P_1)$ from the flux network. This process must be then repeated until the total flux F has been substracted and the network is free of pathways. This process is very useful as it provides a vision of the means by which the system transits from A to B. The strongest pathway, this is the one carrying maximum flux is of special importance, particularly if it has a flux comparable to the total one. The most convenient strategy is usually to identify first this pathway and remove it from the network, repeating this process from then on.

Chapter 3

Coarse-Grained Protein Models: the BPN_{46} as a Particular Non-Native Centric Model

This chapter aims to motivate the use of coarse-grained models for studying proteins and other biomolecules. We review briefly some of the existing approaches to build such models. Second, we present the protein model we use using extensively through this Part, the BLN_{46} protein model. We define the model, simulation protocols, characterizing it also from the thermodynamic and mechanical point of view.

3.1 Coarse-Grained Protein Models

3.1.1 Coarse-Grained Representations

Coarse-Grained models focus on the essential features of the particular system of interest and average out the "unneccesary" details. This provides a smaller system, significantly improving the efficiency in over three orders of magnitude, when compared to atomistic models. In this regard, and despite the improvement in computational tools in the last years, they have maintained a significant popularity, specially in soft matter and biomolecular systems [63–65].

Considering a particular biological macromolecule, a coarse-grained model represents it asa set of interaction sites (or "superatoms") that correspond to group of atoms in the system. This process is the "coarse-grained mapping", which captures the essential features of the original system while integrating out the irrelevant details. This mapping should fulfill at least two conditions: 1) preserve the basic features necessary for describing the phenomenon of interest, and the relevant slow, large amplitude motions of the system; 2) eliminate sufficient detail in order to provide a gain in computational efficiency, and filter out the high frequency, low amplitude fluctuations, which provide little information about the global properties of the system. Nevertheless there is little improvement in developing systematic mappings which fulfill such conditions. In this regard, most CG models rely on the physical or chemical "intuition" of the researcher.

For the particular case of proteins, one of the most popular ways to coarse grain is to represent each amino acid by one or few sites. Often, each amino acid is described as the α-carbon, keeping the backbone of the protein. This allows

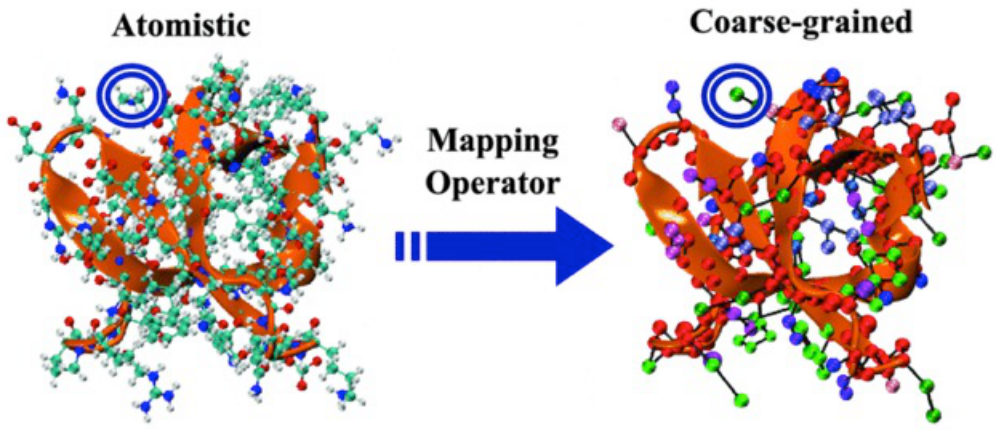


Figure 3.1: **Mapping of an atomistic protein structure to a coarse-grained structure**. Here, the amino acids are represented as few particles, capturing the essence of the system. Additionally, atomistic model often employ explicit solvents, increasing hugely the number of particles. This is usually eliminated in coarse-grained models, (picture taken from [63]).

to identify secondary structure elements in an unambiguous way, as the dihedral angles in the Ramachandran plot can be mapped to new coarse-grained angles in the α-carbon representation [66].

Ideally, it should be desirable to build coarse-grained protein models such that they represent the whole phenomenology of proteins, but more that an order of magnitude less particles. Unfortunately, this is not possible up to date, so alternative strategies are considered, like relying on the native structure of the protein, or classifying the amino acids into a small number of types based on properties such as hydrophobicity. We review briefly these strategies in the next two Sections.

3.1.2 Native Structure-Based Models

The development of protein folding theory and simulations of coarse-grained models have provided valuable information about the principles by which a protein folds rapidly to its unique native state. One of the strongest assumptions is that only native contacts play a significant role in the folding mechanism [11, 41, 44, 45], a statement motivated by the "principle of minimal frustration" [11], already discussed Chapter 1. This consideration is based on the supposition that the folding energy landscape has been designed by evolution such that the energy is correlated as far as possible with the nativeness of the structure, being misfolded traps or those ruled by nonnative interactions reduced or eliminated.

Native structure-based coarse-grained models lay on this principle, and are built or designed starting from the native structure, as seen in the Protein-Data-Bank. The parametrization of the model, or the definition of the "allowed interactions" is based on the concept of native contacts. Given the native coarse-grained representation $\mathbf{r}_0$, two residues are said to be native contacts if they are closer than a certain threshold distance. Based on this criterion, the interaction between pairs will be labelled as native or nonnative, independently of any physicochemical characteristic of the residues. The potentials involved in the model are set such that the energy is minimal for the native structure, stabilizing this structure.

Despite the apparent simplicity of the assumptions and the definition of the

model, this sort of coarse-graining has been tremendously successful for studying folding, fluctuations or interactions between proteins, also because of the computer efficiency they allow [9, 41, 67]. The particular definition of the native interactions allows different models. We focus on two kinds, network models and Gō (or native-centric) models.

Network Models

Network models are perhaps the simplest approximations, as they represent the protein as an elastic network based on the native configuration [68–72]. The interactions between contacts are simply quadratic functions, either of the distances between sites (Elastic Network Model, ENM [68]) or cartesian displacements of sites (Gaussian Network Model, GNM [71]). This kind of model provide insight on the elastic properties about the crystalized structure, particularly about the fluctuations of each site, which can be correlated with experimental properties.

For the ENM model, given the native configuration $\mathbf{r}_0$ [68],

$$U_{ENM}(\mathbf{r}|\mathbf{r}_0) = \frac{1}{2}\sum_{i<j} k_{ij}\Delta_{ij}(\mathbf{r}_0)|r_{ij} - r^0_{ij}|^2, \tag{3.1}$$

where r_{ij} and r^0_{ij} are the distances between sites i and j in configuration $\mathbf{r}$ and native configuration $\mathbf{r}_0$, respectively, while $\Delta_{ij}(\mathbf{r}_0)$ is the matrix of native contacts, being $\Delta_{ij} = 1$ if contacts i and j are native (closer than the a cutoff distance in the native structure) and 0 otherwise.

The GNM potential is expressed as [71, 72]:

$$U_{GNM}(\mathbf{r}|\mathbf{r}_0) = \frac{1}{2}\sum_{i<j} k_{ij}\Delta_{ij}(\mathbf{r}_0)|\mathbf{r}_{ij} - \mathbf{r}^0_{ij}|^2, \tag{3.2}$$

where $\mathbf{r}_{ij}$ and $\mathbf{r}^0_{ij}$ are the vector displacements from site i to site j in configurations $\mathbf{r}$ and $\mathbf{r}^0_{ij}$, respectively.

Clearly, the folded structure corresponds to the minimum in both models. In the case of the GNM model the the average fluctuations of each site and their covariance can be analytically determined from Gaussian integrals. The ENM model cannot be easily analyzed, but provides more realistic modes [70].

These network models represent usually each amino acid as a single site corresponding to the α-carbon, and employ single spring constants for bonded atoms. The only parameters involved in the model are the spring constant k and the cut-off distance r_c. These models reproduce reasonably experimental B-factors—which are easily calculated redefining the contact matrix [73]. Extensive developments have been produced for improving the model, including fitting of the constants [74], distance-dependent constants [69], as well as modifications of the model to include the possibility of breaking contacts. Dynamics are reproduced by simply integrating the equations of motion with a Langevin thermostat (or any other).

Gō-Models

This sort of models are based on the assumption that nonbonded interactions in the protein are ruled by the folded native structure, rather than by the character of the residues. In this regard, the Hamiltonian minimizes the potential of the folded

structure. Unlike network models, Gō-models include nonlinear potentials looking for describing interactions or the folding mechanism [9, 67].

Based on this simple idea, there are several ways to build a Gō-model. A simple and popular one, is to represent each amino acid as its α-carbon. Then, the nonbonded interactions between native contacts are represented as an attractive Lennard-Jones potential (or similar), and those of nonnative contacts as a simple repulsive excluded-volume potential. Bonds between adjacent residues in the backbone are modeled as stiff springs. Angle-dependent potentials, such as bending and dihedral interactions, can also be included, as they are useful for maintaining the preferred geometry of the peptide bonds.

In this way, we eliminate any energetic frustration, leaving a "funneled" landscape towards the native state [11]. This sort of models are useful for monitoring the evolution of a protein from an unfolded structure to the native one. Additionally, it has been often argued that they are able to reproduce successfully the folding mechanism, as well as correlating the folding rates for small proteins [41, 44]. Due to this, and also to the computer efficiency—reduced by several orders of magnitudes the number of particles to integrate, as N is the number of residues in the molecule–they enjoy a remarkable popularity, which has even increased in the last years.

3.1.3 Knowledge-Based Models

Native structure-based models work fairly well for the set of problems mentioned before. Naturally, they are not the end of the story. One must define a Hamiltonian for *each* single protein, and obviously proteins with unknown structures, or unstructured ones cannot be addressed by this approach. Also, the role of nonnative contacts on protein dynamics remains still as an interesting question. For example, misfolded structures could be driven by the formation of nonnative interactions, which stabilize the system in a nonnative structure [64].

In this sense, a different approach is to build transferable models, which are protein-independent potentials and work for modeling multiple proteins, for ideally reaching some sort of general coarse-grained protein model (if possible).

There exist different strategies for developing such models. For example, these potentials could be derived from the PDB statistics, this is, estimating the effective interactions between pairs of amino acids based on the statistical frequency of finding these contacts in the PDB data set [75]. Another strategy is to propose potentials which are optimized on the basis of known structures which appear PDB, and choose parameter sets such that α-helices or β-sheets become stable structures [76].

Although some advance has been made on these lines, the choice of potential and parameter sets is still *ad hoc*, and thus the models are built according to the system we wish to study.

Through this Part, we study in detail a protein model which falls onto this category. On the one hand, it constitutes a useful test model for different analysis techniques or explorations which can be hard to apply on atomic-detailed models due to the sampling problem. Also, providing they resemble the behavior actual proteins, they can be used for learning and gaining insight onto some of the structural or kinetic properties which govern this biomolecules.

3.2 The BPN_{46} Protein Model: Origin and Description

In this section, we present the BPN_{46} model, that we study through Chapters 4 and 5, following reference [55]. This protein model, despite its apparent simplicity, exhibits a rather rich behavior, becoming an ideal model to explore an analyze the different free energy techniques we have been describing so far.

3.2.1 Description of the Model

The BPN_{46} model is a coarse-grained off-lattice protein model which was introduced by Honeycutt-Thirumalai [77] and successively generalized by Berry *et al* to include harmonic interaction between next-neighbor residues [78]. This model has been widely studied through time, suffering different modifications, and being explored both in the context of thermally driven unfolding and subject to mechanical forces [79, 80]. It has 46 monomers which mimic the residues in a protein, in the α-carbon positions. These monomers belong to three different categories, hydrophobic (B), polar (P) and neutral (N). The residue sequence is given by 46 amino acids: $B_9N_3(PB)_3N_3B_9N_3(PB)_5P$.

This protein folds successfully into an stable four-strand β-barrel structure, stabilized by the hydrophobic core formed by the interaction between the two hydrophobic β strands which run parallel to each other and anti-parallel to the second and fourth strands (see Fig. 3.2 (A)). We can number the β strands, being β_1 the N-terminal hydrophobic strand and β_4 the C-terminal one. Numbering with the neutral turns T with a similar criterion, the protein structure can be represented as $\beta_1T_1\beta_2T_2\beta_3T_3\beta_4$.

The Hamiltonian of the system is defined by four different interaction terms:

1. **Next-neighbor interaction**: Harmonic springs sett the backbone of the protein i

$$V_1(r_i, r_{i+1}) = \frac{1}{2}K\sum_{i=1}^{N-1}(r_{i,i+1} - r_0)^2, \tag{3.3}$$

   where $N = 46$ is the number of monomers, and, in adimensional units, $K = 50$ is the spring constant, and $r_{i,i+1}$ the distance between neighbor residues i and $i+1$, with $r_0 = 1$ the equilibrium distance between residues. These parameters set a very stiff spring, which maintains the distance between residues almost constant.

2. **Bending interaction**: Three body angular potential, which accounts for the energy associated to bond angles,

$$V_2(\theta_i) = \sum_{i+1}^{N-1}\left[A\cos\theta_i + B\cos 2\theta_i - V_0\right], \tag{3.4}$$

   where θ_i is the bending angle formed by residues $i-1$, i and $i+1$ and $A = -k_\theta\cos\theta_0/\sin^2\theta_0$, $B = k_\theta/4\sin^2\theta_0$ and $V_0 = A\cos\theta_0 + B\cos 2\theta_0$, with $k_\theta = 20$, $\theta = 5\pi/12$ rad. This potential term, corresponds, up to second order, to a harmonic interaction term $\sim(\theta_i - \theta_0)^2/2$.

3. **Dihedral interaction**: Four body interaction, corresponding to the dihedral angle potential

$$V_3(\phi_i, \theta_i, \theta_{i+1}) = \sum_{i=1}^{N-3} \left[C_i(1 - S(\theta_i)S(\theta_{i+1}) \cos\phi_i) + D_i(1 - S(\theta_i)S(\theta_{i+1}) \cos 3\phi_i) \right], \tag{3.5}$$

where ϕ_i is the dihedral angle formed by the two planes defined by residues $(i-2)$-$(i-1)$-(i) and $(i-1)$-(i)-$(i+1)$ respectively. The parameters choice is $C_i = 0$ and $D_i = 0.2$ if two or more of the residues in the planes are neutral, and $C_i = D_i = 1.2$ otherwise. The tappering function $S(\theta_i) = 1 - \cos^{32}\theta_i$ is introduced to cure a problem of dihedral potentials, which appears when $\theta_i = 0$ or $\theta_i = \pi$. When three residues lay on the same plane, it is impossible to define the dihedral angle, which leads to a discontinuity in V_3. The tappering function does not introduce any extra minima in the potential, having little influence in the dynamics.

   This potential has three minima for $\phi = 0$ (trans state) and $\phi = \pm 2\pi/3$ (gauche states), being mainly responsible for the formation of secondary structures (see Fig. 3.2 (B)).

4. **All-residue potential**: A long range sequence dependent Lennard-Jones interaction between every pair i and j of residues,

$$V_4(r_{ij}) = \sum_{ij} \epsilon_{ij} \left(\frac{1}{r_{ij}^{12}} - \frac{c_{ij}}{r_{ij}^{6}} \right), \tag{3.6}$$

   where r_{ij} is the Euclidean distance between residues residues i and j and the parameters depend on the nature of the interacting residues, being attractive between hydrophobic residues and repulsive otherwise. Particularly:

   - $c_{ij} = 0$ and $\epsilon_{ij} = 4$ if i or j are neutral.
   - $c_{ij} = 1$ and $\epsilon_{ij} = 4$ if i and j are hydrophobic.
   - $c_{ij} = -1$ and $\epsilon_{ij} = 8/3$ otherwise.

Now, the Hamiltonian of the system, simply reads:

$$\mathcal{H} = \sum_{i=1}^{N} \frac{\mathbf{p_i}^2}{2m} + V_1(r_i, r_{i+1}) + V_2(\theta_i) + V_3(\phi_i, \theta_i, \theta_{i+1}) + V_4(r_{ij}). \tag{3.7}$$

Where m is the mass of each residue. In every moment we employ adimensional units. Nevertheless, real units might be recovered. Our distance unit is the distance between monomers, which can be taken to be that of α-carbons, thus $r_0 = 1 = 0.38nm$. The energy units might estimated as those of an H-bond, so $\tilde{\epsilon} = 1.7kT$. Now, mass units are taken as the average value for and amino acids, namely $m = 3 \times 10^{-22}kg$.

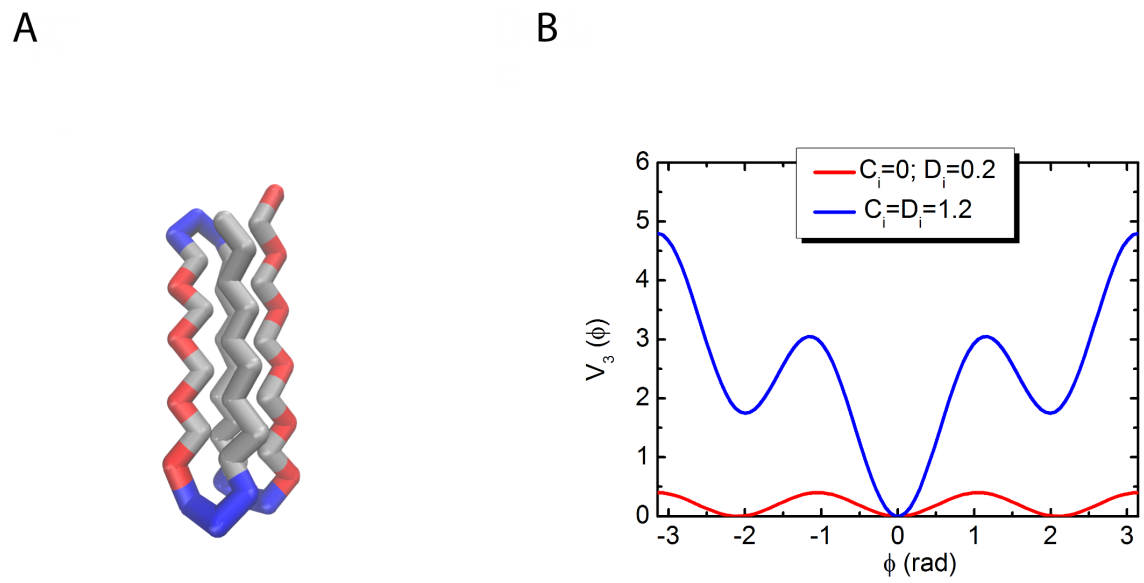


Figure 3.2: **A) Representation of the native structure**: The protein folds into a β barrel-like structure. Hydrophobic residues are colored in grey, neutral in blue and hydrophilic in red. **B) Dihedral potential for the two cases**: Red line shows potential for having at least two neutral residues, and blue other wise. The three stable states (one trans, two gauche) are clearly shown.

3.2.2 Simulation Protocol

The model is studied by performing molecular dynamics simulations integrating the Langevin equations of motion under different protocols [55]. For the equilibrium canonical simulations at temperature T:

$$m\ddot{\mathbf{r}}_\mathbf{i} = -\gamma\dot{\mathbf{r}}_\mathbf{i} - \nabla_i V_{BPN} + \eta_i, \tag{3.8}$$

where γ is the friction coefficient ($\gamma = 1$ in adimensional units), V_{BPN} the total intramolecular potential discussed above, and η_i a Gaussian white noise, of zero average and holding fluctuation-dissipation theorem $\langle \eta_i \eta_j \rangle = 2k_B T\gamma\delta(t-t')\delta_{ij}$.

The equations of motion are integrated with a second-order Runge-Kutta algorithm [81], using a timestep of $\Delta t = 0.005$. Our time units can be estimated as $\tau \approx 3ps$.

For thermal simulations, we run the dynamics starting from the native configuration, allowing the system to equilibrate for 10^6 timesteps, and running trayectories of $\sim 10^9$ timesteps. Usually, several trajectories from different initial conditions are run in order to obtain a better sampling. The particular details will be specified when convenient.

We are interested in the study of the system under the presence of external forces, with the spirit of single molecule experiments (see part III). This force might be applied in an in equilibrium or out-of-equilibrium protocol. In every case we attach the first monomer to a fixed spring, while the last one to another spring, responsible of setting the external force. The force is then applied in one direction, namely the z direction.

The out-of-equilibrium protocol sets a constant loading rate to the system. This is, the spring retracts at a constant velocity V, being the external force $F_{ext} = k(x - Vt)$, where k is the constant of the spring, here $k = 30$ in adimensional units (see part III for discussion on the influence of this spring constant). This pulling

protocol induces an increasing loading force which unfolds mechanically the protein, being the pulling velocity V a critical parameter for the study of the system.

The equilibrium protocol uses the same set-up but setting a constant force, often known as the force-clamp protocol. Here, the spring attached to the last monomer is simply pulled a distance z_0 so that $F_{ext} = kz_0$ is the desired force. This protocol is equivalent to "tilting" the free energy landscape of the system in the direction of the pulling force. This is the pulling mode we use in Chapters 4 and 5.

3.3 Thermodynamic Properties and Behavior Under Force

The BLN_{46} protein exhibits three different transition temperatures, at is has been reported [79, 80], the glassy temperature T_g, the folding temperature T_f and the collapse or critical temperature T_c. The native configuration is degenerated, with multiple alike configurations which are separated by large barriers [82]. The glassy temperature T_g indicates the temperature below which the system can be trapped in local minima of the potential, freezing in non-native conformations. The folding temperature might be defined in different ways. For example, the temperature at which the probability to visit the native configuration is 1/2. Finally, the critical temperature T_c is the "proper" thermodynamic transition identified by the peak in the heat capacity. It distinguishes between random coiled configurations and collapsed or structured ones.

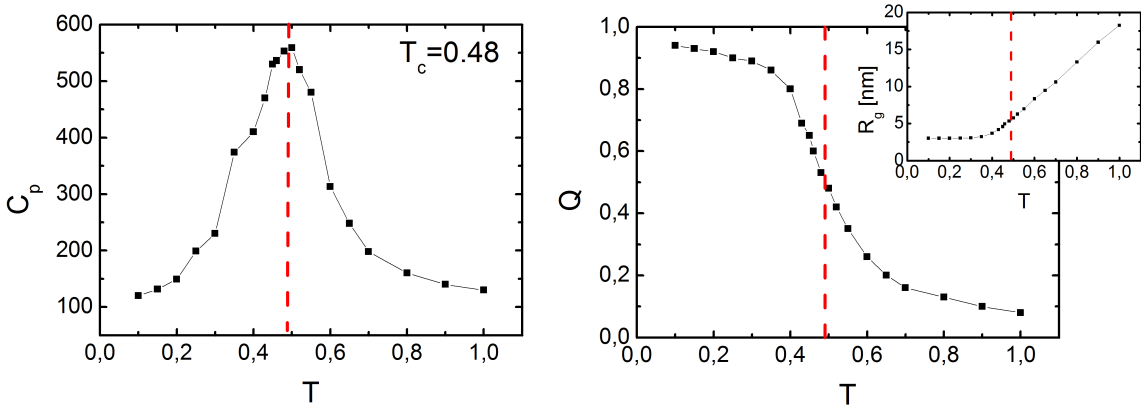


Figure 3.3: **Heat capacity and fraction of native contacts as a function of temperature**: The critical temperature (separating folded configurations from random coiled ones) is identified as the peak of the heat capacity, coinciding also with the temperature at which the fraction of native contacts is 0.5.

Figure 3.3 shows the heat capacity and the fraction of native contacts as a function of temperature, allowing both magnitudes to identify the critical temperature $T_c = 0.51$. As we show below, the protein exhibits a complex configurational space, with many allowed metastable configurations. As temperature increases, the system starts to populate other configurations rather than native-like ones. The decrease in the fraction of native contacts reveals this feature. Over the transition temperature it nearly drops to zero, revealing that the system is dominated by random coiled-like conformations, as suggested by the peak in the heat capacity.

We simulate the system in the presence of an external mechanical force. Figure 3.4 shows the average fraction of native contacts as a function of the constant pulling force. The chosen temperature is $T = 0.35$, above the glassy temperature but below the critical one. We see two types of transition in this curve. First, around $F \approx 0.5$ there is a drop in the fraction of native contacts to $Q \sim 0.5$, due to the population of a metastable state (the half-stretched configuration) which we discuss further on. The unfolding force is at $F_U = 1.1$, when the system unfolds totally to $Q \sim 0$.

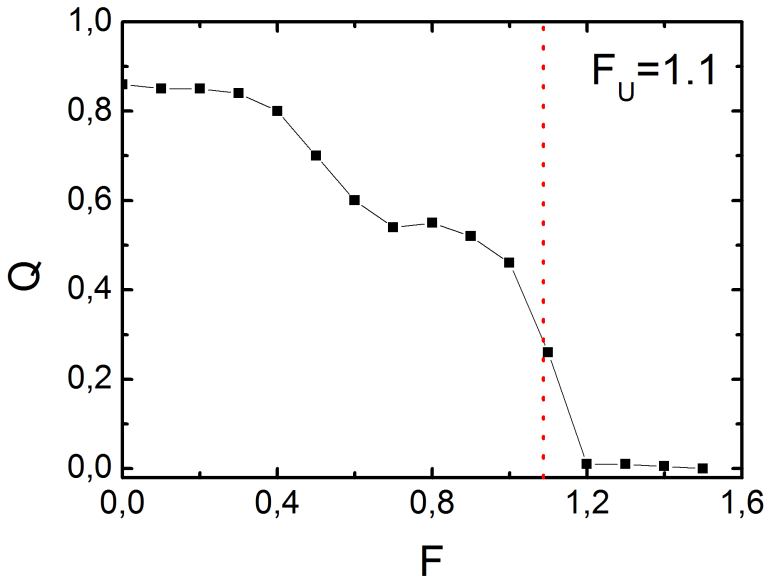


Figure 3.4: **Fraction of native contacts as a function of the pulling force**: Two drops are clearly seen, one at $F \approx 0.5$ and other at $F \approx 1.1$. The first one is due to the excitation of an intermediate structure. The second one is the unfolding force F_U.

Recall that in real units, the critical temperature is $T_c \approx 1.15kT$ (little bit above room temperature) and the unfolding force $F_U \approx 30pN$, in the range of actual unfolding forces for actual proteins (for not to mechanically stable ones, though).

Chapter 4

Mechanical Unfolding of BPN_{46}: Reaction Coordinates and Free Energy Profiles

In this chapter, we describe the mechanical unfolding of protein BPN_{46} employing low dimensional representations of the free energy landscape. First, we explore two reaction coordinates, the end-to-end distance ξ—the natural reaction coordinate due to the presence of the external force—and the fraction of native contacts Q. We employ the Bayesian test to rate their quality as reaction coordinates. Next, we explore the use of PCA to find order parameter with which describe the system. Finally, we build two-dimensional free energy surfaces with different combinations of collective variables.

4.1 Motivation

The present and following chapters are related to studies of the equilibrium ensemble of coarse-grained protein BPN_{46} under the presence of a constant mechanical force. This protocol—known experimentally as the force-clamp set-up—allows equilibrium transitions between different states by "tilting" the free energy landscape in the direction of the pulling force [83, 84].

The main interest of the present and next Chapters is to explore the different analysis alternatives we have discussed in Chapter 2. Particularly, we confront two different frameworks to describe the free energy landscape of the system. First, the use of low dimensional projections along different order parameters (present Chapter). Second, a Markov state model description (Chapter 5). We focus on a proper description of the conformational state and the unfolding mechanism, opposing the both methods. These two chapters follow reference [55], with extended containts.

4.2 Simulation Protocol

The molecular dynamics trajectories we analyze in Chapters 4 and 5 are produced in the following way. We integrate the Langevin equations of motion for the BPN_{46} protein in a force-clamp protocol (see Section 3.2.2), fixing monomer one and applying a constant force to the last monomer. In this way, we run equilibrium simulations.

The simulations are performed at a force of $F = 0.8F_U$ (applied in along the z coordinate), and temperature of $T = 0.55T_c$, between the critical temperature but above the glassy one. The chosen force optimizes the number of configurations visited by the system, particularly the number of unfolding events. Larger forces would tilt the landscape to the stretched configuration, hindering refolding events. Lower forces would keep too large barriers towards the stretched configuration, which would hardly be visited.

We run a total of six equilibrium trajectories of $\sim 3\,\text{ms}$ long with a preheating time of $\sim 3\,\mu\text{s}$. Configurations are stored every 1000 time steps, which is $\sim 15\,\text{ps}$.

4.3 One Dimensional Descriptions: the Free Energy Landscape Along Different Reaction Coordinates

We start describing the system through one-dimensional free energy profiles along two different reaction coordinates, the end-to-end distance and the fraction of native contacts. The first one is the natural reaction coordinate as the force imposes a privileged direction, narrowing the landscape along the pulling coordinate. The fraction of native contacts is widely used to describe protein folding, although our model is not based on any native-centric assumption

4.3.1 The End-to-End Distance and the Fraction of Native Contacts as Reaction Coordinates

Free Energy Profile Along the Fraction of Native Contacts and the End-to-End Distance

Figure 4.1 shows the free energy profile along the end-to-end distance ξ and the fraction of native contacts Q for a constant force $F = 0.8F_U$. The profile along ξ shows four clear minima that can be identified with four different configurations, considering that each of the four β strands have a length of $\xi \sim 3\,\text{nm}$. Clearly, the native configuration (N) has $\xi \sim 0\,\text{nm}$, as the extremal β strands are oriented in the same direction. The fully stretched configuration (S) has $\xi \sim 12\,\text{nm}$, as the protein is fully extended as an stretched polymer ($\xi \approx 0.38 \times 46$). In between, we find two different intermediates. The aligned configuration (Al) has the second strand $(PB)_4$ bent, so the ends of the molecule are oriented in the pulling direction and $\xi \sim 3\,\text{nm}$, the length of a β strand. The most stable configuration is the Half-Stretched configuration (HS), where $\xi \sim 6\,\text{nm}$ and the fourth $(PB)_5$ strand is unfolded, aligning both extremal strands in the direction of the pulling force.

These four states are also identified in the Q profile. State S has all contacts broken, so $Q \sim 0$. States HS and Al appear almost overlapped in the profile, with $Q \sim 0.5$ and $Q \sim 0.4$ respectively. Finally the N configuration corresponds to the well in $Q \sim 0.9$.

The HS configuration is the most stable state, according to both profiles, and the N configuration has similar stability. The stability of the HS is easy to understand. First, this configuration allows to have both extremal strands oriented in the direction of the pulling force. The energetic cost of separating the β_4 strand is

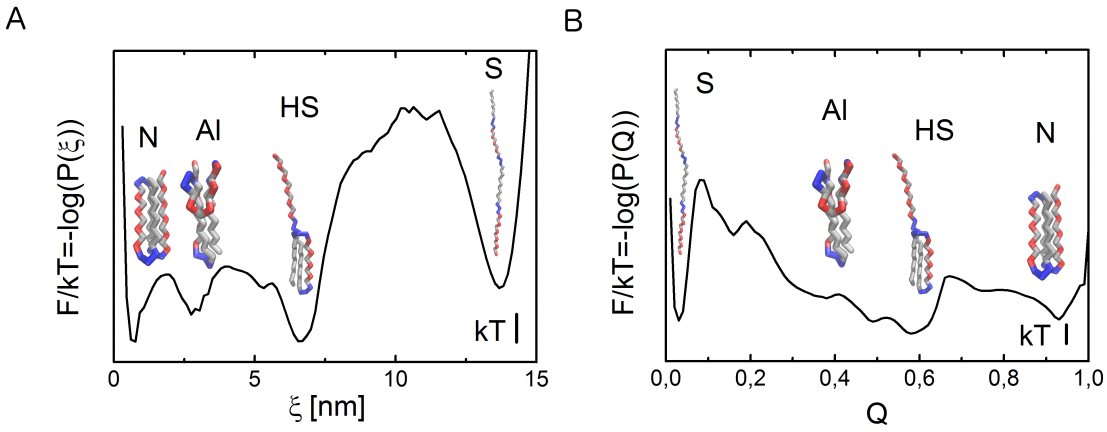


Figure 4.1: **Free energy profile along the end-to-end distance and the fraction of native contacts**: Both profiles depict a similar scenario, with four main free energy wells which can be associated to four different configurations, the native N, the stretched S and two intermediates, the half-stretched HS and the Aligned Al states.

low, as the only involved interaction are between its hydrophobic residues and the hydrophobic core. This configuration maintains intact this hydrophobic core, the main responsible of the stability of the folded structure. The aligned configuration is an alternative solution for aligning the ends in the direction of the force. The β_2 strand is bent in the middle, keeping also the hydrophobic core, but breaking more interactions than in the HS configuration.

Regarding the free energy barriers, the N and HS are separated by low barriers of around $2kT$. This implies that the transitions between both states would be relatively fast. The S configuration is separated by a large free energy barrier of $\sim 7kT$, so the waiting time to reach this configuration is expected to be large.

This conclusion is tested by simple inspection of a trajectory along the ξ coordinate. Figure 4.2 shows a piece of 120 µs of trajectory where this scenario is clear. In a time scale of ~ 100 ns the system is involved in fast transitions between the N, HS and Al states. Then, in a µs scale the system unfolds completely, visiting the S configuration. This vision agrees with the free energy profile shown in 4.1.

The remaining question is to unveil the unfolding mechanism of the protein. The free energy profiles in Fig. 4.1 suggest that the HS plays a relevant role as a mechanical intermediate. Starting from the native state, the β_4 strand unfolds as a first step, to then surmount the free energy barrier of $\sim 7kT$ and reach the fully unfolded state. Physically, this means to break the hydrophobic core, implying thus a big energetic cost.

We highlight two particular unfolding pathways in Fig. 4.2. In the first pathway, the system starts in the N configuration to jump to a state with $\xi \sim 7$ nm, which coincides with the HS state, and then hops to the S configuration, supporting the picture we proposed. Nevertheless, the second transition shows a different mechanism, involving the visit of at least three different intermediates, first one with $\xi \sim 3$ nm (probably the Al configuration) then $\xi \sim 7$ nm (probably the HS configuration) and finally one unidentified at $\xi \sim 13$ nm.

Figure 4.2 suggest a possible multiplicity of unfolding pathways. This feature can never be described with a one-dimensional profile, as the projection onto any reaction coordinate would be non-Markovian. This complexity is of particular interest for the present system, as the pulling force constraints the free energy landscape along

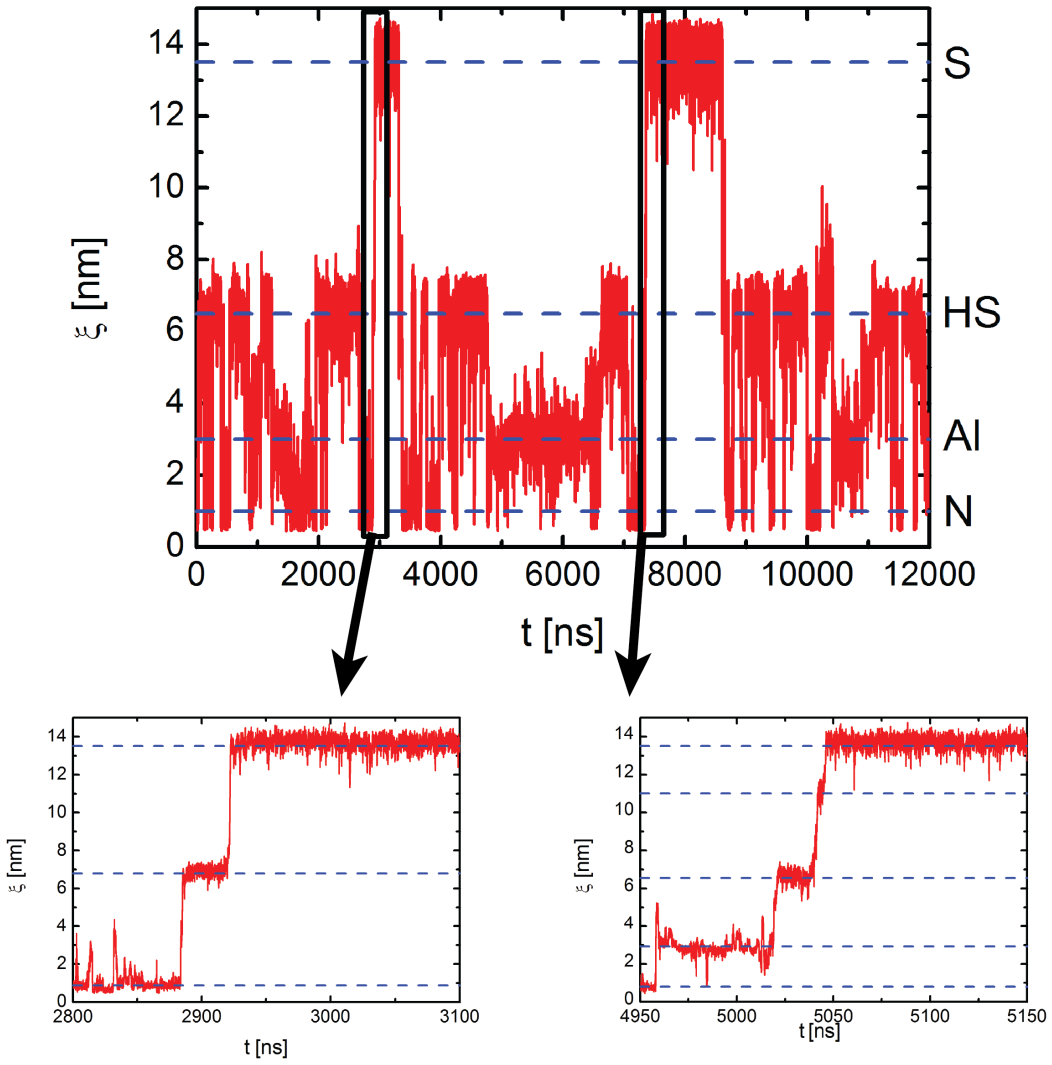


Figure 4.2: **Snapshot of the trajectory along the end-to-end distance with two transition pathways highlighted**: The trajectory shows two main events in the molecular trajectory. The first one involves fast time scales, and is associated with transitions between the N, Al and HS, within times of ~ 1 ns. The second is the unfolding transition, which occurs at slow time scales of $\sim 40 - 50$ µs. Highlight of two different unfolding pathways suggests that unfolding might occur through more than a single reactive pathway, as different intermediates seem to be involved in the process.

ξ, and still the system seems to evolve through other degrees of freedom.

Bayesian Test on ξ and Q

We apply the Bayesian test described in Section 2.3.1 to test the quality of Q and ξ as reaction coordinates. We inspect the trajectories $Q(t)$ and $\xi(t)$ to identify transition pathways, as those fragments of trajectory where the system leaves the native state to go to the stretched state without any recrossing to the native state. The native and stretched states are straightforward to define, just from the histograms showed in Fig. 4.1. A total of nine transition paths are found, covering a fraction of trajectory of $p(TP) = 0.024$.

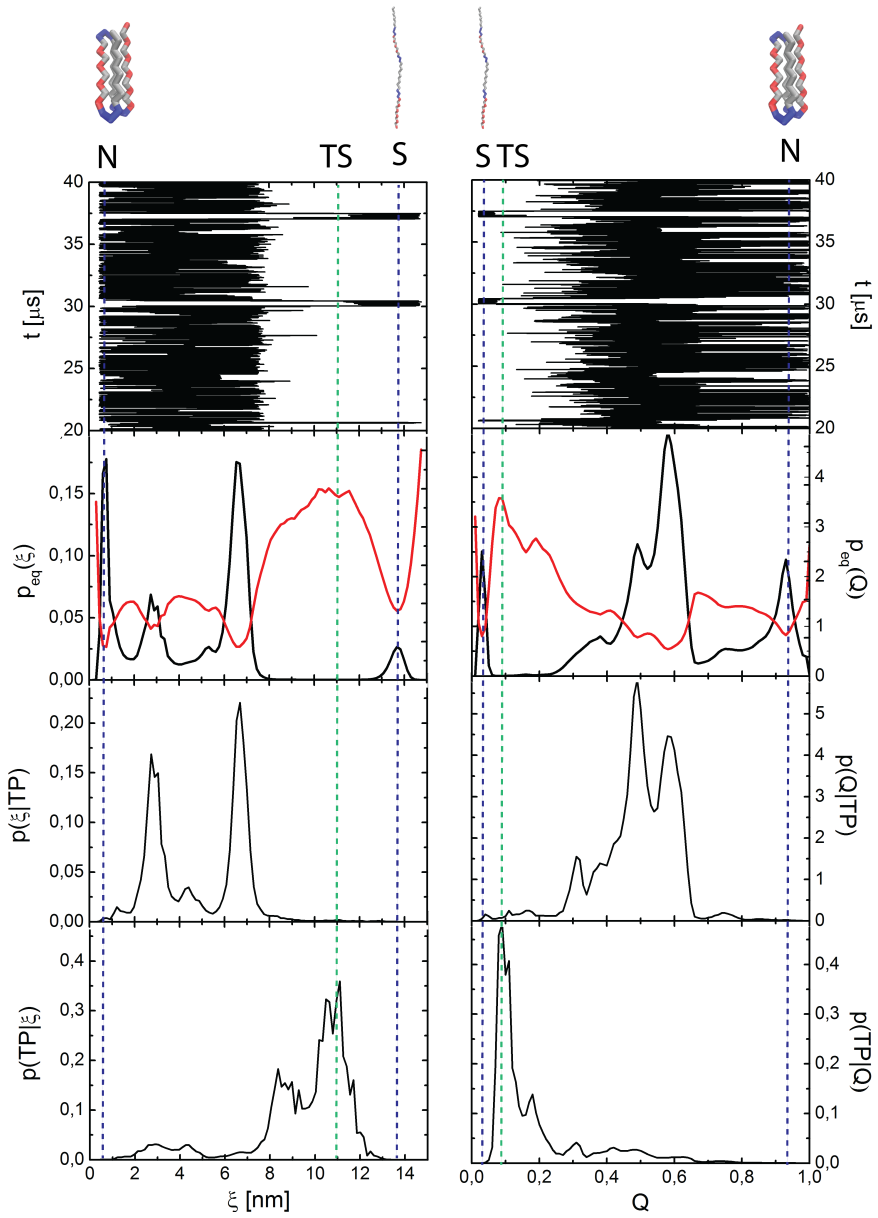


Figure 4.3: **Bayesian criterion to quantify the quality of ξ and Q as reaction coordinates** Left column shows the plots for ξ while the one in the right for Q. From top to bottom we show, a fraction of trajectory involving some transition paths; density probability distribution along the reaction coordinate (black) and free energy profile (red); density probability distribution of the transition pathways; and probability of being on a transition pathway being at a value of the reaction coordinate. Blue dashed lines indicate the N and S state (reactant and product state) and the green dashed line the position of the transition state according to the Bayesian criterion.

Figure 4.3 shows the Bayesian test on Q and ξ in an analogous way as done in [38]. Left panel shows results for ξ and right one for Q. From top to bottom we show a fragment of trajectory with explicit transition paths; the density probability distribution p_{eq} (black) and the free energy profile (red); the density probability distribution along the TPs, $p(q|TP)$ (being q, ξ or Q) and the density probability distribution $p(TP|q)$ as given by Eq. (2.14).

The reactant (N) and product states (S) are identified (blue dashed lines) as free energy wells or p_{eq} maxima. The TS (green dashed line) corresponds to the maximum of $p(TP|\xi)$ and $p(TP|Q)$, appearing at $\xi \approx 11\,\text{nm}$ and $Q \approx 0.1$.

Nevertheless, as mentioned in section 2.3.1, the indicative of a good reaction coordinate q is a unimodal distribution $p(TP|q)$ with a sharp single peak at q_{TS} close to the diffusive limit $p(TP|q) \approx 0.5$. Otherwise, this implies the existence of multiple TSs or overlapping of states at the same value of the reaction coordinate.

Figure 4.3 shows a multimodal representation for $p(TP|\xi)$ with at least three different peaks where the one at $\xi \approx 11\,\text{nm}$ stands out, close to the diffusive limit. This plot suggests that ξ does not provides a proper description of the system, even though is the natural reaction coordinate. The multiple peaked structure implies the presence of possible multi-transition states, and thus of multiple unfolding pathways.

The distribution $p(TP|Q)$ shows a single sharp peak at $Q = 0.1$, close to the diffusive limit 0.5. In this sense, the test ratifies it as a good reaction coordinate. The TS is a structure with just the 10% of the native contacts. This is quite close to the S configuration, so the test fails in describing how does the system reaches the TS, as we can have overlapped pathways in the transition from N to the TS. The most surprising feature is the the HS configuration seems to play no role in the unfolding mechanism, in opposition from the intuition given by the free energy profiles.

4.3.2 PCA as a Method to Find Order Parameters

In this section we employ PCA for two different purposes. First, the eigenvectors contain information about which regions of the molecule have more important contributions to the fluctuations. Second, PCA is a way to find a new set of coordinates which can offer useful order parameters for describing the system. We build the free energy profile along the first few PCs and compare to the ones in Section 4.3.1. We do not refer to the PCs as reaction coordinates, given that no reaction occurs along them. The term order parameter is more appropriate as they are useful collective variables for monitoring the state of the system.

The i-th PC is defined as:

$$q_i(t) = \mathbf{v}_i \cdot (\mathbf{r}(t) - \langle \mathbf{r} \rangle). \tag{4.1}$$

Here, subindex i stands for the PC eigenspace index, and PC eigenvector $\mathbf{v}_i$ is a 3N component vector (where $N = 46$, the number of residues in our protein). $\mathbf{r}(t)$ stands for the the 3N component time dependent trajectory of the protein, $\mathbf{r}(t) = (x^{(1)}, y^{(1)}, z^{(1)}, x^{(2)}, \cdots)$, and $\langle \mathbf{r} \rangle$ is the average structure of the protein calculated along the trajectory.

Study of the Eigenvalue Distribution and Eigenvectors

We study the PC eigenvalues and eigenvectors to gain the information they contain. As mentioned in Chapter 2, the eigenvalues are the autocovariances of the PCs $\lambda_i = \sigma^2$, which we order in decreasing value. Observation of the cumulant autocovariance along the eigenvalue index provides a dimension reduction cutoff criterion. Typically, keeping between the $80 - 90\%$ of the total autocovariance is enough for a proper description of the system. We define the cumulant autocovariance ζ as,

$$\zeta_i = \frac{\sum_j^i \lambda_j}{\sum_i^N \lambda_i}, \tag{4.2}$$

where N is the number of coordinates of the system. Panel (A) in Fig. 4.4 shows the cumulant autocovariance as a function of the eigenvalue index i. The blue dashed line shows that for three PCs we gather up to the 80% of the total autocovariance. The three first eigenspaces are responsible for the majority of the fluctuations of the systemm and the first PC already gathers more than 50% of the total.

For a more intuitive visualization, it is useful to calculate the contributions of each monomer to the autocovariance, rather than of each coordinate. We define the following magnitude:

$$\delta_i^j = (v_i^j)^2 + (v_i^{j+1})^2 + (v_i^{j+2})^2, \tag{4.3}$$

where v_i^j is the j-th component of the i-th eigenvector. In this sense, δ_1^j accounts for the magnitude of the contribution of monomer j to the first PC, and so on. In our particular case, keeping just the z component should be enough, as most of the motion occurs through this coordinate, given the direction of the force.

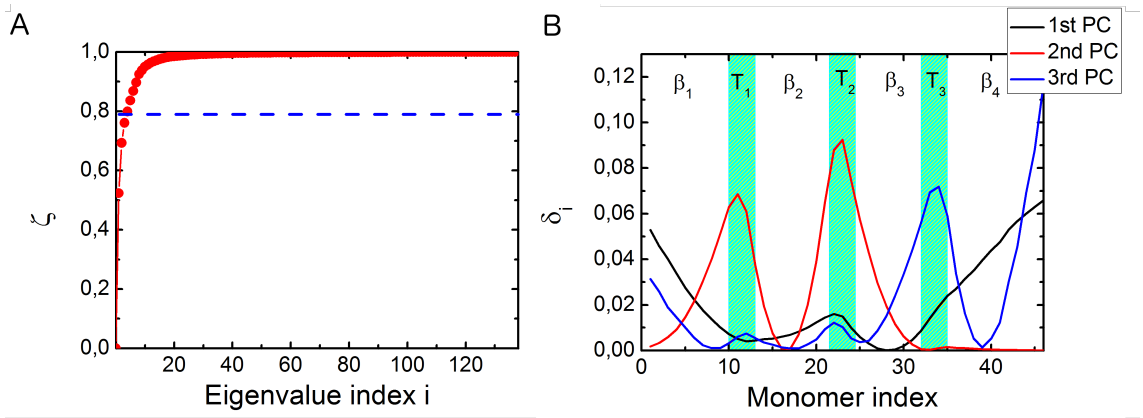


Figure 4.4: **Cumulant autocovariance and eigenvector representation**: Panel (A): Cumulant autocovariance ζ as a function of the eigenvalue index. Largest part of the autocovariance accumulates into the first few eigenvalues, with the first three ones gathering up to the 80% of the total. Panel (B): δ_i quantities computed, where the regions which account for the largest part of the fluctuations can be clearly seen. First component is due to motion in the extremal β_4 strand, while second and third due to movements in the neutral turns.

Panel (B) in Fig. 4.4 shows the quantities δ_i for the first three PC, revealing in which regions of the protein do fluctuations concentrate. First PC accounts for motions in the extremal β strands, mainly in the β_4 strand, the one that unfolds in the HS configuration. A description through the first PC should be enough for differencing between the N, HS and S states, as they rely on large configurational changes associated with the extremal β strands.

The contribution of the second PC accounts for movements in the neutral turns, particularly in the T_1 and T_2. Interestingly, this component is separating the motion associated with unfolding of the β strands, giving more subtle fluctuations which are associated with arrangements in the neutral turns. Finally, the third component focuses on the third neutral turn and part of the β_4 strand. This accounts for possible flexibility in this strand (as there is a minimum around monomer 40) and the third T_3 turn.

Free Energy Profiles Along the Principal Components

We compute the PCs q_i and study their suitability as order parameters. First, they can be useful to understand the unfolding mechanism of protein BPN_{46} and compare the conclusions with those derived from the reaction coordinate candidates ξ and Q. Second, they can be used as input parameters for the discretization process prior to a Markov state model building (see Chapter 5).

Studying the probability density—or the free energy profile—along each PC reveals states associated with large amplitude motion or fluctuations from the average structure. Typically, the first PCs would have structured distributions, with multiple peaks revealing different states. The last PCs become gaussian distributions about the average structure, accounting just for the symmetric thermal fluctuations, so are to be discarded.

The interpretation of the PCs is non-intuitive, and they should be analyzed jointly with the average structure and the PC eigenvectors. There are the following three options for the overall behavior:

1. **Null values of the PC**: Peaks centered at $q_i = 0$ indicate two possibilities. The first one is no motion about the average structure, indicating the presence of the average structure itself. For example, the last PCs are gaussians distributions about $q_i = 0$, indicating thermal fluctuations about the average structure. The second option is to have motions orthogonal to the associated eigenvector $\mathbf{v}_i$. The former option is the most likely, as latter structures are typically eliminated in the aligning process.

2. **Positive values of the PC**: These states indicate presence of motions which are symmetric to the ones described by the eigenvector.

3. **Negative values of the PC**: These states indicate presence of motions which are asymmetric to the ones described by the eigenvector.

Figure 4.5 shows the PMF along the first two PCs, with the associated probability density distributions in the insets. As predicted earlier, the first PC gives a general idea about the behavior of the system, in agreement with the profiles shown in Fig. 4.1. The second component gives a more detailed picture of the system, accounting for motions associated with the turns.

The free energy profile along q_1 shows three free energy wells. The first one has a large negative value ($q_1 \approx -50$), the second one $q_1 \approx -10$ while the third one $q_1 \approx +20$. This profile resembles the one along the end-to-end distance, and these three states correspond to the S, HS and N configurations, respectively. This is concluded by two reasons. First is the mere observation of the population of the three states (see Inset in Fig. 4.5). Second, the first PC accounts for motions in the extremal β_4 strand, and the conformational changes among these three states are associated with large changes in this strand. The average conformation is close to the HS state, so free energy well at $q_1 \approx -10$ corresponds to the HS conformation. State at $q_1 \approx -50$ is a large negative motion, so likely the S conformation. Finally, state at $q_1 \approx +20$ corresponds to the N state, given an opposite sign motion of the strands. Observation of the trajectory $q_1(t)$ compared to $Q(t)$ or $\xi(t)$ certify these conclusions.

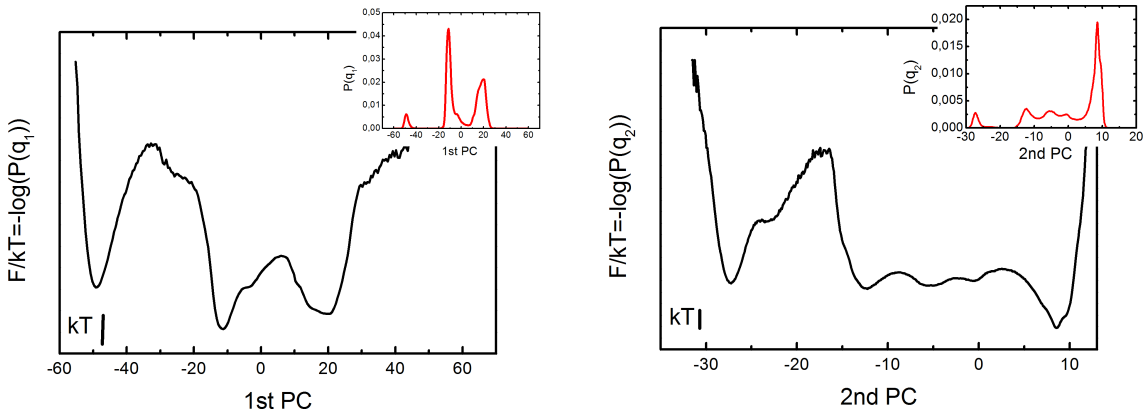


Figure 4.5: **Free energy profiles along the first and second PC**: Free energy profile along q_1 and q_2 and probability distribution (inset). The profile along the first PC reveals the three major states already identified, while the one along the second a rugged landscape with probably more states not found in the representations along ξ and Q.

The free energy profile along q_2 is harder to interpret, but the free energy wells should be associated with motions in the turns. Several free energy wells appear, separated by low barriers, revealing a rugged landscape. This suggest that the multiplicity of states could be larger than that found in the simple one-dimensional profiles.

4.4 Two Dimensional Free Energy Landscapes

We show two dimensional free energy surfaces along different pairs of order parameters, namely ξ and Q, and the first two PCs. While one-dimensional projections showed the basic features of the system, some states can be overlapped in this simple projection. Two dimensional surfaces are useful for resolving possible overlappings and providing a more complete vision of the free energy landscape.

Figure 4.6 shows the two-dimensional free energy surfaces along the fraction of native contacts Q and the end-to-end distance ξ. The three major states N, HS and S are identified as three free energy wells in the surface. The Al state appears as another differentiated well, well resolved in with the ξ coordinate but not with Q. Also, a different state with $\xi \sim 5\,\text{nm}$ and $Q \sim 0.4$ is revealed. Being a two dimensional surface, it offers a greater variability of pathways. While one dimensional descriptions suggest HS as the natural intermediate—which is a plausible option, given that one of the strands is unfolded—other low free energy routes are possible here, for example following the Al configuration.

Figure 4.7 plots the free energy surface along the first two PCs. It is clear how the PCs are better reaction coordinates than Q and ξ as up to eight distinct states can be identified, as eight valleys in the free energy landscape. In an overall view, the surface shows three major regions, in correspondence with the N, HS and S states. Nevertheless, they have a rugged structure, with multiple free energy wells inside. For example, the native region is divided in three different wells. Moreover a set of two wells connects this native region with the stretched one, suggesting a possible pathway not seen in the one dimensional profiles.

The dashed lines suggest two possible unfolding pathways. First, the system would transit to the HS configuration to unfold completely later. Nevertheless,

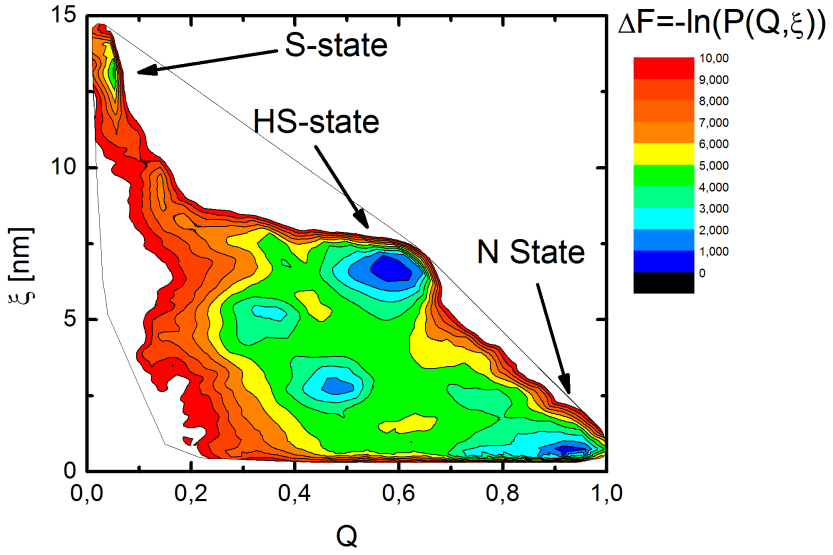


Figure 4.6: **Two-dimensional PMF along Q and ξ**: The three main states are clearly identified (N, HS and S) as three wells in the bidimensional landscape. The Al configuration is also found as a separate valley. Apparently, no new state state is recovered in this representation, meaning that the one dimensional projections along these two respective coordinates were not missing any information.

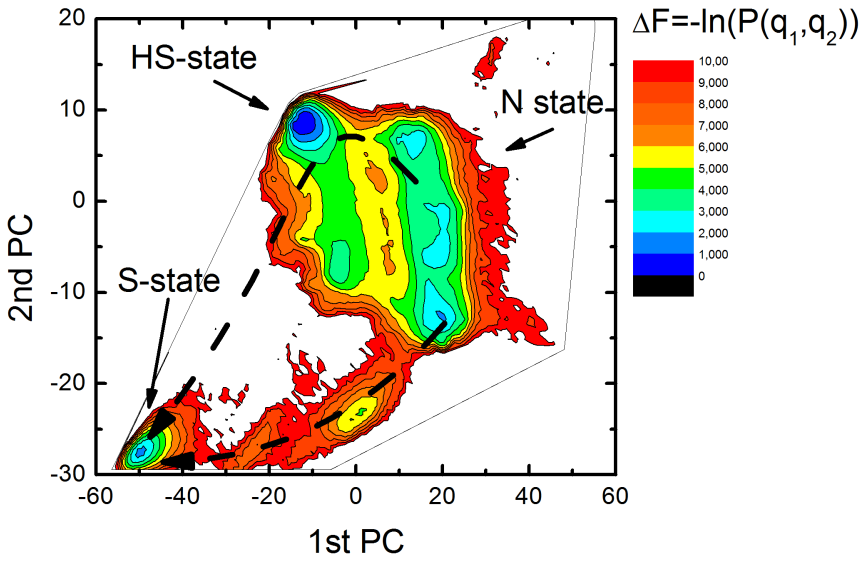


Figure 4.7: **Two-dimensional PMF along the first two PCs (q_1 and q_2)**: The N, HS and S configurations appear as three clear valleys in the landscape. Nevertheless, a number of additional shallower wells suggest the presence of further configurations. Two possible unfolding pathways are suggested. The first pathways crosses the HS region (as suggested by the one dimensional profiles), following a high free-energy route. The second pathway moves directly from the native states through a set of new states until it unfolds.

this pathway is forced to go through high free energy regions, so it should have low probability. The second pathway seems more straightforward. The system

would move directly from the native state to the stretched one by crossing two new configurations connecting them. These regions have a higher population and lower barriers, and thus, this pathway seems more likely. Remarkably, this is hard to find in the $Q - \xi$ representation.

Chapter 5

Mechanical Unfolding of BPN_{46}: Markov State Model and Unfolding Pathways

In this chapter we employ a Markov state model description of the free energy landscape of the system. In opposition to the approach of Chapter 4, this requires no any prior knowledge about the system, or arbitrary definition. We proide a clear picture of the configurational space of the system, allowing to identify all relevant macrostates as free energy basins. Additionally, we apply TPT to unveil the unfolding pathways. Finally we compare the results from both descriptions.

5.1 The Markov State Model of the System

5.1.1 Markov State Model Construction

We build a Markov State Model following the method described previously. PCA is used in order to reduce the dimension of the system. We take the first three PCs as our configurational space (keeping the 80% of the total autocovariance), which we discretize to build the microstate network. We discretize each PC in 20 bins, so the our configurational space is made of 8000 possible microstates.

The conformational Markov network is built by mapping the molecular dynamics trajectories onto a complex network discretizing the trajectory according to the defined bins. The conformational Markov network we obtain is made up of 1867 bins (23% of the possible microstates) which are connected through 23995 links, including self-links. This is less than the 1% of the possible links. This is no cause of surprise or concern for possible misconvergence of the network or lack of enough statistics. Typically, the number of links of a converged network go with $N \ln N$ [12], which is far satisfied here.

In order to build a more intuitive model of our system, we cluster the microstates (nodes) onto macrostates by applying the Stochastic Steepest Descent algorithm [32]. From this clustering criterion, we define the macrostate network, where each new node represents the basins of attraction of the underlying free energy landscape of the system. This smaller network is more significative from a physical point of view and allows us to relate each node with conformations adopted by the system through the simulation. We take the basin network as the equilibrium ensemble of

our system.

The basin network is made up of 30 basin connected through 1290 edges. Those basins with a population lower than the 0.001% of the trajectory ($\pi < 10^{-5}$) are subtracted from the network in order to avoid pathological or extremely rare states. After this refinements, we keep 13 macrostates connected through 65 edges, including auto links.

5.1.2 Study of the Eigenvalues and Eigenvectors of the Transition Matrix

As discussed in Chapter 2, the transition matrix is a discrete version of the transfer operator of our dynamical model. Thus, the study of its eigespectrum provides useful information about the kinetic processes of the system. We address directly the spectrum of the coarse-grained transition matrix, considering that the state space is composed of 13 macrostates. In such way we already rule out any fast process (which are intra-basins transitions).

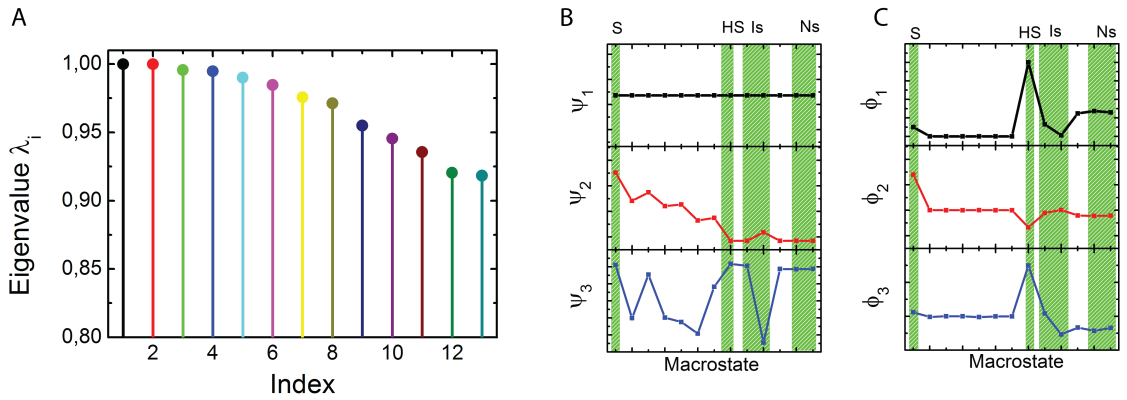


Figure 5.1: **Eigenspectrum of the transition probability matrix and first three eigenvectors** : Panel (A) shows the 13 eigenvalues for the coarse-grained transition probability matrix. Largest eigenvalue is $\lambda_1 = 1$, associated with the equilibrium distribution. Panel (B) and (C) shows first three eigenvectors ψ_i and ϕ_i. ϕ_1 shows the stationary distribution and the next ones the following two slowest dynamical processes, associated with the transition to the stretch process and transition to the half-stretched state.

Panel (A) in Fig. 5.1 shows the eigenvalue spectrum for the 13×13 transition matrix. All 13 dynamical processes are slow processes (high eigenvalues), where first eigenvalue $\lambda_1 = 1$ corresponds to the equilibrium distribution. Panel (B) shows the eigenfunctions ψ_i of transition matrix $\tilde{T}$ while panel (C) shows the eigenfuctions weighted with the equilibrium distribution ϕ_i (see Chapter 2). We focus on the three slowest dynamical processes.

The first eigenfunction recovers the equilibrium distribution (constant in the ψ_1 representation and equal to π in the ϕ_1 representation). We highlight four groups of states which are clearly identified in next sections. State with index 1 corresponds to the stretched state S shown in the one-dimensional descriptions. State 8 is the half-stretched conformation HS while 12 and 13 correspond to native N conformations. States 9 and 10 are intermediates Is, to be further discussed later.

Second eigenfunction accounts for the transition to the stretched state, which is the slowest dynamical process, apart from the equilibrium relaxation. This agrees

with the large barrier seen in the one-dimensional profiles. ϕ_2 is associated with a dynamical transition from HS state to S state, according to the observed signs. This does not means that state HS jumps directly to state S, as this is weighted with the equilibrium distribution, and the HS state has a large population. Actually, ϕ_2 shows subsequent excitation of the intermediate low-populated states $2-7$.

The third eigenfunction accounts for transitions from the intermediate Is and native states Ns to the HS configuration. It is associated with transitions from native-like states to the HS basin, and thus to the second slowest transition, as discussed in Chapter 4.

In this way, we see how the transition matrix is capturing the two main processes in which the system is involved, transition to the stretched state, and transitions from native-like configurations to the half-stretched configuration. This agrees with our first vision of the system, as understood from the trajectories along the reaction coordinates.

5.1.3 Description of the Markov State Model: Topology, Macrostates and Involved Transitions

Figure 5.2 (upper) shows a pictured vision of the basin network with a significative structure represented by each node (lower panel). The size of each bead (node) is proportional to its occupation π_i. The spatial arrangement of the nodes and links has been calculated by applying a *Force Atlas* algorithm [85], where an artificial dynamics is simulated in order to relax the network to an equilibrium arrangement. Each link is considered to be a linear spring, while a certain repulsion is set between nodes. The system is left to interact until an equilibrium configuration is reached. Nodes which are *kinetically* close would appear near each other in the final arrangement, while nodes which are far away (meaning that transitions between them are unlikely), would appear separated in space. A modularity algorithm [86][1] is applied in order to rank the nodes according to their modularity class. This is indicated in their different colors, associated to five different modularity classes.

Lower panel shows a representative structure encoded in each of the nodes of the network. Configurations N_1 and N_2 correspond to native-like structures. Despite their structural similarity, they were not distinguished in the Q or ξ landscapes. Interestingly they play a rather different role in the kinetic behavior of the system, as we discuss later on. Particularly, macrostate N_1 resembles more the actual native structure, while N_2 shows a higher flexibility in the T_1 and T_3 turns of the structure, as it can be derived from the structures shown in Fig. 5.2. Both represent the native ensemble of our conformational space.

Basin HS represents the half-stretched configuration. It is the most populated state in the network, as could already be expected. The aligned configuration determines another populated basin, which belongs to the same modularity class as native state N_2 but not to N_1, meaning that it is kinetically closer to the former. The stretched configuration S is far topologically from the native region. Remarkably, 8 intermediates, something not found when using Q or ξ as reaction coordinates.

[1]Modularity algorithms are a popular way to define communities in complex networks. The idea is to perform a partition on a network measuring the density of links inside communities compared to links between communities. Starting from some random choice, this quantity is optimized until the best partition is found

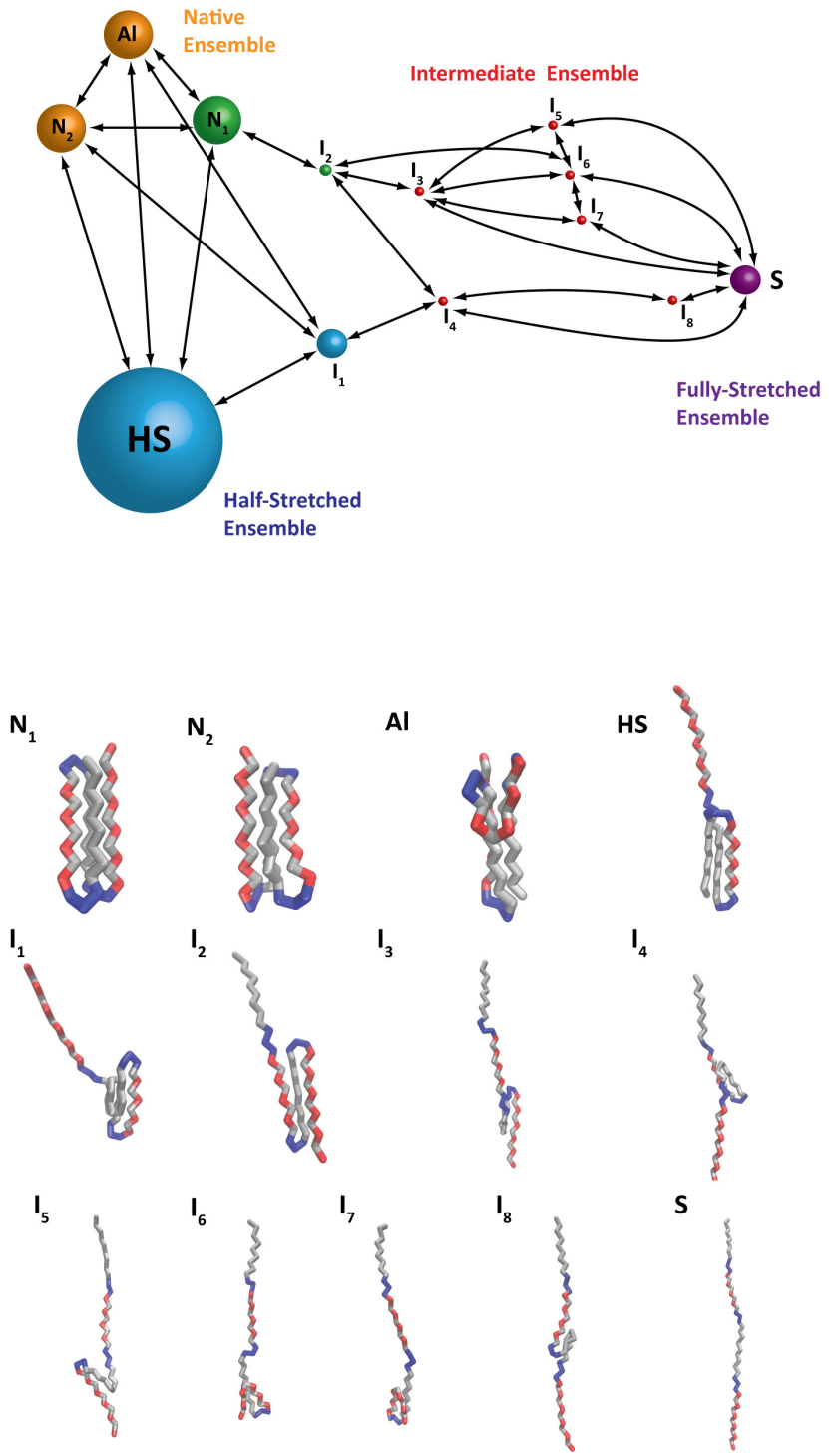


Figure 5.2: **Basin Network and associated structures**: Upper panel shows a representation of the 13 basins with $\pi > 10^{-5}$. The size of the nodes is proportional to its population π_i. Bidirectional arrows indicate allowed transitions, but the magnitude of T_{ij} is not shown. Self-loops are not drawn, being always present. Each basin is labelled according to the structure they encode. Lower panel shows a representative structure for each of the basins. We identify two different native configurations N_1 and N_2, the half-stretched configuration HS and the aligned configuration Al, the stretched structure S and a total of eight low populated intermediates $I_1 - I_8$.

This can be due to its low population or because they were overlapped in different regions of the profile. These intermediates seem to play a relevant role in the

unfolding mechanism as they connect the Native Ensemble with the Stretched one.

Table 5.1: **Description of the basins of attraction**. Characterization of each of the macrostates of the system in terms of their population π_i, mean escape time $\langle t_s \rangle$, mean fraction of native contacts $\langle Q \rangle$, mean end-to-end distance $\langle \xi \rangle$, mean fraction of *non-native* contacts $\langle f_{NN}$ and committor probabilities from the native to the stretched ensemble q_i^+.

#	π_i	$\langle t_s \rangle$	$\langle Q \rangle$	$\langle \xi \rangle (nm)$	$\langle f_{NN} \rangle$	q_i^+
N_1	0.15	559	0.85	0.8	0.13	0.0
N_2	0.14	495	0.83	0.9	0.30	0.0
Al	0.14	272	0.40	2.6	0.60	1.4×10^{-4}
HS	0.44	2982	0.46	6.5	0.18	9.2×10^{-4}
I_1	0.07	362	0.25	4.8	0.66	1.2×10^{-3}
I_2	0.01	2586	0.35	6.8	0.40	0.12
I_3	6.67×10^{-5}	120	0.12	9.0	0.23	0.29
I_4	1.3×10^{-4}	198	0.11	10.1	0.54	0.34
I_5	1.9×10^{-5}	64	0.10	9.6	0.60	0.51
I_6	3.9×10^{-4}	163	0.14	8.55	0.30	0.53
I_7	3.3×10^{-4}	176	0.13	9.35	0.50	0.58
I_8	2.5×10^{-5}	56	0.09	10.5	0.70	0.71
S	0.06	75000	0.01	13.7	0.00	1

Table 10.1 shows different characteristics about each of the 13 macrostates of the equilibrium ensemble of the system. π_i stands for the population of each basin, $\langle t_s \rangle$ is the mean escape time, $\langle Q \rangle$ the mean fraction of native contacts, $\langle \xi \rangle$ the mean end-to-end distance, $\langle f_{NN} \rangle$ the mean fraction of *non-native* contacts and q_i^+ the forward committor probabilities from the native ensemble (N_1 and N_2) to the stretched ensemble S. We have introduced here the magnitude f_{NN}, which checks how many of the contacts in each configuration are non-native, this is, do not appear in the native contact map.

$\langle Q \rangle$, $\langle \xi \rangle$ and $\langle f_{NN} \rangle$ are the average values calculated from the marginal distributions from the configurations adopted in each basin. In most cases, these average values are not enough to characterize each state, as the marginal distributions are not unimodal. The committor probabilities have been calculated by defining the initial state as the native one and the final as the stretched one. This quantity coincides with the probability of unfolding p_{unf}, and the backwards commitor probability $q_i^- = 1 - q_i^+$ is the probability of folding p_{fold}. Both quantities are necessary when applying TPT in next section. The magnitude f_{NN} is relevant as the model is allows non-native contacts to form. Some structures might be stabilized by the creation of interactions which did not appear in the native structure.

The native ensemble is a good example of this. Despite their overall structural similarity, they play a rather different role in the kinetic behavior of the system. In addition, they show a similar value of $\langle Q \rangle$ but state N_2 is stabilized by a larger number of non-native contacts. We can say that N_1 is "more native" than N_2, as seen in the arrangement of the neutral turns. Even though they display a rather similar stability (similar weights and escape times), N_1 is surprisingly closer to the Intermediate states. Indeed, it belongs to the same modularity class as I_2,

while N_2 is kinetically closer to the HS configuration. If we compare the transition times computed from the transition matrix T_{ij}, ($\tau_{N_2 \to HS} = 557ps$ and, $\tau_{N_1 \to HS} = 13.5 \times 10^6 ps$), state N_1 is scarcely connected to the HS configurations, while N_2 has a high probability of jumping to this state. Additionally, both native states are separated by large barriers, as $\tau_{N_2 \to N_1} = 14 \times 10^3 ps$ and $\tau_{N_1 \to N_2} = 15 \times 10^3 ps$. Despite their structural similarity—which makes them undistinguishable from each other by any structural based reaction coordinate—their role in the system dynamic behavioris very different.

This is a first contradiction with respect to conclusions yielded by the one-dimenional descriptions. While the main features of the equilibrium ensemble of the system are correctly insinuated by both methods (the three major states, N, HS, and S, the fast kinetics $N \leftrightarrow HS$ and the slow kinetics $N \leftrightarrow S$), the role of such states and the presence of additional relevant configuration is hidden in the one-dimensional projections.

5.2 The Unfolding Pathways: Transition Path Theory

Markov state models allow to computethe pathways connecting two subsets of the network in a rather straightforward way. This is possible thanks to TPT [10, 61, 62]. Particularly, we employ this method to unveil the actual unfolding mechanism of our protein under the effect of a mechanical force.

We define the "reactant" subset A as the native ensemble, made up of basins N_1 and N_2 together, while the "product" subset B is determined by basin S. According to this definition, we calculate the forward committor probabilities q_i^+, already shown in Table 10.1. From an intuitive point of view, q_i^+ can be interpreted as the probability, being in state i, to unfold (moving to subset B) without folding back (going back to subset A). By definition, q_i^+ is 0 for states N_1 and N_2 and 1 for state S. Then, the states can be sorted accordingly to "how close" are they kinetically to the unfolded state.

From the committor probabilities, the effective and net fluxes are calculated and the flux network built, as explained in the methods section. Figure 5.3 shows the net flux network, where the thickness of the arrows is proportional to the net flux f_{ij}^+. The total unfolding flux is $F = 2.9 \times 10^{-7}$ per lag time τ, meaning that we expect an unfolding transition every 51.7 µs. Accordingly the rate constant is $k_{NS} = 2.1 \times 10^{-8} \mathrm{ps}^{-1}$, which means that, being in the native state, the system would jump to the stretched state in an average time of 47.6 µs. The conclusions yielded by both quantities are similar given that the probability to be close to the N state is very high $\sum_{i=1}^{N} \pi_i q_i^- \approx 1$.

The flux network can be decomposed into the individual unfolding pathways. In order to do so, the strongest pathway are identified and removed from the network. This process is repeated until no path connects subsets A and B. In principle, for large networks, this process is not trivial, nevertheless in our case it can be done by hand, due to the limited size of our network.

We identify a total of 9 pathways leading from A to B. After decomposing the network into these 9 paths, unconnected regions can be still observed. These regions remain due to the presence of *trap states*, which in our case carry near the 20% of

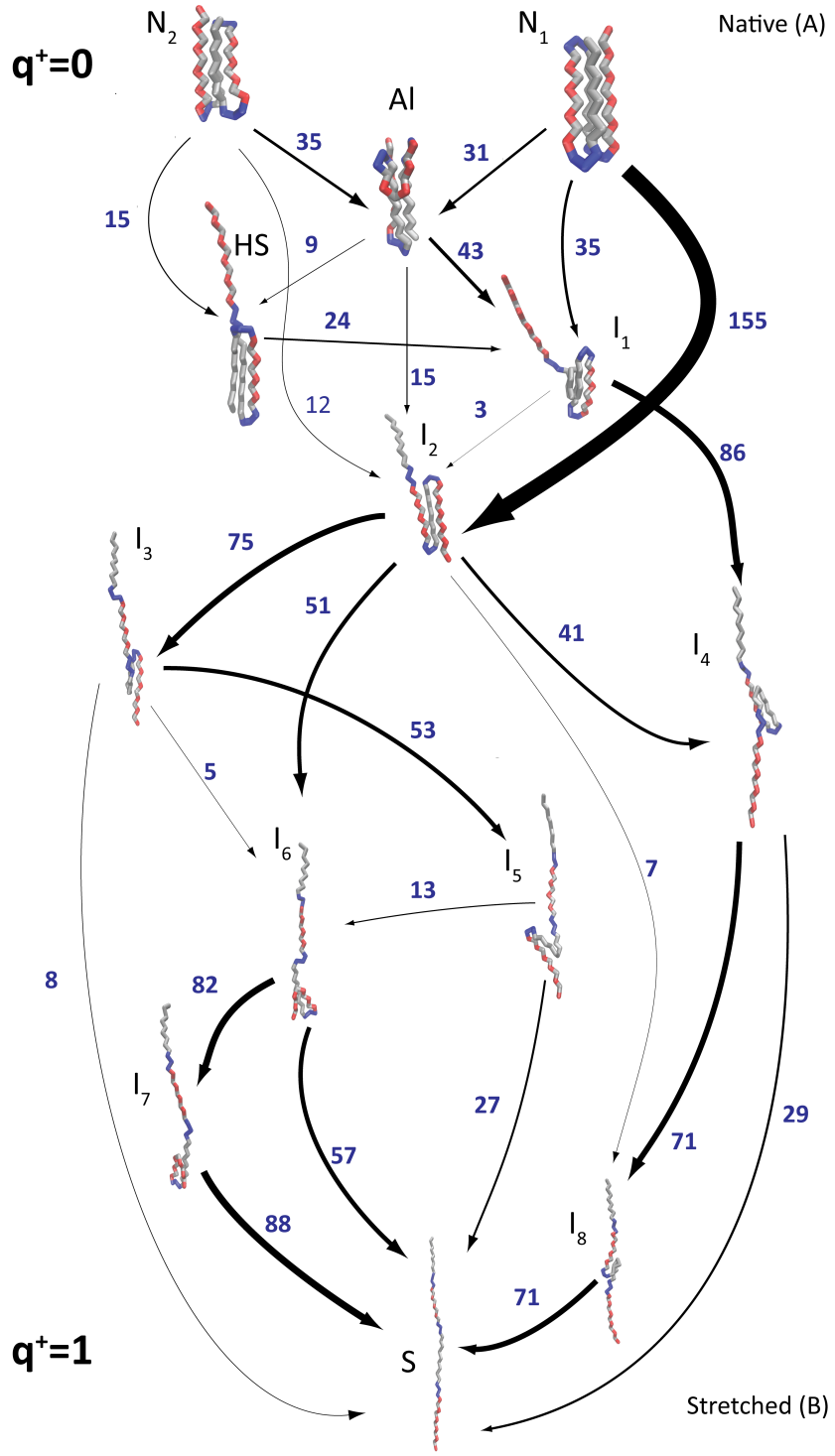


Figure 5.3: **Unfolding flux for the model protein**: The 13 configurations are arranged vertically according to the value of the committor probabilites (not in scale). The arrows connecting configurations represent the unfolding net flux, where the thickness of the arrows is proportional to its magnitude. Numbers next to the arrows show the net flux in units of $10^{-9}ps^{-1}$.

the total flux. We show in Fig. 5.4 the 6 more relevant pathways, carrying the 89% of the total *unfolding* flux.

From these 9 pathways, 7 start from conformation N_1, while just 2 from N_2.

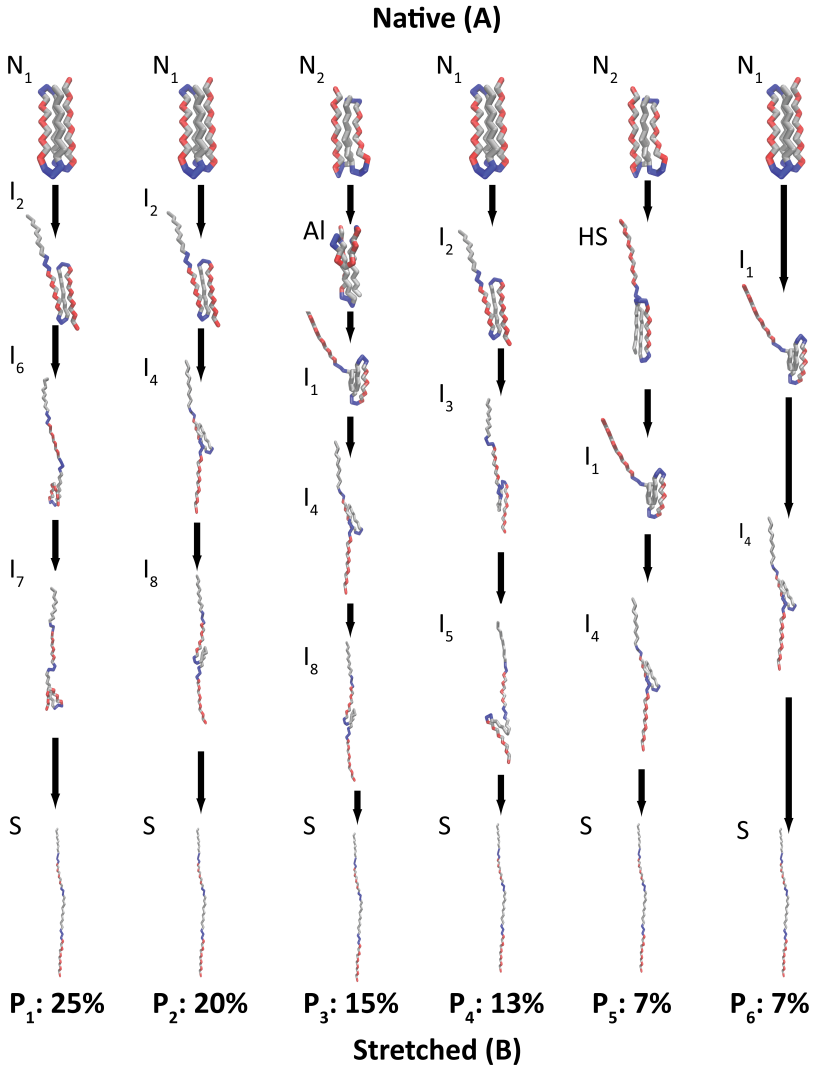


Figure 5.4: **Six main unfolding pathways**: The six pathways carrying a larger fraction of the total flux (up to a 89%) are shown here.

This is a remarkable fact, as N_1 is closer to the native state, as discussed previously. In addition, states I_1 and I_2 appear as the actual intermediates of the unfolding mechanism, as by removing both states, the transition $A \to B$ becomes forbidden. Out of the 9 pathways, 6 go through I_2 and 3 through I_1, defining two major unfolding routes, one driven through state I_1 and other through state I_2.

5.3 Discussion

Through the present and previous chapter, we have analyzed and discussed the behavior of protein BPN_{46} under the presence of a mechanical bias. In particular, we were interested in unveiling the unfolding mechanism. In order to do so, we have employed two complementary approaches. First, a low dimensional representation of the free energy landscape, the free energy profiles along two different reaction coordinates and two-dimensional free energy surfaces. Second, a Markov state model, which allowed us to obtained a detailed description of the configurational space of the system, and to calculate the unfolding pathways. We discuss now the vision

that each of the methods provided us, the common points, and the differences.

Finding reaction coordinates that allow to describe faithfully complex dynamical process, is an appealing problem for a number of reasons. Currently, molecular dynamics simulations produce a large amount of data, whose magnitude itself becomes a main difficulty when it comes to understand the process subject to study. Hence, being able to describe the system through a one-dimensional description is a natural worthwhile question. Ultimately, one could be able to represent a complex system as a diffusive (overdamped Langevin) process along this reaction coordinate [41]. Since reaction coordinates constitute often the slow variables of a process, this dynamical coarse-graining construction determines a simple description for rather a complex system. Additionally, a free energy profile along a good reaction coordinate, not just resolves the reactant and product state, but also the position of the transition state and the height of the free energy barrier, and so the kinetics of the reaction [38, 39].

In the present case, we have studied the *mechanical unfolding* of a protein. Hence, the force sets a natural reaction coordinate through which the system evolves, the pulling direction. The actual multi dimensional free energy landscape of the system is *collapsed* onto this *slow* variable, as the additional degrees of freedom relax faster, and contain little relevant information. This assumption is also taken in single molecule force spectroscopy experiments (see Part III), as the pulling direction is usually the only available observable. In this sense, it is useful to think in this terms, in order to bridge the gap between experiments and molecular simulations.

The analysis of the low dimensional representations yielded some important conclusions. From the point of view of the identified states, the unfolding transition is not a simple two state problem, as at least four relevant free energy minima were identified, the native, and stretched states, and the two intermediates, the aligned and half-stretched configurations. The latter one corresponds to be the most stable configuration under the set conditions, as the extremal β strands were aligned in the direction of the force and the hydrophobic core left intact. Kinetically, we observed two different involved time scales, first fast transitions between the N, Al and HS configurations, at a time scale of $\sim 100\,\text{ns}$, and then a slow time scale associated to transitions to the S state, within times of few tenths of µs.

This latter observation is backed up by observing the free energy barriers, as the first states where separated by low barriers of $\sim 2kT$, while the stretched state needed to surmount barriers of $\sim 7 - 8kT$. Importantly, the HS configuration defines as a likely mechanical intermediate. Nevertheless, a more careful analysis of the trajectories, and the application of the Bayesian criterion suggested that surprisingly ξ was not an appropriate reaction coordinate, and that Q was a more appropriate one, with the TS located at $Q \sim 0.1$, very close to the actual unfolded structure.

In this sense, description of the system through the two reaction coordinate yields a correct overall description of the states of the system, and also the two involved kinetic scales. Nevertheless, it fails in giving an appropriate vision of the unfolding mechanism, which occurs through more than a single reaction pathway.

The Markov state model description provides a correct vision of the configurational space of the system and of the unfolding mechanism. Regarding the equilibrium ensemble, several features can be pointed out. In a coarse vision, the three major states identified previously appear as well differentiated nodes, or set of nodes.

Nevertheless, the actual conformational space has more free energy basins which play different kinetic roles.

This is the case of the native ensemble. We identify two different native basins, N_1 and N_2 with very similar values of ξ and Q, and thus overlapped in the free energy profile. Nevertheless, their role is rather different. N_2 is very connected to the HS configuration and thus is responsible of the fast kinetic transitions between these two macrostates. N_1 and N_2, despite their structural similarity, are indeed separated by a large free energy barrier.

The HS configuration is ratified as the most populated configuration. Nevertheless its role in the unfolding mechanism is rather marginal. It appears just in one of the found unfolding pathways, path P_5, carrying the 7% of the total unfolding flux. Being a quite stable conformation, it rarely unfolds completely, as the hydrophobic core of the molecule is maintained.

The actual mechanical intermediates are states I_1 and I_2. They are hard to identify in the representations along coordinates ξ or Q due to their low population, which disguised them into the background. Additionally, I_2 has similar ξ value to HS and Al configuration, so both states would overlap in the one dimensional description. Structurally, this state is symmetric to HS, as the β_1 strand is unfolded instead of β_4. From the point of view of the stability of the system this is a big difference, as the hydrophobic core is broken, which makes the protein unstable, leading the unfolding mechanism.

The other major unfolding route includes state I_1, similar to the HS state, as both have the β_4 strand unfolded. Nevertheless, in I_1 the core adopts a compact globular structure that is sustained by a large fraction of non-native interactions, specially between the hydrophobic residues of strands β_1 and β_3. This configuration drives a second unfolding route through states I_4 and I_8, also sustained by a large fraction of non-native contacts. The possibility of forming non-native contacts is responsible also of structure Al, which has a relevant stability ($\pi_i = 0.14$), and a 60% of non-native interactions. This structure has a certain role in the unfolding mechanism, allowing to reach intermediate I_1 from native state N_2), but participates mainly in the fast dynamics between HS and the native set, as can be also directly observed in Fig. 4.2.

Finally, we can try to connect the description with free energy profiles with the Markov state model one. Markov state models provide an appropriate description of the configurational space of a molecular system. Reaction coordinates do not intend to do so, but rather to provide a direct comprehension about a reaction mechanism, by appropriately picturing the reaction pathway, the position of the transition state and the free energy barriers. Markov state models usually fail on this latter point, as it is hard to locate the transition state within the network description or to obtain free energy barriers. We were able, nevertheless, to estimate unfolding rates finding that the unfolding reaction occurred every $\sim 50\,\mu s$. This seems an appropriate answer on the problem, just by looking at a trajectory along a reaction coordinate, where unfolding transitions are found to occur on this time scale. Also, the TS is located at $Q \approx 0.1$ and $\xi \approx 11\,nm$ (although the test was not completely satisfactory). These conformations correspond to states with a committor probability of $q_i^+ \approx 0.5$, which in our case are states I_5, I_6 and I_7. In this sense, the TS can be an imaginary surface along these three states in the network picture of the system. Remarkably, the Q values for these three states are $Q \approx 0.1$ while the end-to-end

distance $\xi \approx 9\,\mathrm{nm}$, considerably lower than the proposed one by the Bayesian criterion. This means, that Q describes correctly the unfolding transition, although many states are overlapped within the same Q values. This is a remarkable feature, given that the model does not considers any native-centric assumption, yet the protein seems to be "designed" to have the native contacts as a primary source of stability.

Part II

Mesoscopic Modeling of DNA: of Peyrard-Bishop-Dauxois Model and Beyond

Essentially,
all models are wrong,
but some are useful

GEORGE E. P. BOX

Chapter 6

Brief Overview on the Molecule of DNA

This Chapter is a short introduction on the molecule of DNA, both from the biochemical and biophysical point of view. We focus the relation of the physical properties with the biological function. In particular, we stress the importance of transient local openings—bubbles—on processes such as replication or transcription. These aspects will be of importance for the work developed through the next Chapters.

6.1 Deoxyribonucleic Acid, the Book of Life

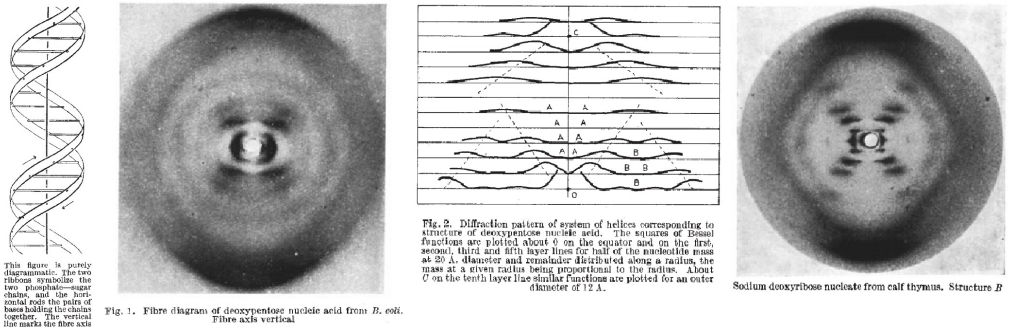
6.1.1 What is DNA?

Life depends on the ability of cells to store, retrieve and translate the genetic instructions required to make and maintain a living organism. This information must be passed on from a cell to its daughters in a faithful way. Also, it must be accessible for the machinery of cells to be "read", in order to perform the different functions there codified.

The Deoxyribonucleic acid (DNA) is the molecule responsible of this, as it carries most of the genetic information stored in all known living beings and some viruses [2, 87]. Chemically, it is a nucleic acid, which together with proteins and carbohydrates are the main macromolecules essential for life.

DNA was identified as the likely carrier of the genetic information in the 1940s. Nevertheless, the best known milestone regarding DNA is the famous paper from 1953 by Watson and Crick—actually a total of five classic papers—which described and presented evidence for the double helix structure [88–92]. Watson and Crick proposed that DNA was composed of two entangled biopolymers with helical structure coiled round a common axis. These two strands were held together by the hydrogen bonds formed by the bases, which followed an specific pairing. The key aspect in this discovery is that with the structure came a mechanism for *copying* this information. This relation of structure-function is a common feature in biological molecules, as, for example, with proteins (see part II).

In a very simple picture (which has proven to be rather an exception) DNA constitutes a linear sequence of "instructions" called genes, having each of them the necessary information for synthesizing a protein, which is in charge of performing one particular task. A gene is divided in two parts, the first one, the promoter, regu-



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Fig. 1. Fibre diagram of deoxypentose nucleic acid from *B. coli*. Fibre axis vertical

Fig. 2. Diffraction pattern of system of helices corresponding to structure of deoxypentose nucleic acid. The squares of Bessel functions are plotted about 0 on the equator and on the first, second, third and fifth layer lines for half of the nucleotide mass at 20 Å. diameter and remainder distributed along a radius, the mass at a given radius being proportional to the radius. About *C* on the tenth layer line similar functions are plotted for an outer diameter of 12 Å.

Sodium deoxyribose nucleate from calf thymus. Structure *B*

Figure 6.1: **Xray diagrams as published by Watson and Crick**: (Left) Proposed structure. (Second from the left) X-ray diffraction photograph and (3rd from the left) diffraction pattern from Wilkins, Stokes and Wilson [89]. (Right)X-ray diagram of structure B from Franklin and Gosling.

lates the expression of such protein. The second one, the coding region, contains the information necessary to synthesize the protein. This information is read by a protein (the RNA polymerase) which synthesizes an intermediate RNA molecule. This molecule is ultimately employed by the cell machinery—namely the ribosome—to synthesize the protein from the amino acid sequence encoded in the RNA molecule. This picture is often referred to as the *Central dogma of molecular biology*, first proposed by Crick in 1956, which gives the first plausible picture of the information flow in biology [93].

This simple proposal is now known to have a higher complexity. First, information can flow in different directions rather than the simple DNA→RNA→ protein, appearing transfers from RNA→DNA or DNA→ protein. Second, the DNA molecule is far from being a linear sequence of genes. They can be overlapped, run in different directions, and there exist also a large fraction of "junk" DNA which does not code any protein, but seems to have a rather relevant role in genetic regulation [94]. Additionally, DNA does not have all the available information of the cell as, for example, post transcriptional modifications play a relevant role by "editing" the amino acid sequences and affecting the final protein function. These exceptions become more significant as we increase the complexity of the organism. In this sense, the simpler picture described above can be thought to be approximately true for simple living systems, such as prokaryotes.

6.1.2 Chemistry and Structure of the DNA Molecule

Chemistry of the DNA Molecule

At the lowest level, DNA is a polymer made up of repeating units, which are called nucleotides. These nucleotides are the monomers of all nucleic acids, and are composed of three different molecules, a nitrogenous base, a five-carbon sugar and a phosphate group (see panel (A) Fig. 6.2). The phosphate and sugar residues form the backbone of each DNA strand. Particularly, the sugar is a 2-deoxyribose, and they are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of contiguous sugar rings. This gives the strand a directionality (3' to 5' or 5' to 3'). The two strands within a DNA molecule run in opposite directions, and thus are antiparallel (see Fig. 6.2).

The nitrogenous bases, (also called nucleobase or base to shorten) are kept inside

the double helix. The bases from each strand keep the helix together by establishing a series of hydrogen bonds, and they arrange in a perpendicular way, like steps on a ladder. In DNA, there are four possible bases, adenine (purine, A), thymine (pyrimidine, T), guanine (purine, G) and cytosine (pyrimidine, C). Pyrimidines only pair with purines, doing so in a particular way. Adenine just pairs thymine through two hydrogen bonds while guanine pairs cytosine through three hydrogen bonds. This is the so-called Watson and Crick pairing. The particular four-letter sequence of bases through the DNA molecule encodes the genetic information.

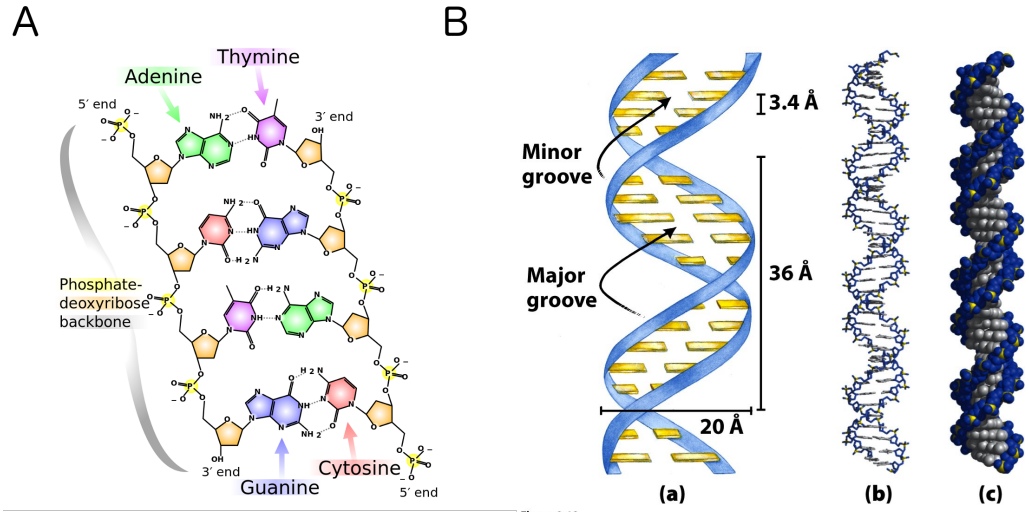


Figure 6.2: **DNA structure**: Panel (A) shows the DNA chemical structure, identifying the four bases, plus the deoxyribose and phospate in the backbone. Panel (B) shows the structure of the B-DNA, highlighting the involved distances, helix turn and the major and minor groove. (Picture modified from Madeleine Price Ball, Wikicommons), and [2])

Interactions within the DNA Molecule

Into the DNA molecule, there exist many interactions which are responsible for its particular double helix structure. In few words, we can consider two major sources for the stabilization of the double helix.

- **Hydrogen bonds between nucleotides**: In the Watson and Crick pairing (A-T, C-G) the two strands of the DNA molecule are held together by the hydrogen bonds formed by the complementary nucleotides. In this case Adenine and Thymine form two hydrogen bonds, while Guanine and Cytosine form three (see Fig. 6.2). This is a weak interaction, compared to covalent bonding. For example, a typical O-H$\cdots$ O bond has a length of 2.75 Å and an energy of $\sim 5k_BT$ [95].

- **Base-stacking interaction**: This interaction is more complex, and comes mainly from the overlap of the π electrons of the bases, and also from the hydrophobic interactions. The conjugated π bonds of the nucleotide bases align perpendicular to the axis of the DNA molecule, minimizing the interaction with the solvation shell. This imposes a well defined distance between the axis of the molecule, and gives rise to the high rigidity of the DNA along this axis. Nevertheless, bases can be pulled out the stack by sliding on each other, perpendicular to the axis [95].

The DNA Double Helix and Alternative Structures

The most abundant DNA form (that proposed by Watson and Crick) is the so called B-DNA. This structure has a 2.37 nm diameter and the double helix is right-handed, turning completely its axis every 10.4-10.5 base pairs in solution (see panel (B) Fig. 6.2). As the backbone of the molecule runs in the helical structure, it leaves "void" spaces in between, called the grooves (see Fig. 6.2 (B)). There are two kinds of grooves depending on their size, as the strands are not symmetrically located with respect to each other. The major groove is 2.2 nm wide, while the minor groove is 1.2nm wide. This spaces are quite important for proteins to access the information encoded in the bases, and they bind usually though the major groove.

Additionally to B-DNA other structures can be found for DNA, namely A-DNA and Z-DNA. A-DNA appears commonly under dehydration conditions, and it has a biological function. It is also right-handed but with a more compact structure than B-DNA (11 bases per turn), which causes the bases to tilt inside the structure. Z-DNA is more different, as it is left-handed, with a repeating structure every two base-pairs. It has also some biological function but its structure is far more unfavorable. As additional structures, we can name DNA-quadruplexes, local arrangements which form in the telomeres (ending regions of chromosomes), and are thought to help in protecting the DNA ends.

Organization of the DNA Molecule

DNA is a very long macromolecule–about 3 mm in the E.coli bacterium, which has itself a length of 2 µm. In the case of eukaryotic cells, where the DNA is longer and is packed inside the nuclei, it forms a complex structure made by the DNA itself, proteins and RNA. This is the chromatin. The chromatin does not pack the DNA exclusively, but has a number of function related to facilitate mitoses, prevent DNA damage or control the replication.

The chromatin is hierarchically organized into a number of substructures, where one of the key elements are the histones, which compact the DNA. Histones are a family of proteins which are able to interact with the DNA molecule by different means—mainly hydrogen bonds and salt bridges between basic amino acids and the negatively charged phosphate groups in the DNA—so that DNA is able to wrap around them.

Starting from the bare DNA helix, the first structural level (see Fig. 6.3) is the nucleosome, which is also the repeating element in the chromatin. This is a segment of DNA wounded around eight histone protein cores. Repetitions of this entity separated by fragments of DNA form the "Beads-on-a-String" picture, where the DNA molecule is wounded around histone molecules (beads), separated by unwounded segments.

This structure coils into the 30 nm fibre or filament when the histone H1 is added (see Fig. 6.3). The exact details of this structure are still not completely known. The next level is a special conformational arrangement with the aid of some scaffold proteins, to ultimately form the chromosomes, the highest level structure, which appears during cell division.

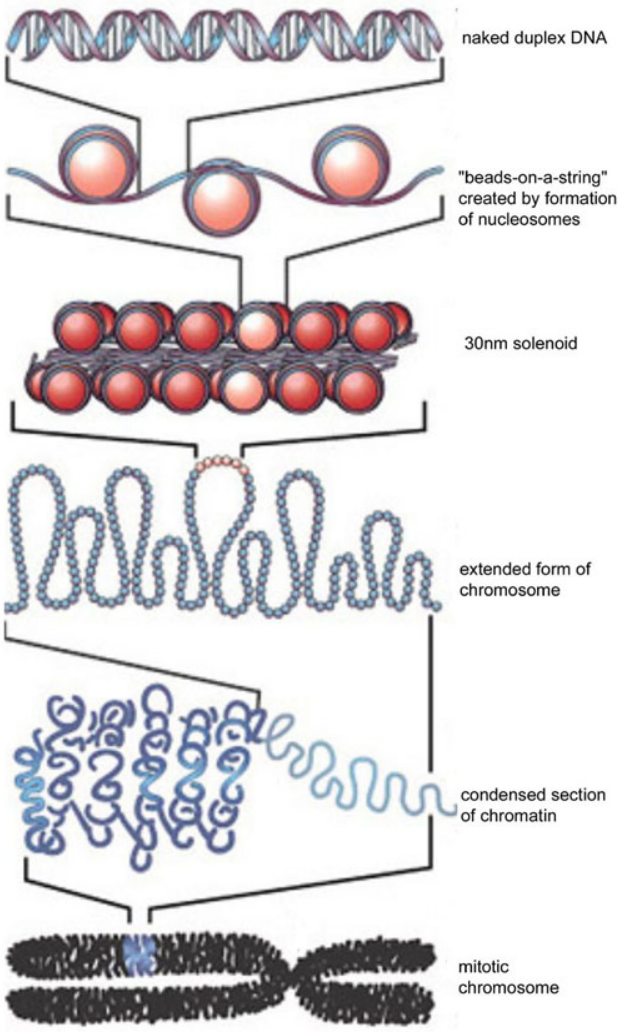


Figure 6.3: **Major chromatin structures**: DNA is compacted in a highly hierarchical structure called chromatin. Chromatin is a combination of the DNA molecule and several proteins, which help in organizing the long molecule and prepare it for different cellular processes, (taken from Felsenfeld and Groudine, Nature 2003)

6.2 DNA Function

The DNA molecule has a twofold mission. First, it has to store this information unchanged through subsequent cell divisions, second to have this information accessible to the cell machinery.

Thus, DNA is involved in two major molecular processes associated with the mentioned functions, replication—the process by which a copy of the molecule is synthetized—; and transcription—the first step of gene expression, which a particular segment is copied into RNA.

6.2.1 Replication

During replication, two DNA molecules are produced from a mother molecule. Each of the two strands of the mother DNA is conserved and serves as template for the daughter molecules, so it is a semiconservative replication. This process is the basis of the genetic inheritance.

DNA replication occurs thanks to a large number of proteins and enzymes which are responsible of unwinding the DNA molecule, stabilizing the replication bubble, reading or synthesizing the daughter strand, among other functions (see Fig. 6.4). The enzymes responsible of the replication itself are the DNA polymerases. They cannot initiate the synthesis of new strands on their own, but can extend DNA or RNA strands paired with a template one. In this sense, an RNA primer is needed to initiate replication. Also, the DNA polymerase can just move in 5' to 3' direction, creating an asymmetry of the two strands (leading and lagging strand, see Fig. 6.4). The replication mechanism is thus slightly different for the two template strands.

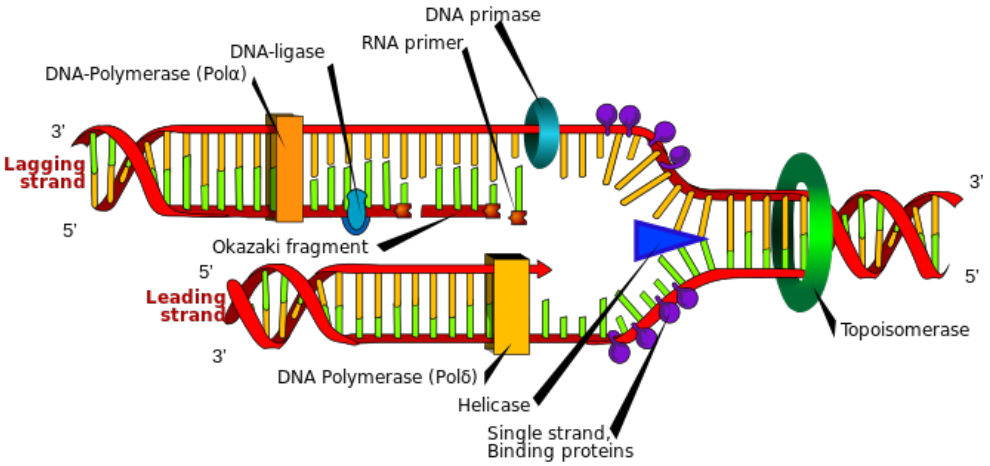


Figure 6.4: **Picture of the replication process**: The replication fork and with the leading and lagging strands are depicted, highlighting the difference the replication direction for each of them. Also, some of the involved proteins are sketched. (Picture by Mariana Ruiz, Wikipediacommons)

The replication starts with the *initiation process* which implies the formation of a protein complex which unwinds the DNA molecule, opening the replication bubble. Several proteins are involved here, responsible of binding to a specific DNA site and opening it. The elongation is the replication process itself, where the DNA polymerase has the enzymatic activity and synthesizes the DNA by adding nucleotides complementary to the template strand. On the one hand, the synthesis of the leading strand is continuous. Starting with the RNA primer, the DNA polymerase simply ads complementary nucleotides in a sequential way. On the other hand, the lagging strand is replicated in discontinuous way, by adding the so called Okazaki fragments, which must be ligated to form the new strand, requiring a veryprecise coordination of the cell machinery. Many other proteins take part in the process, as helicases which unwind the DNA or the topoisomerases, which release the strain.

6.2.2 Transcription

The transcription is the first step by which the information stored in the DNA is read and translated into proteins. In a nutshell, the whole process of gene expression is done in two steps. First DNA is read and copied into mRNA. Then, this mRNA is read by the ribosome and translated into the final protein. Obviously the actual process is quite more complex, and it involves the participation of several proteins and complexes.

Before reviewing briefly the process by which DNA is copied into RNA, let us look at the basic structure of a gene, in other words, how is the information written along the DNA sequence.

Simple Vision of a Gene Structure

In a simple picture, a gene is made of three different regions, the initiation, the coding and the termination region. The first one is the promoter, and is the region where the transcription starts. The coding region contains the sequence that would be transcribed to mRNA, while the latter indicates the end of the gene and thus the detaching of the transcription machinery.

The promoter is a sequence of around 100 base-pairs long in prokaryotes where the RNA polymerase binds to start transcription. It contains the Trascription-Starting site, where the sequence that would be transcribed to mRNA starts, and thus from where the transcription bubble forms. By convention, the DNA base pairs that correspond to the beginning of the RNA transcription are given positive numbers, and those preceding the Transcription-Start site negative numbers. Thus, the Transcription-Start site is labelled as $+1$, and promoters usually span from -70 to $+30$, approximately.

The function of promoters is not just to indicate the initiation of a gene, but also to promote transcription. Additional proteins, called transcription factors, bind to promoters at specific sites to facilitate binding of the RNA polymerase. Also, some transcription factors might activate or repress the synthesis of the particular gene by binding to specific sites. The particular features change a lot from prokaryotes to eukaryotes or even from organism to organism. Most bacteria promoters reveal the importance of sites -35 and -10 for binding of factors that recruit the RNA polymerase [2]. Many archaea and eukaryotes show another regulatory element known as TATA box, which is responsible of binding of the TATA-binding protein, which unwinds the DNA at this site. The TATA box has a typical sequence 5'-TATAAA-3' or variant, and is usually located between -25 to -35 position. Due to the AT rich content, this site is particularly weak, so the double-helix breaks with more ease, facilitating the role of the TATA-binding protein [96].

The Transcription Process

The transcription process is simpler than the replication, provided just one of the strands is "read" (see Fig. 6.5). As anticipated before, the RNA polymerase is responsible of synthesizing an RNA molecule complementary to the DNA sequence of the gene. A transient transcription bubble must be formed—of around 15 base pairs—that allows the enzyme to access to the base pair sequence. The DNA is read in the 3'-5' direction, and the RNA produced in 3'-5'. The RNA polymerase needs no primer, unlike the DNA polymerase.

6.3 Biophysical Properties of the DNA Molecule

The DNA molecule is a very long and thin polymer, made up of two entangled chains, DNA has a large number of interesting properties related with the shape it adopts in three dimensions, with the possibility of breaking locally the double helix

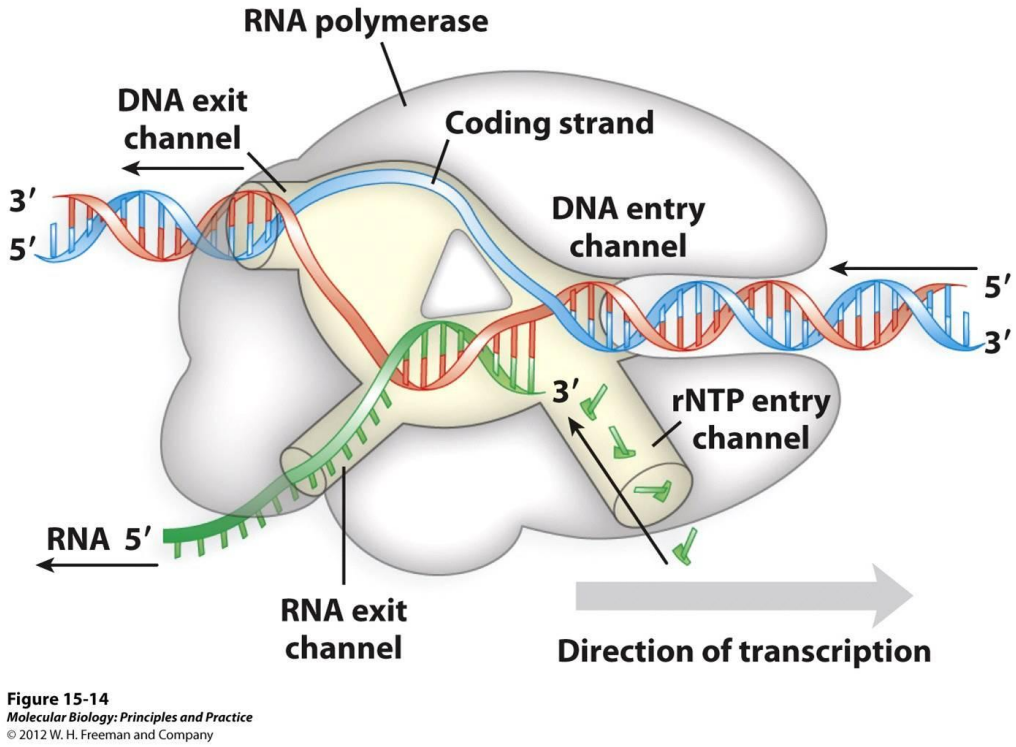


Figure 6.5: **Picture of the transcription process**: The RNA polymerase is responsible of reading the DNA sequence and to synthesize the mRNA. A transcription bubble must be formed, in order for the enzyme to access the information (taken from Molecular Biology, 2012).

structure, and with its behavior under the presence of an external force. This latter feature has won popularity since the advent of single molecule techniques.

6.3.1 Twisting and Curving DNA

Although DNA molecules by themselves are already quite interesting objects for the biophysical study, the actual *in vivo* functions can only be understood with the action of a huge number of proteins which interact with DNA to perform every function in which it is involved. Twisting and curving are two actions of the DNA molecule which are directly related with its biological functions, and particularly with protein interaction.

Bubble Formation

Twisting, or to be more precise, untwisting is a feature of primary interest and it is directly related with the double helix formation. This action of the DNA molecule was already suggested in the brief discussion about replication and transcription, and it involves the local unwinding of the double helix to form a bubble of unpaired bases, exposed to the solvent, and thus to the cell machinery. These bubbles form in replication and transcription processes, where the DNA or RNA polymerase bind to the bubble, reading the sequence and synthesizing the corresponding match molecule. These bubbles are typically around 15-20 bases—the energetic cost of breaking the double helix and expose the bases to water is very high—and they travel along the DNA molecule at a relatively high speed (around 100 bases per second) [95].

The DNA double helix structure implies that, in order to form a bubble, either the traveling protein screws around DNA, or DNA screws around itself. It is currently known that the polymerase is kind of stationary, while it is the DNA which screws [95]. Bubbles can form spontaneously, but are also stabilized by a large number of proteins which help in unwinding the double helix, release the stress and maintain the bubble over significant times in order to perform the pertinent biological process.

The DNA sequence is non homogeneous, and thus, the local physical properties change from site to site. They are not randomly distributed, and as a general rule, we known that nature makes the double helix openings to start at energetically weak places, where the separation between the two strands is likelier to begin [95]. The hydrogen bonding between the two strands and base stacking are the main sources of stability. Some pairs are known to be weaker than others, particularly A-T pairs have two hydrogen bonds, and so large stretches of these pairs form weak sites, where DNA would open with higher probability. These prototypic weak sequences are something like 'TATATATA' or 'TAATAATAA', which appear in promoter sequences and are known to have a key role for in transcriptional processes [95, 97, 98] .

DNA Denaturation

The extreme case of untwisting of the DNA molecule is a bubble which spans over the whole chain, separating the two strands completely. This is a phenomenon called denaturation, as the DNA molecule is loosing its structure, becoming two random coiled polymers. Denaturation can be caused by a number of elements, and one of them is temperature. Thermal denaturation or DNA melting is the process by which the two DNA strands separate due to heating.

This process is interesting from a physical point of view, as it involves a phase transition in a one-dimensional system, the DNA chain. This transition can be monitored experimentally, as the exposition of the base-pairs involve an increase of the UV light absorbance at 260 nm [99, 100]. Experiments on artificial homopolymers show a sharp increase in the UV absorbance, certifying that DNA melting is a genuine phase transition [99–101].

The denaturation process starts with few local bubbles that soon would become more frequent, often fusing to each other. Finally a unique denaturating bubble would span over the whole chain, separating fully both strands. In some way this can be thought as a nucleation process (see Fig. 6.6).

Curving

DNA is a rather stiff polymer, with a persistence length of about 50 nm, which means that it is a rigid rod for about 150 base pairs. Sometimes, DNA must curve around proteins -like histones to form the nucleosome. This costs obviously some energy, given that we have to stretch the outer part of the polymer. As DNA is made of discrete steps, this requires base pairs to roll a certain angle in order to adopt a curved conformation. As this carries an energetic cost, it is obviously sequence dependent.

Particularly, A-A/T-T sequences curve more easily, compared C/G rich ones [95]. We can expect again to find them in a large number of situations when DNA requires such conformation. Easy example is that of the nucleosome, where stretches of DNA

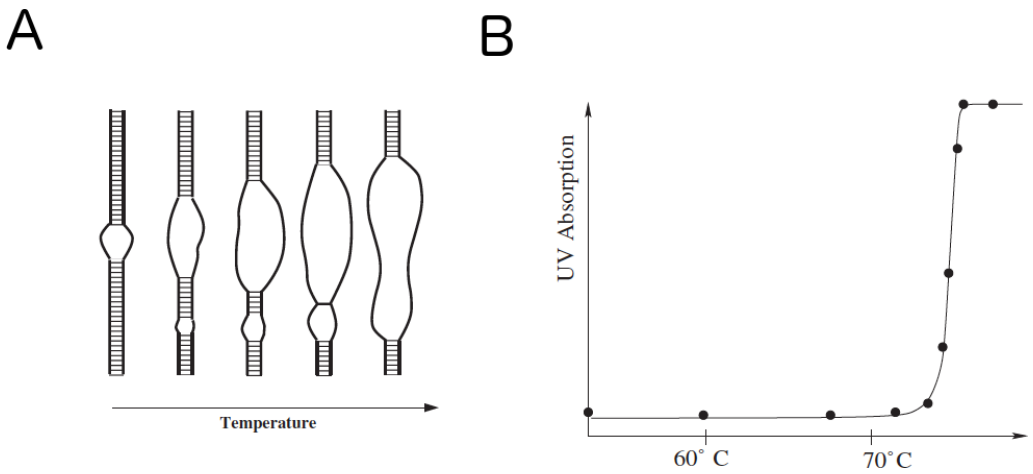


Figure 6.6: **Denaturation of the DNA molecule**: Panel (A): The melting of the DNA is driven by the formation of bubbles, which merge, and ultimately span the whole molecule. Panel (B) The denaturation of the DNA molecule can be observed experimentally with UV absorption, suggesting a phase transition, given the sharp rise in the absorbance (Picture modified from [100], adapted from [99]).

curve around the histones, separated by uncurved stretches. Other popular examples are binding of repressors or Zinc fingers, which are able to recognize particular sequences of DNA.

6.3.2 Topology of DNA

By topology of polymers, we refer generally to properties such as linking o entanglement which are invariant under smooth geometrical deformations. These properties are related to what is known as DNA supercoiling, important feature for many biological processes.

In the relaxed B form, DNA has 10.4 to 10.5 base pairs per helical step. Nevertheless, this arrangement changes in many situations, when the DNA must add or subtract twists, this is to wound or to unwound DNA. This imposes a strain which, would lead to adoption of new shapes, such a figure-eight, in the case of closed chains (see Fig. 6.7). This is known as DNA supercoiling.

Mathematically, one can describe DNA supercoiling by the linking number L_k, the number of crosses a single strand makes across the other. For a closed circular chromosome, this number cannot be altered without breaking the strands. The linking number can be written as the sum of the Twist Tw (number of twists or turns of the double helix) and the Writhe Wr (number of coils or writhes the strand does, see Fig. 6.7). This means that for a closed chain, changes in one imply changes in the other $L_k = Tw + Wr$ [95, 102].

Unwinding of the DNA as in bubble formation, implies a change in Tw, and is forces the DNA molecule to adopt some supercoiling in order to change Tw. *In vivo*, the DNA molecule might change its L_k in order to relax the stress, specially in special situation during the replication and transcription processes, where the excessive stress the bubble creates could stop the enzymatic activity of RNA or DNA polymerase. This is possible due to a set of proteins known as topoisomerases, which perform topological changes in the DNA molecule by cutting the phosphate backbone, and changing the linking number [95, 103].

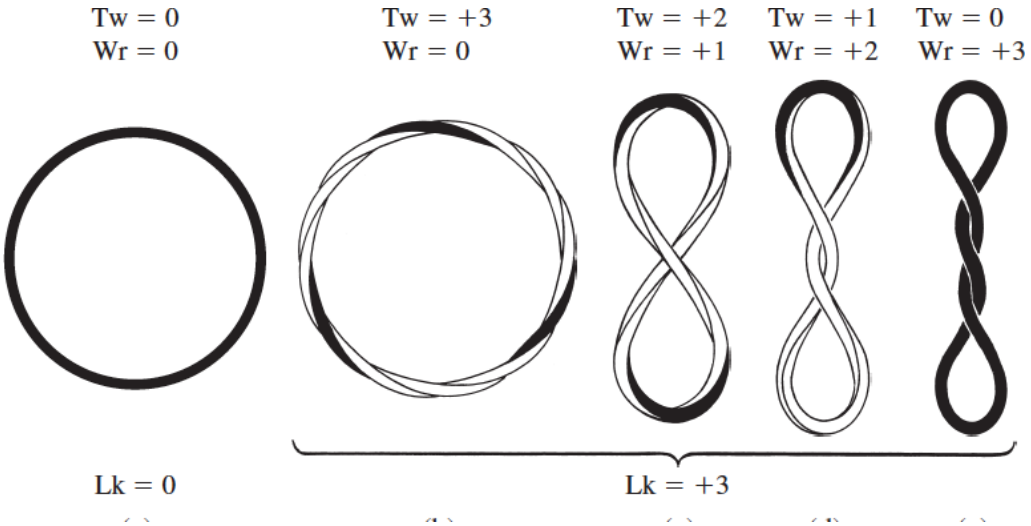


Figure 6.7: **Graphic example of supercoiling on closed DNA molecules**: First molecule is an open relaxed circle with $L_k = 0$. Remaining four ones, have same L_k but in different configurations, according to the interchange between Wr and Tw (Picture taken from [95]).

Chapter 7

Peyrard Bishop Dauxois Model: DNA at the Mesoscale

In this chapter, we introduce and review the basic aspects of the so-called Peyrard-Bishop-Dauxois model. This DNA-model at the mesoscopic level starts with very simple assumptions, yet is able to reproduce some properties of the molecule at the base-pair level, namely the melting transition, and also bubble formation. We start defining the model as it was first conceived. Next, we focus on the practical issues, stressing its simulation to reproduce equilibrium and dynamic properties.

7.1 Modeling DNA, Different Questions, Different Levels

The molecule of DNA exhibits a large number of properties and behaviors which span over a wide spatial and temporal range. Proposing a mathematical model to reproduce some experimentally observed properties implies first the choice of the appropriate level of description. In the present case, it covers over six orders of magnitude in length, from the atomic level, to the chromosome arrangement.

For example, if we wish to study the properties of DNA as a polymer, a continuous model, such as the Worm-Like-Chain model, would serve well. This is contrained to a micrometer scale, where the actual structure of the DNA is meaningless, and effects such as the separation of the double strand are not incorporated. On the other extreme, all-atom descriptions appear as the highest resolution choice, incorporating every degree of freedom (neglecting quantum effects) so, the full complexity of the system is taken into account. The real limitation comes from the computation time, which currently makes it unfeasible to simulate sizes over few tenths of base pairs, within times larger than hundreds of ns [104]. If we are, for example, interested in studying bubble dynamics, which have a typical size of 10-20 base pairs and last for several µs [95], this is not practical. In addition, a physical property such as the melting transition, which is a cooperative effect and would need hundreds of base pairs, could also not be tackled with such a fine level of description.

At the intermediate level, there exist other options, such as coarse-graining or mesoscopic models, where some microscopic description is maintained, while many degrees of freedom are integrated out. Coarse-grained models typically gather several atoms up, leaving "effective atoms", which decrease the detail but also the

number of particles [63, 64]. Mesoscopic models move to higher levels, and are usually based on some physical assumption on the system which tackles directly the problem subject to study. These options allow naturally for larger scales and longer simulation times. For example, in the case of DNA this would allow describing properties at an intermediate level, such as bubble formation or the melting transition. This choice has obviously an straightforward flaw, which is the election of the relevant degrees of freedom and the right parametrization of the model. Often, fitting to the experimental phenomenology is a satisfactory verification.

The scale of DNA in which we are interested is that of the base pair. This element is a key characteristic of the DNA molecule, as it is the entity that encodes the information. Such level is the appropriate one to understand a large number of interesting properties which are representative of the DNA, such as transcription or replication. In addition, the melting transition can be also tackled under such level consideration.

The simplest proposal at this point is the one-dimensional Ising model [105]. Here, the degrees of freedom of the system are represented as the discrete states of each base pairs, equal to 0 if the base is closed and 1 if it is open. This choice has the obvious advantage of its simplicity, as we can use the well-known toolkit of statistical mechanics to undertake the problem. Unfortunately, it suffers from some drawbacks, namely the estimation of the parameters involved in the problem which cannot be predicted from the known phenomenology (such as the coupling between the base pairs) or its two-state status, which clearly restricts greatly the actual picture of the system.

These kind of toy models have been used as a simple approach to study the thermal denaturation [105], an interesting phenomenon from the theoretical point of view, as a phase transition takes place in a one-dimensional system. Defining a model with bigger predictive ability, requires a more sophisticated formulation. The Peyrard-Bishop-Dauxois model sits somewhere in between, keeping that statistical mechanics model spirit, yet proposing a more realistic and meaningful way to approach to the DNA molecule.

7.2 The Peyrard-Bishop-Dauxois Model: a Simple Model for DNA at the Base-Pair Level

The Peyrard-Bishop-Dauxois model (PBD from now on), keeping the same degree of freedom (the base pair state), takes on the next step in complexity and represents it with a continuous variable y_i, the base pair opening, in other words, how much does the i-th base pairs departs from its equilibrium position.

Originally the PB model [106], it was born as an analytical model able to predict the DNA melting transition, explaining also its driving effect. Soon, in its extension to the PBD model, it incorporated the nonlinear character of the current model, enhancing the cooperativity of the system. This modification allowed to reproduce more faithfully the melting transition, particularly the denaturation rate or width of the transition [107, 108].

The PBD model has also been used to explore the dynamical properties of the DNA molecules, showing its ability to reproduce the formation of bubbles in the DNA chain, feature driven mostly by the nonlinear character of the mode. The

main interest here is the study of heterogeneous biological sequences, which carry the genetic information. The correlation between bubble formation and the existence of biologically relevant sites was soon proposed. This latter point is probably the most controversial one, with some works claiming this direct relationship, while others preferring to remaining cautious (see references [109–113]).

7.2.1 Description of the PBD Model

The PBD model is a DNA model at the base-pair level, where the only degree of freedom is the stretching y_i of the i-th base pair. The value $y_i = 0$ corresponds to a closed base-pair. Positive values indicate increasing opening of the base-pairs, as in DNA denaturation. The variable y_i can also take negative values corresponding to a compression of the linking bonds from its equilibrium position. Negative values are forbidden by steric hindrance.

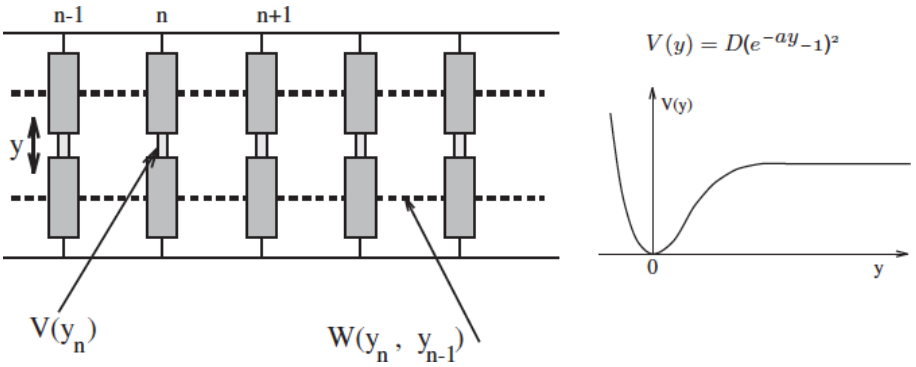


Figure 7.1: **Schematic picture of the PBD model.** The whole complexity of the DNA molecule is reduced to a single degree of freedom per base pair, the base pair opening y_i. The Hamiltonian is reduced to two interaction terms, the stacking potential $W(y_i, y_{i+1})$ and the inter base pair potential $V(y)$. (Picture taken from [100]).

A schematic view of the model is depicted in Fig. 7.1, and is defined by the Hamiltonian

$$\mathcal{H} = \sum_{i=1}^{N} \left[\frac{p_i^2}{2m} + V(y_i) + W(y_i, y_{i+1}) \right], \tag{7.1}$$

where i is the index labeling each base pair of an N base-pair chain, $p_i = m dy_i/dt$ and m its reduced mass.

In this equation we identify two energy potential terms which account for the main sources of stability described in previous chapter: $V(y_i)$ describes the interaction between the two bases defining a base-pair, and it is an intra base-pair potential term, while $W(y_i, y_{i+1})$ describes the interaction between adjacent bases along the DNA chain, setting an inter base-pair interaction. $W(y_i, y_{i+1})$ is often termed as the *stacking* potential (as it represents the staking interaction in the molecule).

The potential $V(y_i)$ is modeled with a standard expression for chemical bonds. Physically it accounts for the interaction between the two nucleotides from a base pair. Coarsely, this interaction are the hydrogen bonds formed to keep the two chains together, two in the case of A-T pair and three in the case of C-G pairs. In this sense, the sequence content has been introduced generally in this potential term. Nevertheless, being a mean force potential potential, it integrates multiple

contributions in an effective way, such as electrostatic repulsion of the phosphate groups, the solvent effect, the solution ions screening or entropic effects. The original PBD model considers a Morse potential for this interaction,

$$V(y) = D(e^{-\alpha y} - 1)^2, \tag{7.2}$$

where D is the dissociation energy of a base-pair ($V(0) = 0$ and $V(\infty) = D$) and α sets the amplitude of the potential well ($V''(0) \equiv \kappa_V = 2D\alpha^2$, being $\omega_V = \sqrt{\kappa_V/m}$) the fundamental frequency of the equilibrium state $y = 0$). Qualitatively, this term has the appropriate shape, an strong repulsive part for $y < 0$—accounting for steric hindrance—, a minimum well at $y = 0$ and a flat region for large y as the force between the base-pairs vanishes allowing for a complete dissociation. This expression is the original one of PBD model, nevertheless, in subsequent revisions, some modifications have been introduced [114–116] (see Chapter 8).

The potential $W(y_i, y_{i+1})$ accounts for the stacking interaction, of key importance for the DNA molecule stability. It has different physical origins. First, the sugar-phosphate strand sets a rather rigid polymeric structure connecting the bases. Pulling a base pair apart from its bond tends to pull all the neighbors due to this connection, and thus an energetic cost. Second, the direct interaction between the base-pair plateaux, due to an overlap of the π-electron orbitals of the organic rings making up the bases.

The original PB model considered a harmonic potential for this contribution. This is a good approximation if the staking interaction is strong enough to keep y_i close to y_{i+1} at all times, which is not true for DNA. Nevertheless, this approach is convenient as it allows for an analytical study of the model. Years later, the expression was modified to include a nonlinear term, which enhanced the cooperativism of the model. This is a desirable effect as it helps in reproducing more faithfully the melting transition, particularly sharpening it. In this sense, the nonlinearity of the model becomes a key aspect of its performance. The stacking potential takes the following expression [107, 108]:

$$W(y_i, y_{i+1}) = \frac{1}{2}K\left[1 + \rho e^{-\delta(y_i + y_{i+1})}\right](y_i - y_{i+1})^2. \tag{7.3}$$

This potential shape sets a position dependent coupling constant, setting a hard spring of constant $K(1 + \rho)$ for small openings, and a softer spring of constant K for large openings. The parameter ρ sets the intensity of this nonlinear interaction, while δ the length scale for this behavior.

7.2.2 Parameter Choice

In mesoscopic models, the critical step is the choice of the parameter set. As the Hamiltonian of the system includes effective potentials, one cannot derive the parameters from first principles, as it shall be done in an atomic level or even coarse-graining. Typically, the proposed strategy combines two considerations. First, the parameters must remain in a physically reasonable interval, given the energy and length scales of the given problem. Second, on tunes them in order to reproduce successfully the properties we are modeling given the experimental evidence. In our case this is the melting transition temperatures of different sequences, but also an appropriate shape of this transition, as well as the formation of transient bubbles

of proper length and lifetimes. In addition, the model must be robust upon the parameter set, this is, slight changes in the parameters must not produce dramatic changes in the model output.

A suitable choice is that of Campa and Giasanti in their work of 1998 [117]. There, experimental tests on some short DNA chains were made, melting them and comparing the melting profiles with the prediction of PBD model. This allowed a proper tuning of the set of parameters, namely: $K = 0.025eV/\mathring{A}^2$, $\rho = 3$, $\delta = 0.35\mathring{A}^{-1}$, $D_{AT} = 0.05eV$, $D_{CG} = 0.075eV$, $\alpha_{AT} = 4.2\mathring{A}^{-1}$ and $\alpha_{CG} = 6.9\mathring{A}^{-1}$.

Nevertheless, along this work, we modify the original model in different aspects, so a slight retuning of the parameter set will be done.

7.2.3 Adimensionalization of the Equations

Although previous definition of the parameters of the model sets a link between it and experimental evidence, dimensionless quantities are often employed, both for theoretical calculations and numerical simulations. This helps in reducing the number of parameters involved in the model, as well as setting a more confortable definition of the involved units. We define a dimensionless stretch of the base pairs $Y = \alpha_{AT} y$, and we measure the energy units with the depth D_{AT} of the Morse potential for A-T pairs. The dimensionless Hamiltonian is $\mathcal{H}' = \mathcal{H}/D_{AT}$, which defines the quantity $S = K/D_{AT}\alpha_{AT}^2$, and the dimensionless time $\tau = \sqrt{D_{AT}\alpha^2/mt}$. The adimensional Hamiltonian is:

$$\mathcal{H}' = \sum_i \frac{1}{2}P_i^2 + \frac{1}{2}S(Y_i - Y_{i-1})^2 + (e^{-Y_i} - 1)^2, \tag{7.4}$$

where $P_i = dY_i/d\tau$. Now there is a single parameter involved, S.

7.3 Simulating the PBD Model

Most of the progress on the PBD model has been made simulating numerically the equations of motion of the model. This allows to deal with the intrinsic nonlinearities of the model, but also to study heterogeneous sequences, impossible to be tackled with any analytical approach.

In order to simulate numerically the model, one must consider thermal fluctuations, which are rather important at the scale of behavior we are in. In order to do so, many approaches exist. For instance, Monte-Carlo simulations are an efficient strategy to compute ensemble averages, like, for example, to reproduce the melting transition. Dynamical properties such as bubble formation (in which we are interested) cannot be undertaken via Monte Carlo methods, and thus, real dynamics should be simulated.

7.3.1 Dynamics of the PBD Model: Integrating the Langevin Equations of Motion

Through this work, we simulate the dynamical behavior of the PBD model by integrating numerically the Langevin equations of motion For our system, we have:

$$m\frac{d^2y_i}{dt^2} = -m\gamma\frac{dy_i}{dt} - \nabla E + \eta_i(t), \tag{7.5}$$

where y_i is the coordinate of each base pair, m its mass, $E = V + W$ the potentials acting on the system, γ the effective damping of the system and $\eta_i(t)$ the random noise force, of zero average (white noise) $\langle\eta_i(t)\rangle = 0$ and correlation given by $\langle\eta_i(t)\eta_k(t')\rangle = 2m\gamma k_B T\delta_{ik}\delta(t-t')$, where T is the temperature of the thermal bath. The integration of these set of equations can be done via several algorithms. Particularly, we will use an stochastic fourth-order Runge-Kutta algorithm [81].

7.3.2 Observables to Characterize the Melting Transition

In order to characterize the melting transition, we compute different observables based on the thermodynamic properties of the system, and also on the particular arrangement of the base pairs. For instance, the average energy $\langle u\rangle$ and the heat capacity C_v are defined as:

$$\langle u\rangle = \frac{1}{Nt_s}\sum_{n,t}^{N,t_s}[W(y_n, y_{n-1} + V(y_n)], \tag{7.6}$$

$$C_v = \frac{1}{kT^2}\left(\langle u^2\rangle - \langle u\rangle^2\right), \tag{7.7}$$

where N is the total number of base pairs to study and t_s the total simulation time. Additional topological magnitudes as the mean displacement $\langle y\rangle$ are useful, as we expect a sudden rise of the average opening, once the melting transition is overcome. We define $\langle y\rangle$ as:

$$\langle y\rangle = \frac{1}{N}\sum_{n}^{N}\langle y_n\rangle; \qquad \text{where} \qquad \langle y_n\rangle = \frac{1}{Nt_s}\sum_{n,t}^{N,t_s} y_n(t). \tag{7.8}$$

Additionally, the fraction of open chain P, can be calculated from the probability that the n-th base-pair is opened $P_n(y_{th})$, by defining a threshold y_{th} over which a base-pair is defined to be open. Thus,

$$P = \frac{1}{N}\frac{1}{t_s}\sum_{n}^{n=N}\sum_{t}^{t_s}\Theta(y_n(t) - y_{th}), \tag{7.9}$$

where $\Theta(x)$ is the Heaviside step function, such that $\Theta(x) = 0$ for $x < 0$ and $\Theta(x) = 1$ for $x \geq 0$. At low temperatures we expect $P = 0$, as the chain is closed on average. At increasing values, it would start to take higher values, reaching $P = 1$ above the melting transition.

7.3.3 Observables to Study Bubble Dynamics

The formation of bubbles is another interesting feature. Particularly, we are interested in computing which regions are more likely to be open along the simulated trajectories. For this regard we calculate the average opening of the chain:

$$\langle y_n\rangle = \frac{1}{Nt_s}\sum_{n,t}^{N,t_s} y_n(t), \tag{7.10}$$

and also the probability that a base pair opens:

$$P_n(y_{th}) = \frac{1}{t_s}\sum_{t}^{t_s}\Theta(y_n(t) - y_{th}). \tag{7.11}$$

These two quantities provide rather similar information. Those regions where bubbles form more easily should show peaks in the $\langle y_n \rangle$ and $\langle P_n \rangle$ profiles.

7.3.4 Principal Components Analysis

PCA is also a useful tool for understanding the PBD model [58, 59, 116]. First, the study of the eigenvector spectrum as a function of temperature gives us information about the melting transition, as we study in Chapter 8. Also, direct inspection of the eigenvectors helps us finding more flexible regions in the sequences, as those which experience large amplitude motions. Finally, we also use PCA as a dimension reduction tool, in order to build Markov state models, as done in Part I.

Briefly, we remember the basic features of PCA (see Chapter 2 for further details). Mathematically, it is based on building the correlation matrix. If our system is described by a set of N coordinates y_i, then the correlation matrix:

$$C_{ij} = \langle y_i y_j \rangle - \langle y_i \rangle \langle y_j \rangle. \tag{7.12}$$

Diagonalizing this matrix, we obtain an ordered set of eigenvalues $(\lambda_1 > \lambda_2 > \cdots, \lambda_N)$, with their associated eigenvectors $(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N)$. λ_i is a measure of the amount of fluctuations corresponding to coordinate given by $\mathbf{v}_i$. In this sense, by keeping just the first few PC, we describe faithfully the system in what regards to fluctuations from the average behavior.

Here, it is useful to define the "principal frequencies" associated to each eigenvalue λ_i,

$$\omega_i = \sqrt{\frac{kT}{\lambda_i}}. \tag{7.13}$$

This definition is analogous to normal modes. If our system is restricted to a linear (harmonic) behavior, PCA coincides with normal mode analysis. For example, this happens for low temperatures, where the principal frequencies coincide with the normal mode frequencies. In this sense, it is hard to say if the defined principal frequencies can be associated to actual *physical* frequencies of the system. Nevertheless, they are a useful tool for characterizing the melting transition of the DNA molecule.

Chapter 8

PBD Model with a Solvation Barrier: Towards a more Faithful Description of the Melting Transition and Bubble Dynamics

This chapter aims to present and explain a modification in the PBD model proposed in [116]. We introduce a potential barrier in the on-site potential in order to account for the solvent effects. Studying the melting transition, we observe a more faithful recreation, as the transition is narrowed. Regarding bubble dynamics, bubbles last longer, in agreement with the expected order of magnitude. We use the modified model in homogeneous sequences and on P5 promoter sequence. Additionally, we use PCA to characterize the melting transition.

8.1 Motivation

In this Chapter, we present a modification of the PBD model which allows for a more faithful description of the melting transition and the bubble dynamics [116]. The inclusion of a solvation barrier in the original Hamiltonian of the model leads to a narrower melting transition, from the equilibrium perspective. From the dynamical point of view, it allows longer-lasting bubbles, and thus to a more reliable description of the phenomenology.

We analyze both properties, discussing the importance of this barrier. The effect of the different parameters of the model is also discussed. In addition, we employ Principal Component Analysis as a powerful tool for understanding our system. This technique allows to characterize the melting transition, with reminiscences to normal mode analysis. In heterogeneous sequences (biological ones), PCA provides a effective mechanism for identifying "softer" regions of the sequence, understood as those where bubbles form with higher probability. The relation of these regions with biologically relevant ones (as Transcription Starting Sites, or binding sites for different regulation factors) is a controversial topic, as mentioned before. We concentrate on a viral promoter P5, which has already been studied within the context of the PBD model [109, 118].

8.2 The Introduction of a Solvation Barrier

The introduction of a solvation barrier in the inter base-pair potential is a successful improvement of the PBD model [114, 116, 119]. This barrier answers to the necessity of including entropic and enthalpic effects in the effective Hamiltonian of the system. As a single base pair flips out of the stack it should lower the effective potential due to the entropy gain. In the same sense, in order to reclose this base pair, an entropic barrier should be surmounted. For a correct dynamical description of the open states, this entropic barrier should be included in the effective potential $V(y)$. Besides this effect, this barrier should contain also pure enthalpic contributions due to the hydrogen bonds it would form with the solvent, which have to be broken upon reclosing [114, 115]. The existence of this barrier has been reported in free energy calculations derived from all-atom simulations of DNA [120].

Mathematically, this barrier should appear just after the equilibrium well, avoiding any effect on the original shape of the potential. Intuitive control of its height and position would also be advisable. Hence, we choose a gaussian barrier term, simply added up to the original Morse potential [116].

$$V(y) = D(e^{-\alpha y} - 1)^2 + Ge^{-(y-y_0)^2/b}, \tag{8.1}$$

where G controls the height of the barrier, y_0 its position and b its width. Regarding the original parameters, a qualitative and quantitative reasonable election is $G = 3D$, $y_0 = 2/\alpha$ and $b = 1/2\alpha^2$. We show in Fig. 8.1 a plot of this potential for the A-T and C-G pairs.

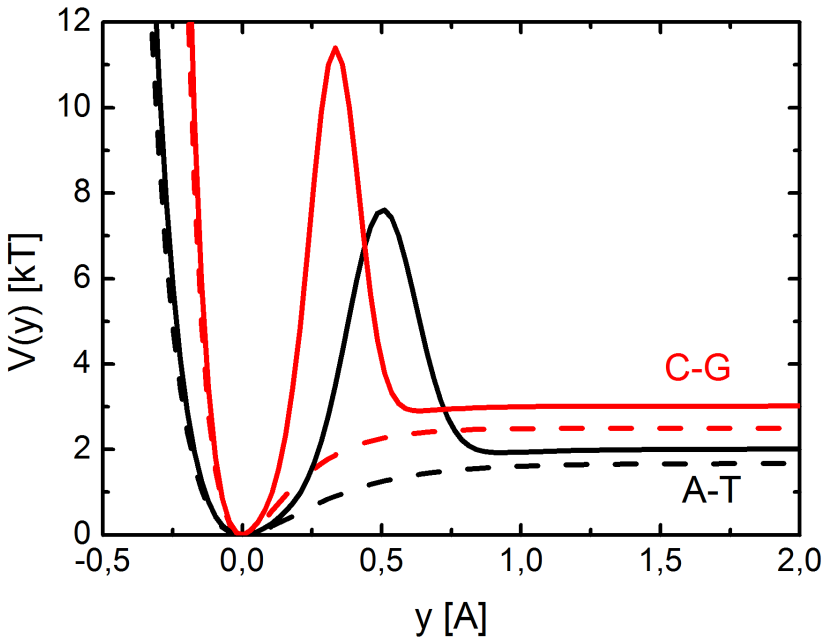


Figure 8.1: **Intra base-pair potential with (solid line) and without (dashed lines) salvation barrier for A-T (black) and C-G base pairs (red).** The potential parameters has been set according to reported melting temperatures of homogeneous sequences.

We use the modified PBD model to study the melting phase transition and bubble dynamics. Both features are to be analyzed first on a uniform chain (*i.e.* A-T or

C-G homogenous chain), to be later applied on real heterogeneous sequences. This first study, although has little biological relevance, allows some analytical treatment and is also useful for tuning the parameter set. This choice is done by comparing the melting temperatures with experimental values.

8.3 The Homogeneous Sequence

8.3.1 Choosing the Parameters of the model: Fitting the Phase Transition on Uniform Sequences

As mentioned in Chapter 8, the set of parameters proposed by Campa and Giasanti in [117] is an appropriate one for the original PBD model. We modify the original model, so the parameter set must be retuned, in order to keep the same features. In order to determine the new parameter set, we compare melting temperatures for homogeneous sequences of A-T and C-G chains, which have a reported to be $T_m^{AT} \approx 310\,\mathrm{K}$ and $T_m^{CG} = 350\,\mathrm{K}$ [121]. Nevertheless, one should be aware that, in general, the melting temperature depends on the length of the sequence, base composition, topological structure and salt concentrations, so these values are merely orientative.

In our case for the model with no barrier, we had $D_{AT} = 0.043\,\mathrm{eV}$, $\alpha_{AT} = 4\,\mathrm{Å}^{-1}$, for the on-site potential, while $K = 0.03\,\mathrm{eV/Å}$, $\delta = 0.8\,\mathrm{Å}^{-1}$, $\rho = 3$ for the stacking potential. When we introduce the on-site barrier, the transition temperatures are shifted, and thus the energy units must be rescaled. In this case, $D = 0.0519\,\mathrm{eV}$, $G = 3D$, $y_0 = 2/\alpha$ and $b = 1/2\alpha^2$. As mentioned earlier, the sequence is set on the on-site potential, even though some recent studies prefer to do so in the stacking constants [121]. In our case the energy and length units are rescaled by a 1.5 factor, $D_{CG} = 1.5 D_{AT}$, $\alpha_{CG} = 1.5\alpha_{AT}$.

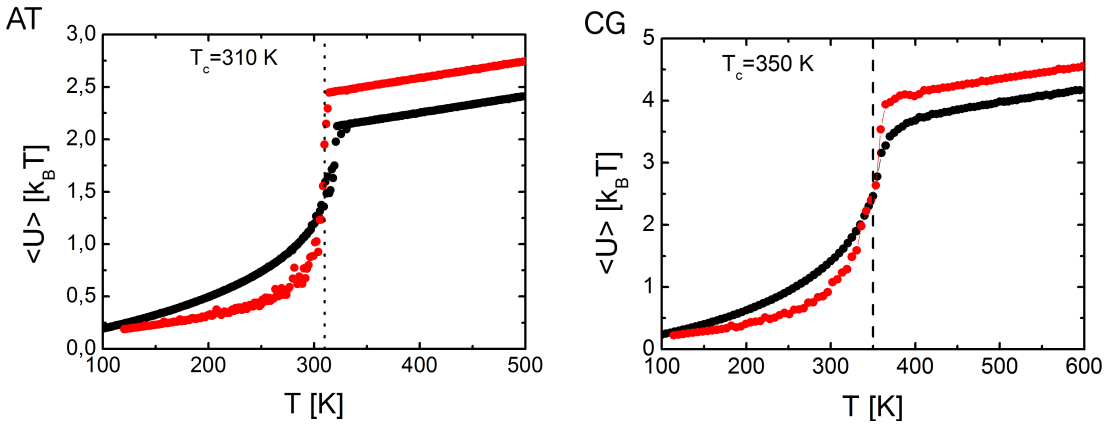


Figure 8.2: **Average energy versus temperature for A-T and C-G homogeneous sequences with and without solvation barrier**: Left panel shows results for and A-T chain of 220 base-pairs and right panel for a C-G chain of equal length. Black points show the curve for the original model with no barrier, while red poins. for the model with barrier. The effect of the barrier is a sharpening of the fase transition.

Figure 8.2 shows the average energy $\langle u \rangle$ as a function of temperature for AT (left) and CG (right) homogeneous sequences. The original model (black) and the

one with barrier (red) have been simulated with the parameter sets mentioned before. At high temperature, the behavior is that of a free gaussian polymer chain, with constant K. A remarkable feature is how the barrier narrows the melting transition, an effect which can be expected to occur when ρ is increased. In order to defined systematically the melting temperature, we use the following computational criterium. We define two temperatures, T_2, defined as the onset of the linear behavior in $\langle u(T)\rangle$, indicating that the chain is completely melted. T_1 estimates the beginning of the transition, and is defined as $\langle y(T_1)\rangle = y_0$, when the chain is on average on the barrier position. The transition width is thus defined as $\Delta T = T_2 - T_1$ and the melting temperature $T_m = \Delta T/2$.

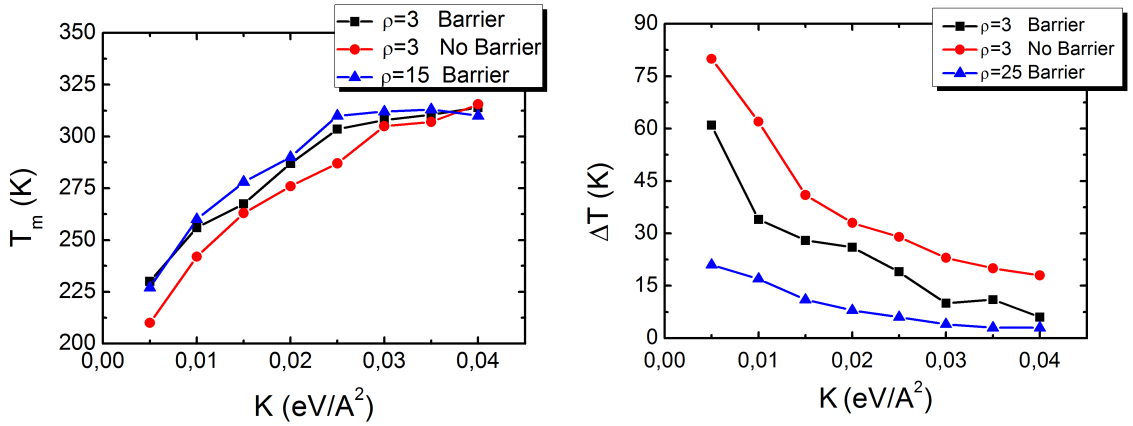


Figure 8.3: **Dependence of the melting temperature and the width of the transition as a function of the stacking constant** K: The effect of ρ and the barrier is insignificant on the melting temperature. Nevertheless, they have a dramatic effect on the width of the transition, decreasing greatly as we increase ρ or set the solvation barrier.

At this point, we discuss briefly the effect of different parameters on the melting transition and dynamical behavior of the system. For the purely harmonic PB model the transition temperature can be analytically computed as $T_m = 2\sqrt{2KD}/ak_B$ [100]. Nevertheless, numerical studies should be carried out for the current model if we want to arrive to equivalent conclusions. As already mentioned D, the Morse potential dissociation energy, affects directly on the value of T_m (higher D leads to higher T_m). The effect of the stacking parameters and ρ on the transition temperature T_m and width ΔT are plotted in Fig. 8.3. A first conclusion is that the non-linear coupling parameter ρ and the presence or absence of barrier do not affect significantly the value of T_m. The transition width ΔT is affected by both effects, as increasing ρ dramatically decreases this width, something also observed for the model with barrier, as already discussed. In this sense, a suitable melting temperature and transition width is obtained at high K values and moderate ρ. Too high ρ values, even narrowing the transition and thus reproducing more faithfully the melting phenomenology, produce too narrow bubbles that are unphysical [122]. With respect to the solvation barrier, its presence makes the bubbles last longer, and thus the complete separation of the strands is facilitated, decreasing the transition temperature. This effect is counter balanced by the increase in D.

8.3.2 PCA of the Phase Transition

PCA provides a powerful method for analyzing the DNA melting phase transition. As already mentioned, if our Hamiltonian involves just harmonic potentials, PCA is equivalent to normal mode analysis. For a uniform chain at low temperatures, the PDB model coincides with a linear chain of particles oscillating in a harmonic potential of frequency $\omega^2 = 2D\alpha^2/m$ and coupled by springs of constant $K(1+\rho)$. Thus, its dispersion relation reads as:

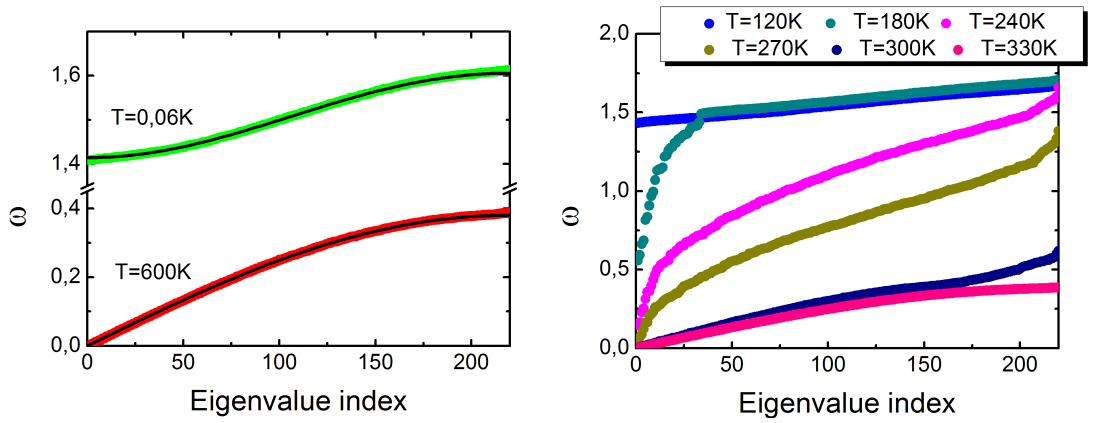


Figure 8.4: **PC frequencies spectrum at different temperatures**. *Left*: PC frequencies spectrum at very low (green) and very high (red) temperatures. Solid lines are the analytical approximation (Eqs. (8.2) and (8.3)). *Right*: PC frequencies spectrum at different temperatures between 120 and 330 K. The transition can be identified as the softest mode drops to zero. In both cases, frequency units are $(D/m)^{1/2}\alpha = 5.15 \times 10^{12} sec^{-1}$.

$$m\omega^2 \approx 2D\alpha^2 + 2K(1+\rho)[1-\cos(\pi n/N)]. \tag{8.2}$$

The principal frequencies correspond to the dispersion relation, while the principal eigenvectors to the normal modes of a linear chain of the mentioned features.

At high temperatures, the interbase potential is irrelevant and the PBD reduces to a free gaussian chain with coupling given by K:

$$m\omega^2 \approx 2K[1-\cos(\pi n/N)]. \tag{8.3}$$

Figure 8.4 (left) shows the principal frequencies plotted with the two analytical expressions, showing an excellent agreement. The intermediate behavior cannot be reproduced analytically. As temperature increases, the nonlinear excitation becomes more important, leading to larger fluctuations which are associated with larger PC eigenvalues, or the lower principal frequencies. Figure 8.4 (right) shows the principal frequencies for different temperatures. Remarkably, as we approach to the transition temperature ($T_m = 310K$ for the A-T chain) a soft mode goes to zero.

The evolution of this mode with the temperature can be used to analyze the melting transition. Figure 8.5 plots this mode as a function of temperature in logarithmic scale. The curve can be fitted using a critical behavior function $\omega \propto (T_m - T)^\nu$, with $T_m = 307K$ and $\nu = 2.2$. The value of T_m is very close to the transition temperature. The dynamic exponent cannot be related to any known family of models.

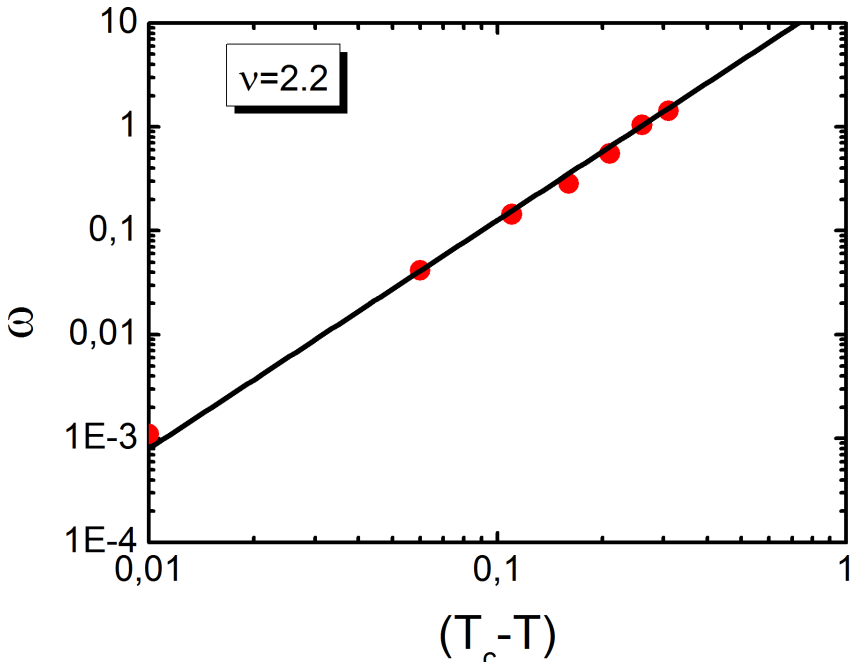


Figure 8.5: **Temperature dependence of the lowest PC frequency.** Fitting to critical behavior function $\omega \propto (T_m - T)^{\nu}$ we obtain $T_m = 307K$ and $\nu = 2.2$ (solid line). T_m shows great agreement with reported melting temperature, while the critical exponent cannot be ascribed to any known family of models..

8.3.3 Bubble Formation

As discussed through Chapter 8, bubbles have a key importance for the biological role of the DNA molecule, and their occurrence in particular sites along the DNA sequence seems to be related with the binding of some proteins. In this regard, PBD model can be used to understand the local dynamic proteins of the DNA molecule at the base-pair level. As the inclusion of the solvation barrier affects importantly to these properties (longer-lasting bubbles and less frequent) it seems an appropriate modification in order to reproduce bubble dynamics and ultimately to relate it with biological features.

Figure 8.6 compares two molecular dynamics trajectories without (upper) and with (lower) barrier. Each panel is made up of three different representations. The upper figure represents the whole trajectory with the base-pair opening y_n plotted in grey scale (white closed and black open). The two other figures show time and position snapshots.

The effect of the solvation barrier on the strand dynamics is clearly demonstrated here. At a given time, the opening profiles are quite similar, meaning that the bubble length (which spans around $\sim 15 - 20$ base pairs in agreement with the biological sizes) is not greatly affected by the inclusion of the barrier. The dynamic profiles are drastically different. Without barrier, base-pair openings correspond to large amplitude oscillations along the Morse potential, where an easy closing is favored. Thus, bubbles are easily formed, but also easily closed. With barrier, the kinetics is controlled by the presence of two different equilibrium states, separated by the salvation barrier. Upon closing, the energy barrier must be overcome and thus

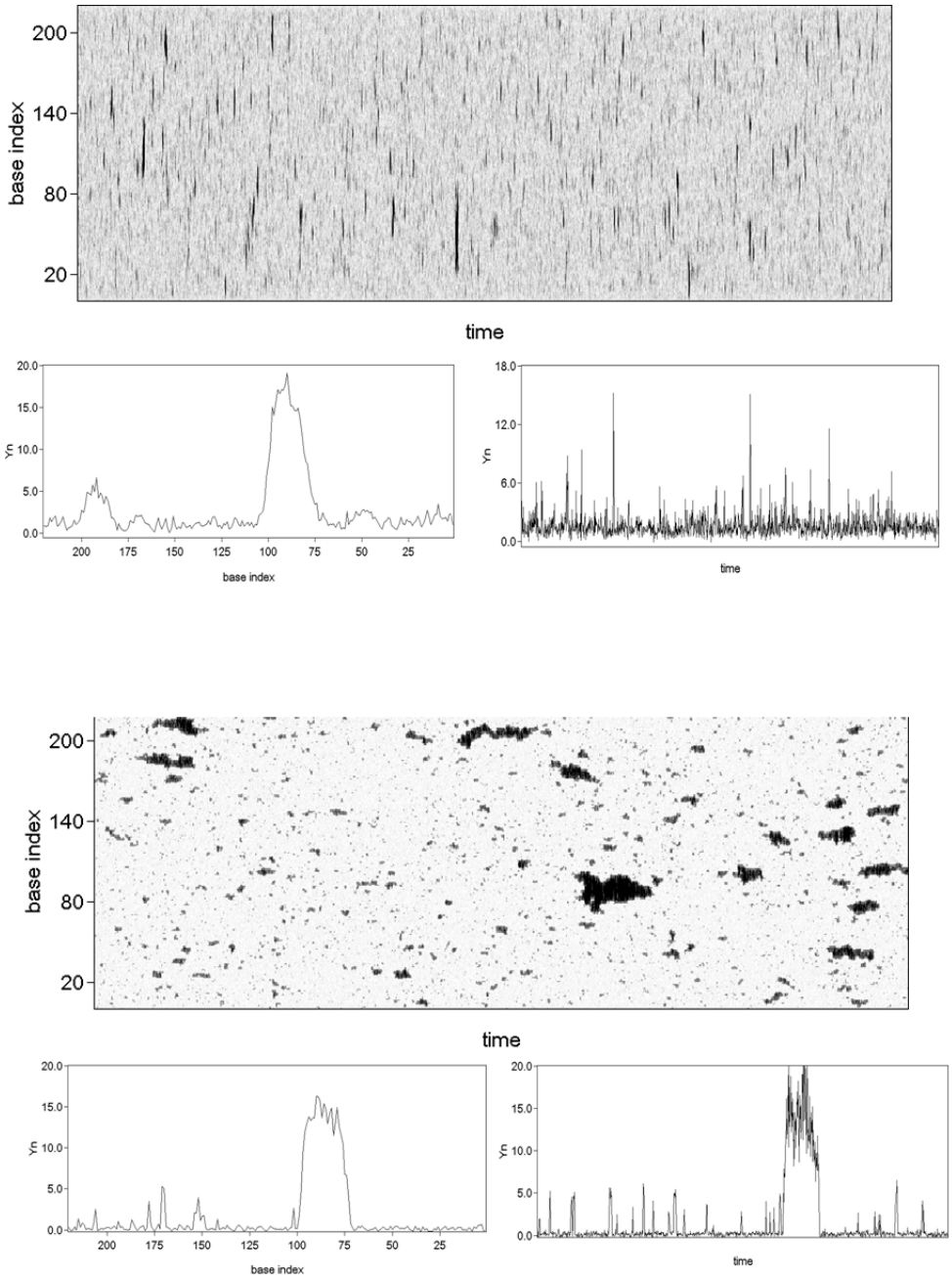


Figure 8.6: **Typical simulation trajectories for a homogeneous AT sequence without barrier (upper panel) and with barrier (lower panel).** The y_n coordinate is plotted in gray scale from white (closed) to black (open). Smaller figures correspond respectively to a time slice (left) and trajectory of a single base-pair (right). The salvation barrier reduces the bubble frequency although it stabilizes them, once they open. Trajectory time is 200ns and y_n is given in units of $\alpha^{-1} = 0.25\mathring{A}$.

bubbles live longer.

This bubble lifetime approaches better to experimental values, reported to be of few tenths of ns [95]. This longer-lasting openings are necessary to drive protein binding upon transcription, replication or regulation processes. Nevertheless, from a computational perspective, this behavior requires longer simulations in order to

gather up the necessary statistics, as bubbles become a rare event.

At this point it is worth to make a commentary regarding the effect of the ρ parameter on the chain dynamics behavior. As discussed already, ρ affects the cooperatively of the model leading to a narrower transition, similar to the effect of the inclusion of a salvation barrier. Nevertheless, large values of ρ lead to long-living bubbles but extremely narrow (around one or two base-pairs, see [122]). In this sense, our choice of $\rho = 3$ including a salvation barrier is good enough for obtaining longer-lasting bubbles but wide enough for reproducing the known phenomenology.

8.4 The P5 Promoter Sequence

Several works have explored the possible link between the formation of these bubble openings and the presence of specific binding sites of regulatory proteins in DNA sequences. For instance, the RNA polymerase binds to the so-called Transcription Starting Site (TSS from now on) located at position +1 in the promoter region prior transcribing the coding region. Additionally, different transcription sites bind to specific sites on the promoter region regulating the transcription process. This process is rather complex and varies from one organisms to another, or even from gene to gene.

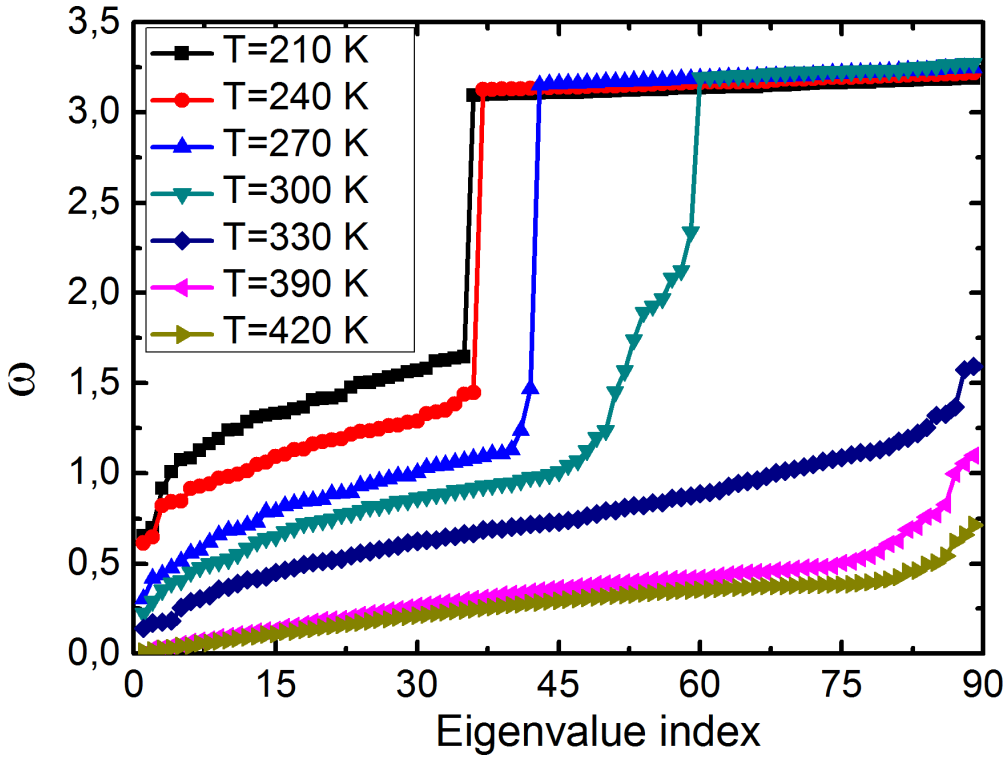


Figure 8.7: **PC frequencies spectrum at different temperatures between 210 and 420 K.** Two bands are identified at low temperatures, corresponding to CG and AT base pairs. Frequency units are $(D/m)^{1/2}\alpha = 5.15 \times 10^{12} sec^{-1}$.

It has often been argued, that local physical properties of the DNA molecule, related with the formation of spontaneous bubbles, are highly correlated with the binding of proteins such as Transcription factors [123–125]. In this sense, the PBD model can be as a useful tool for analyzing promoter sequences, subject to its correct reproduction of the bubble dynamics behavior of the DNA molecule at the

desired level. Particularly, some studies show the correlation between sites with high propensity to form bubbles in the PBD model context, and the position of protein binding sites [59, 97, 126–129] This topic is further discussed in Chapters 9 and 10.

In this section, we analyze a fragment of the adenoassociated viral P5 promoter (P5), widely studied in different works [109, 118]. This promoter is known to contain two main sites exhibiting frequent bubble openings within the context of the PBD model, namely the TSS (+1) and an A-T rich region between -40 and -35, corresponding to the binding site of the transcription factor Ying Yang 1 [118].

We run Langevin simulations at different temperatures on the sequence of the P5 promoter is given by the 69 bp: 5'-GTGCCCATTTAGGGTATATATGGCCGAGTG AGCGAGCAGGATCTCCATTTTGACCGCAAATTTGAACG-3'. In order to isolate this sequence and avoid finite size effects, we add a base-pair clamp of 10 C-G at each of the ends of the promoter, creating "hard" boundary conditions. The first and last base pairs are forced to remain closed in order to avoid a complete opening of the chain.

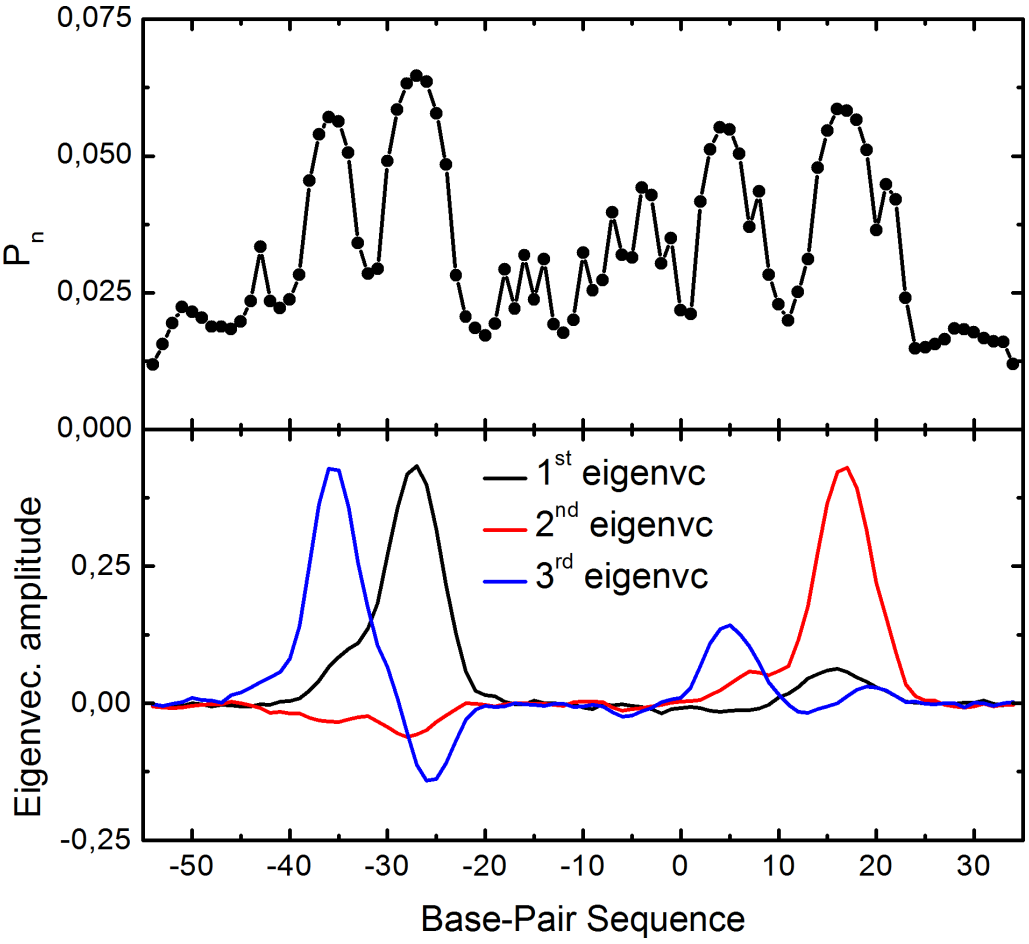


Figure 8.8: **Opening probability (upper) and first three eigenvectors (lower) for sequence P5.** Two clear opening regions with each two different bubbles are clearly identified. Both are settled at biologically relevant spots in the P5 sequence, namely the TSS and the -35 box.

The interest of our work here is twofold. First, we use PCA to analyze the melting transition of this heterogeneous sequence and compare it with the behavior of the homogeneous sequence. Second, we study the behavior of the chain at a fixed tem-

perature, bellow the melting transition, in order to identify those sites with larger nonlinear fluctuations (bubbles) and relate them with the known regulatory sites. We use PCA as a valuable tool for such goal. Whereas forhomogeneous sequences PCA show the normal modes of a chain of oscillators (including nonlinear contributions), in heterogeneous sequences, the eigenvectors show localized contributions at specific sites in the sequence where large amplitude motions are more significant.

Figure 8.7 shows the PC frequency spectrum for different temperatures. This observable is used to identify the phase transition. We observe also some features which distinguish this profile from the one in the homogeneous case (Fig. 8.4). At low temperatures, two bands are identified. The higher one corresponds to the CG base pairs, while the lower one to the AT base pairs. At intermediate temperatures, this gap begins to disappear, as CG pairs surrounded by AT pairs are more likely to open.

Close to the transition ($T \approx 345K$ in this case), several modes detach to low frequencies. These modes correspond to localized regions in the sequence contributing greatly to the fluctuations of the system, and they can be related with zones with high probability of opening. Figure 8.8 shows the opening probability profile (upper) and the first three eigenvectors (ordered according to the eigenvalues) at $T = 290K$, where these features are clearly depicted. We identify two major regions In the probability profile which show each two different peaks with a larger probability of opening compared to the average behavior. These sites correspond respectively to two different regulatory regions, the TSS located at site +1 and the -35 site, binding site of the Ying Yang 1 factor [118].

The first three eigenvectors validate this picture, with an excellent correlation with the probability profile. Each eigenvector shows a highly localized contribution—spanning around 10 base pairs—to the fluctuations of the system. The first and third one show decorrelated contributions in the -35 region while the second one contributes in the +1 to +15 regions. The PCA vision can be useful also to find correlation between global movements of our system. For example, if we focus on the first eigenvectors, the largest contribution is clearly centered around the -25 base pair, but a smaller peak appears also in the +15 region. This means that fluctuations in both sites are slightly correlated in a positive way. On the contrary, the third eigenvector shows negative correlations between the two bubbles that might be formed in the -35 region. This correlations might be of great interest, as modification (mutations) in one of them, could interfere with the behavior showed by other one.

Figure 8.9 plots a typical trajectory on the P5 promoter. Opposite to the case of the uniform chain, where bubbles formed at arbitrary positions along the sequence, here they form mainly at the sites identified by the PC eigenvectors. We are working at $T = 0.85T_m$ and, as the transition is quite sharp, we observe few openings and thus long simulation runs are needed to gather enough statistics. Nevertheless, PC analysis gives good account of the fluctuating regions even though opening are rare events.

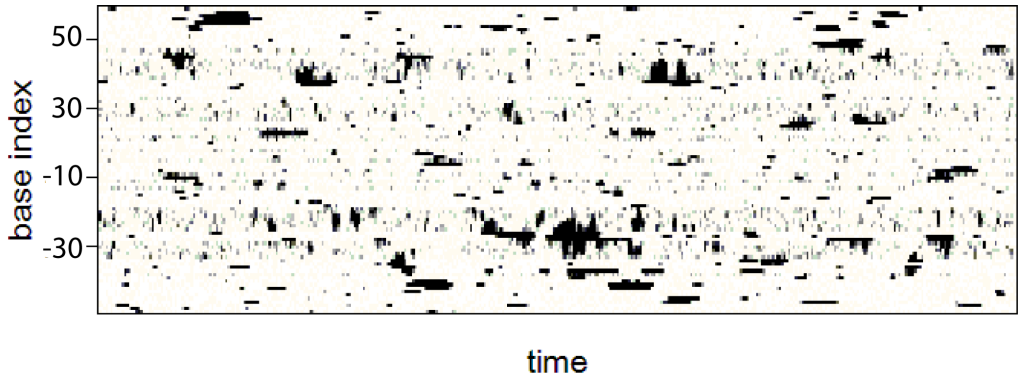


Figure 8.9: **Trajectory of the P5 promoter at $T = 290K$.** Bubbles form mainly around the two biologically relevant regions, already identified by PCA. The simulation time shown is of 400ns.

Chapter 9

A Model for Protein-DNA Interaction at the Base-Pair Level: Analyzing Promoter Sequences with a Mesoscopic Model

In this chapter we introduce a model for protein-DNA interaction [58]. The DNA is represented with the PBD model, and the protein is introduced as a new degree of freedom, which interacts with the chain coupling to open regions. We apply the model to three different promoter sequences, two of them with high RNA production levels (strong promoter), and the other with low production levels (weak promoter). We locate and quantify the binding sites by employing a suitable analysis model, based on Markov state models. These binding sites are correlated with biological relevant regions in the promoters we analyze. By employing a suitable analysis method based on a Markov state model description of the system, Additionally, we distinguish between the strong and weak promoters by analyzing the structure of the free energy landscape of the system.

9.1 Motivation for Developing the Model

The complexity of genetic regulation is very high, and to date it is far from being understood at a molecular level. There is a large number of proteins involved which play different roles, and this changes from organism to organism. In this sense, to propose a general mechanism for protein-DNA interaction seems rather a utopia. Additionally, the classical biochemical picture of "protein A recognizes site B, and binds to it" is a *naive* consideration, especially if we account for the crowded and noise-ruled environment where molecular biological processes take place.

Here, we focus on a particular process, with no intention of generality. It has often been reported, that *some* DNA-interacting proteins couple to the physical properties of the DNA sequence, particularly to bubbles [123–125]. For example, RNA polymerase binds to the so-called Transcription-Starting-Site (TSS) where the transcription bubble starts. Other regulatory sites, such as the TATA box, are known be locally "weak", and thus easier to open and unwound [96]. Here, we explore this aspect, based on the concept that some proteins bind to particular regions in the DNA sequence where bubbles form spontaneously with more facility.

In other words, the presence of regions more likely to be opened are correlated with some protein binding sites, and thus this dynamics have a key role on regulatory processes.

Starting with PBD model, we move a step further and propose a model at the mesoscopic level for the one-dimensional diffusion searching process of a generic regulatory protein[1] [130, 131]. This protein slides along the DNA molecule interacting stronger with open regions in the chain. In turn, it also helps to break the base-pair bonds, opening bubbles, and stabilizing open ones. By analyzing the combined dynamics of the DNA base-pairs and the particle, we can identify the most prominent states the system populates, where the particle bounds to a particular site where a bubble is formed and stabilized. These states will be correlated with already known biological relevant sites, namely protein-binding sites.

We can interpret this model in an additional way. The generic particle we introduce can be seen as a sounding line which runs along the DNA sequence, detecting "softer" sites, which likely can be considered as potential protein-binding sites. This proposal is similar to flexibility maps that are measured with Scanning Probe Microscope on surfaces or even biological molecules [132].

We propose a suitable analysis method which allows to define systematically the states the system occupies through the dynamics and quantify them from an statistical point of view. We are able to compare the importance of such states in the dynamics, and to relate them with the "strength" of the binding sites, in terms of RNA production [133]. This procedure cannot be applied directly on simulations of the PBD model alone, as bubbles are rare excitations of the ground state, where all base-pairs are closed. The introduction of the diffusing particle changes dramatically the free energy landscape of the system, allowing for a richer behavior and further biological consequences.

9.2 Description of the Model

Our model is made up of two ingredients, the DNA chain and the interacting protein. A schematic picture of it can be seen in Fig. 9.1. The DNA chain is described by the modified PBD model presented in Chapter 8, while the interacting protein is modeled as a generic particle undergoing a one-dimensional diffusion along the chain. This particle is coupled to the chain's opening profile in the sense that it interacts more strongly with open base pairs. In turn, the presence of the particle also affects the dynamics of the DNA base-pairs, as it tends to open the chain. The model proposes thus a two fold effect. Softer regions of the DNA sequence are more likely to be visited by the protein, which also helps in opening them and stabilizing the bubbles.

Briefly, the DNA chain is described by the PBD hamiltonian $\mathcal{H}_{_{DNA}} = \sum_{i=1}^{N}[\frac{p_n^2}{2m} + V(y_n) + W(y_n, y_{n-1})]$, where $W(y_n, y_{n-1})$ is the stacking potential given by Eq. (7.3) of previous chapter, and $V(y_n)$ the on-site Morse plus barrier potential given by Eq.

[1] The most accepted mechanism for location of targets in a three-dimensional diffusive environment is a combination of one-dimensional diffusion stages along the DNA chain, and three-dimensional jumps which allow to reach new regions chemically far. This is the process known as facilitated diffusion. Indeed, most of the time is spent in the one-dimensional stage, being the time involved in the three dimensional jumps almost negligible. In this sense, it is reasonable to focus on the former process.

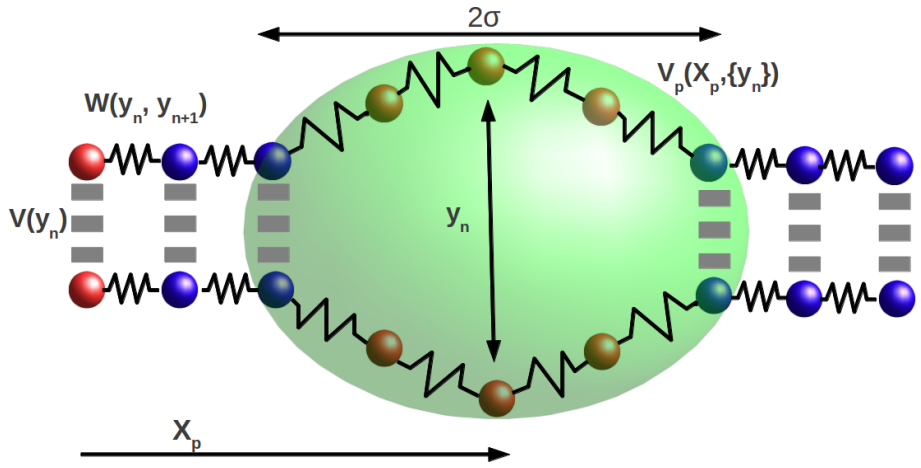


Figure 9.1: **Schematic picture of the protein-DNA interacting model**: The protein or generic particle (big green ball) diffuses along the DNA chain (represented as small balls with springs), interacting with it trough potential term V_p. This interaction increases with the base-pair separation y_n, in such way that the protein couples to bubbles. In turn, the base-pairs are also affected by the protein dynamics, as it also pulls them out of the equilibrium position, helping in opening bubbles and stabilizing them, once open.

(8.1).

Now, the interacting protein is represented by a Brownian particle (see Fig. 9.1) moving through a one-dimensional space with coordinate X_p and interacting with the DNA chain through a phenomenological potential which depends on the coordinate X_p and the instantaneous opening profile given by $\{y_i(t)\}_{i=1}^{N}$. The Hamiltonian for the particle reads:

$$\mathcal{H}_P = \frac{p_p^2}{2m_p} + V_{int}(X_p, \{y_i\}), \tag{9.1}$$

where suffix p stands for protein. Mathematically, we set the following interacting potential:

$$V_{int}(X_p, \{y_i\}) = -\frac{B}{\sqrt{\pi\sigma^2}} \sum_i \tanh(\gamma y_i) e^{-(X_p - ia)^2/\sigma^2}. \tag{9.2}$$

This expression is simply a sum of gaussian wells centered at each base pair, which are separated a distance of a in the coordinate X_p (units along this degree of freedom are arbitrary). Each well has a maximum interacting amplitude of B, where σ is the spatial range of such interaction. The interaction amplitude depends on the state of the $i-th$ base pair, with a $\tanh(\gamma y_i)$ term. This functional form is chosen so that the interaction amplitude is linear with y_i at low openings, and saturates when $y_i \sim \gamma^{-1}$, in order to avoid and indefinitely opening and thus self-trapping of the protein. With this interaction term, the particle tries to open the chain in a length range of σ and with an intensity proportional to the opening profile.

The system can be interpreted as a particle diffusing through a classical field which depends on the instantaneous configuration of the DNA chain $\{y_i(t)\}_{i=1}^{N}$. In this sense, bubbles create wells in the profile and thus, the particle would tends to dwell at these regions. Figure 9.2 illustrates this by plotting an arbitrary average configuration of a DNA sequence within our model, and the potential profile V_{int} created by such profile. Nevertheless, we should remark that the particle itself

is affecting the configuration of the DNA strands, and thus the potential profile through which it is moving changes with time.

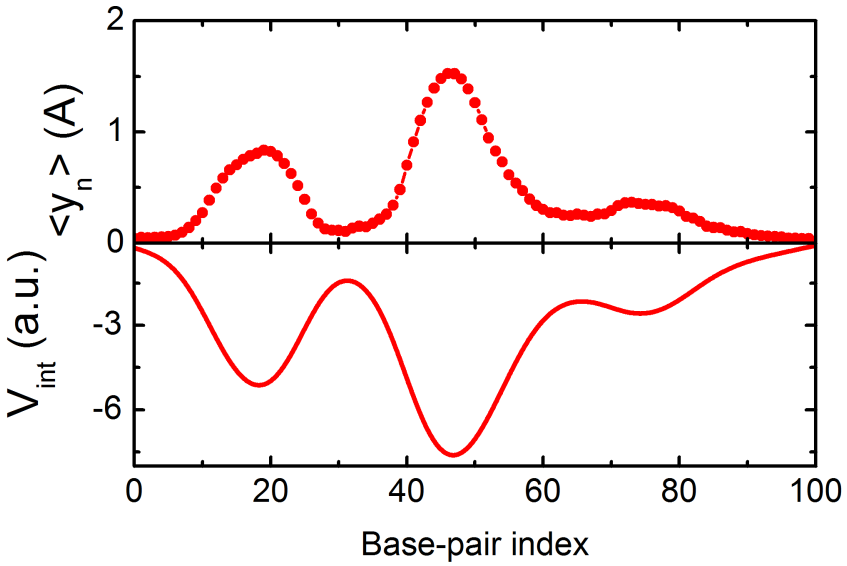


Figure 9.2: **Average configuration of a DNA sequence and associated field created by the interacting protein**: Upper panel shows an average configuration of a DNA sequence and lower panel the instantaneous profile $V_{int}(X_p)$ calculated for such configuration $\{y_n\}_i^N$.

The diffusing protein is a larger entity than the individual base pairs, so we set a higher damping and mass, $\eta_p = 10^{14}\text{s}^{-1}$ and $m_p = 7000\,\text{Da}$, in the order of magnitude of DNA-binding proteins [134]. The intensity of the interaction is chosen as $B = 20k_BT$, providing local interactions of the order of the Morse potential dissociation energy at each base pair. With $\gamma = 0.8\,\text{Å}^{-1}$ the potential saturates at $y = 1.25\,\text{Å}$, which is the position over the on-site barrier. The new degree of freedom X_p has arbitrary units, as a is the base-pair separation along a chain, so meaningless for our purpose here. We set $a = 1$ and $\sigma = 3$, providing an interaction range which spans over 5-6 base pairs. With these interaction parameters, and due to the cooperativity of the model, we observe bubbles of around 15-20 base pairs, which is a typical value for the processes we want to model here [135].

9.3 Simulation Details

The dynamics of the model are defined by the Langevin equations for the base pairs and the particle. For the n-th base pair of the chain, we have

$$m\frac{\partial^2 y_n}{\partial t^2} + m\eta\frac{\partial y_n}{\partial t} = -\frac{\partial[W(y_n, y_{n-1}) + W(y_{n-1}, y_n)]}{\partial y_n} - \frac{\partial V}{\partial y_n} - \frac{\partial V_{int}}{\partial y_n} + \xi_n(t), \quad (9.3)$$

where η stands for the damping and ξ_n for the white thermal noise, so $\langle\xi_n(t)\rangle = 0$ and $\langle\xi_n(t)\xi_k(t')\rangle = 2m\eta k_BT\delta_{nk}\delta(t-t')$ hold.

The protein moves following

$$m_p \frac{\partial^2 X_p}{\partial t^2} + m_p \eta_p \frac{X_p}{\partial t} = -\frac{\partial V_{int}}{\partial X_p} + \xi_p(t), \tag{9.4}$$

where η_p stands for the particle damping and ξ_p for white thermal noise. Analogous fluctuation-dissipation relations stand here.

We integrate numerically previous equations following a fourth order stochastic Runge-Kutta algorithm [81], obtaining a set of $N+1$ molecular trajectories, for the N base pairs and the particle. Each of the sequences is simulated in five different realizations for 40 µs, using a time step of 10 fs and a 1 µs preheating time. These numbers are in agreement with the typical one-dimensional diffusing times for the kind of processes studied here [134]. The simulation temperature is $T = 290K$.

The protein diffuses along the DNA chain with periodic boundary conditions, while we set fixed boundary conditions to the DNA chain, adding a base pair clamp of 10 CGs at the ends of the sequence of study to create "hard" boundary conditions, as discussed in [58, 116].

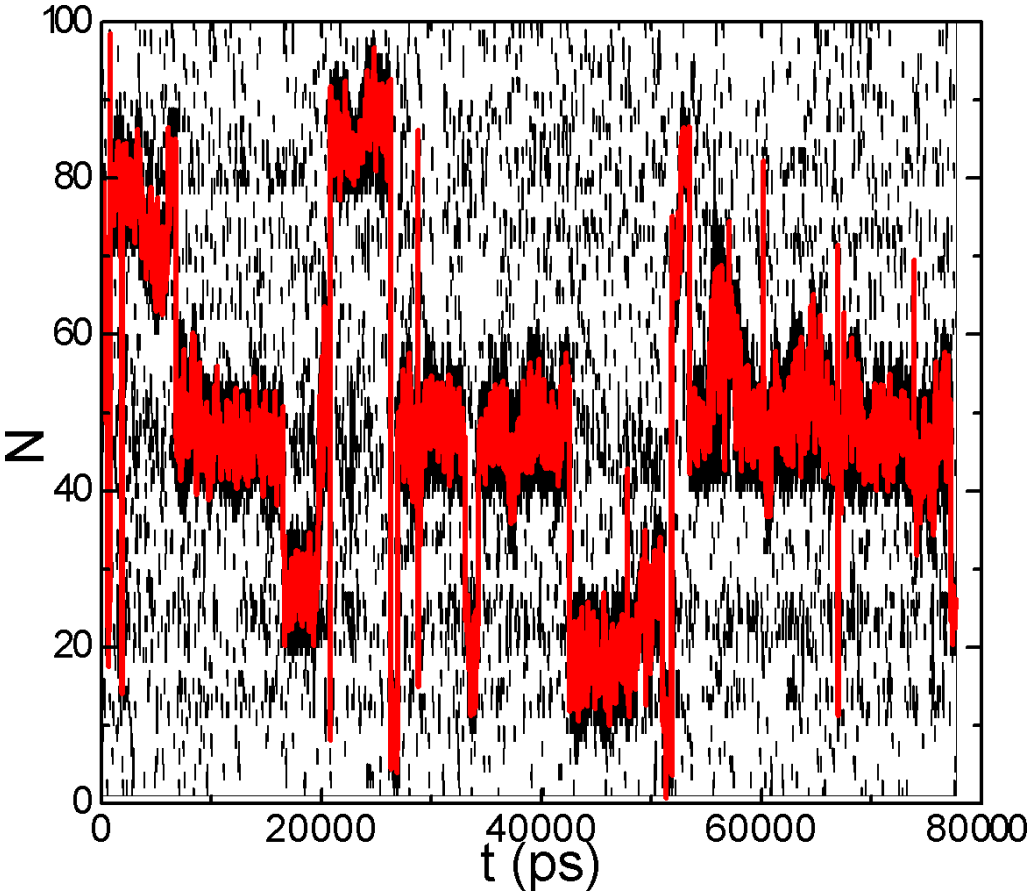


Figure 9.3: **Trajectory of the N base-pairs of a DNA sequence in white (closed)-black (open) code, with superimposed trajectory of the interacting protein (red)**: The dynamics of the DNA base-pairs depends on the position of the particle. In the absence of particle, they show regular dynamics with transient openings. The particle influences greatly this behavior, opening and stabilizing the bubbles, which last for at least two orders of magnitude more. The dynamics of the particle can be described a jumps between different regions where it dwells for a while, opening a long-lasting bubble.

Figure 9.3 shows an example of a trajectory simulated with our model. The base pairs are pictured in a white (closed)-black (open) scale. The particle trajectory

is plotted in red. The particle diffuses along the whole sequence, jumping between different regions where bubbles open. Its presence helps in stabilizing these bubbles, which last considerably longer compared to the PBD model alone (see Chapter 9 or compare with Fig. 8.9). We observe how the dynamics of the base pairs affect strongly the protein dynamics, and the other way around.

9.4 Analysis: Brief Reminder about Markov State Models

We take advantage of Markov state models to analyze the joint trajectories of the N base-pairs plus the particle, and identify systematically states the system visit, and be able to get quantitative information about them. The analysis procedure is quite similar to the one described and discussed in Part I, nevertheless, we are less detailed about the Markov state model validity, focusing on the direct output this technique yields to us. We describe briefly the analysis procedure in the next section. More details about Markov state models can be found in Chapter 2.

9.4.1 Description of the Analysis Protocol

The main steps of the analysis protocol here are: i) Calculate PCs of the chain trajectory to reduce the dimensionality of the system. ii) Calculation of the Conformational Markov Network by discretizing the chosen number of reduced trajectories. iii) Definition of the basins of attraction by clustering the Markov Network with the Stochastic Steepest Descent Algorithm. iv) Build the disconnectivity graphs (or dendrogram) to visualize the free energy landscape. Briefly, we describe each of the steps.

1. **Obtention of the reduced trajectories**

   First, we apply PCA just to the N base-pairs trajectories (omitting the particle) in order to reduce the dimensionality of our system. We project the trajectories $y_i(t)$ onto the principal eigenscapeces in order to define the PCs $q_i(t)$,

   $$q_i(t) = \mathbf{v}_i \cdot (\mathbf{y}(t) - \langle \mathbf{y} \rangle), \tag{9.5}$$

   where $\mathbf{y}(t)$ is the time dependent trajectory written as a vector, $\langle \mathbf{y} \rangle$ the average position of the base pairs as a vector, and $\mathbf{v}_i$ the i-th principal eigenvector.

   We describe our system by keeping just the first five PCs. With the restriction to this subspace, we account for the 75% of the total autocovariance of the system. This is plotted in Fig. 9.4 where the cumulative fluctuation rate ζ_i (or the cumulative trace of the correlation matrix) is plotted, defined as, $\zeta_i = \sum_j^i \lambda_i / \operatorname{tr} \tilde{C}$, where λ_i are the eigenvalues of the correlation matrix $\tilde{C}$.

2. **The Conformational Markov Network**: Next step is to map the trajectories onto a complex network. We calculate the microstate Markov network by discretizing out conformational space into bins for a particular lag time Δt or sampling time. The configurational space is defined by the five first PCs plus

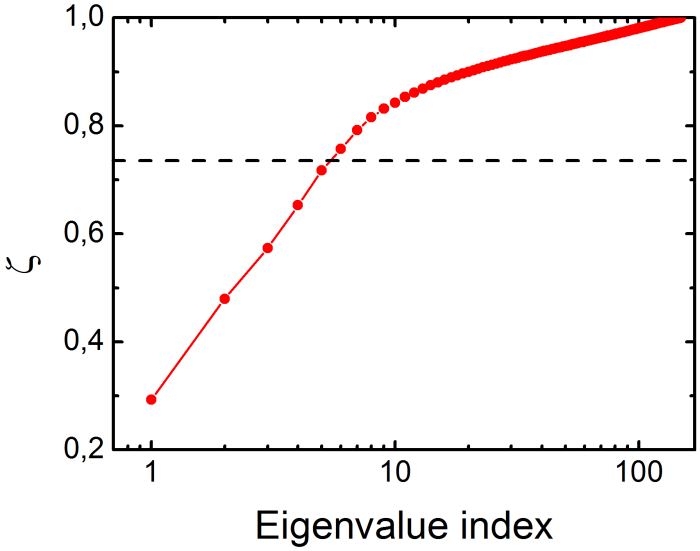


Figure 9.4: **Cumulant autocovariance for the N base-pair trajectories**: The first five eigenspaces gather up to the 70% of the total fluctuations of the system.

the trajectory of the particle. We discretize the five PCs into 20 bins of equal volume and the trajectory of the particle into N bins, corresponding to the domain of each base-pair. This defines a microstate space with a maximum of $20^5 \times N \sim 10^8$ microstates. Next, we map the continuous trajectories into the "bin" trajectory, in order to calculate the transition matrix T_{ij} and the occupation vector π_i.

3. **Stochastic Steepest Descent: obtaining the Basins of Attraction** We cluster the microstate network to define the macrostate network. We use the *Stochastic Steepes Descent* Algorithm (SSD) [32, 56] for this purpose, finding the basins of attraction in the microstate network (see Chapter 2).

4. **Free Energy Landscape as a free energy dendrogram** We represent the Free Energy Landscape as a hierarchical tree diagram, or dendrogram [6]. This representation is built according to the weights and links among basins (see Chapter 2).

   The dendrogram representation provides a qualitative and quantitative hierarchical organization of the basins in terms of free energy and the barriers that separate them in the free energy landscape. Here, we can use them to understand how do the different states in the sequence organize and characterize them from a quantitative point of view, according to different thermodynamic quantities which might be calculated in an straightforward way, thanks to this representation.

9.4.2 Characterizing the Configurational Space

We translate the trajectories (or set of trajectories) on a network of basins of attraction which represent the macrostates of our system. Our initial goal is to provide a solid tool for analyzing promoter sequences in order to find possible binding sites

of DNA-interacting proteins. Each identified macrostate corresponds to a particular configuration of the N base pairs and a position of the particle along the chain. Likely binding sites should correspond to bubbles at concrete regions, with the particle bound there. Our analysis method provides, not only a way to define systematically such states, but also the possibility of characterizing them.

A certain macrostate α is defined by its weight $\pi_\alpha = \sum_i \pi_i$, where $i \in \alpha$ are all nodes from the microstate network belonging to basin α. This weight represents its population in the network, and thus can be related to its relevance from a biological point of view. In the same way, we define the entropy of each basin as $S_\alpha/k_BT = -\sum_i \pi_i \log \pi_i$, with $i \in \alpha$. Intuitively, the entropy of a state means how "wide" the free energy basin is. For example, macrostates made of few very weighty microstates, would lead to low values of S_α, and thus narrow free energy wells, while those made of lots of microstates with lower weight would yield to large entropies, meaning wide wells, even if π_α is the same in both cases.

At this point, we make a distinction between two different categories of states, *specific* and *nonspecific* states. *Specific* configurations represent stages in the dynamics, when the particle is bound to a concrete region in the sequence which is open. On the contrary, *nonspecific* states represent those intermediary stages where the particle diffuses along the sequence without binding to any particular site. Considering the one-dimensional searching process of a particular protein along the DNA strand, these two categories resemble the two different phases in the searching process.

In our context, we define as *specific* states those basins with a weight $\pi_\alpha \geq 10^{-3}$, while *nonspecific* states correspond to the remaining low-populated states. In this sense we define a *nonspecific* basin by clustering all states with $\pi_\alpha < 10^{-3}$. Its weight is $\pi_{NS} = \sum_\alpha \pi_\alpha$, where $\alpha : \pi_\alpha < 10^{-3}$. This *nonspecific* basin is employed as a reference state for defining our quantitative description of the *strength* of the basins, defined as the free energy difference with this state $\Delta F/k_BT = -\log(\pi_\alpha/\pi_{NS})$. In addition, the value of π_{NS} is used as a measure of the *strength* of a particular promoter, as high values of π_{NS} mean that the particle spends a significant fraction of the trajectory dwelling along the chain without binding to particular sites, and thus, the present binding sites are weak.

9.5 Results

We analyze three different promoter sequences of different organisms. Two of them—the already studied P5 viral promoter [136] and the human collagen type I $\alpha 2$ promoter [126]—, correspond to the so-called strong promoters, while the other one—lac operon regulatory region [137]—is a weak promoter. This distinction between strong and weak promoters is related with the level of expression in mRNA, in the sense that strong promoters show higher levels, and their sequences are closer to the consensus sequence [133, 138].

We seek for structural differences in the free energy landscape of the promoters, as interpreted with our analysis method. Specific binding sites should appear as more populated (lower free energy) in the strong promoters than in the weak one.

9.5.1 Control Sequence: Study of a Random Sequence

As a preliminary step, it is worth analyzing a sequence with no biological content as a "control" case. For this purpose we use a random sequence obtained by shuffling the base pairs of the P5 promoter sequence we employ later for our study. This randomization of the sequence removes any biological information contained initially in the sequence, while keeping the A-T/C-G content.

Figure 9.5 plots the three typical representations we will employ for our analysis. From left to right, we have i) the average position of the base pairs and the first two PCA eigenvectors; ii) the free energy dendrogram; iii) the weight distribution of the basin network. First, the opening profile provides little information, as the chain shows relatively large bubble spanning almost the whole chain. The PCA eigenvectors support this vision as they look similar to a homogeneous sequence, providing a delocalized contribution to the fluctuations of the system.

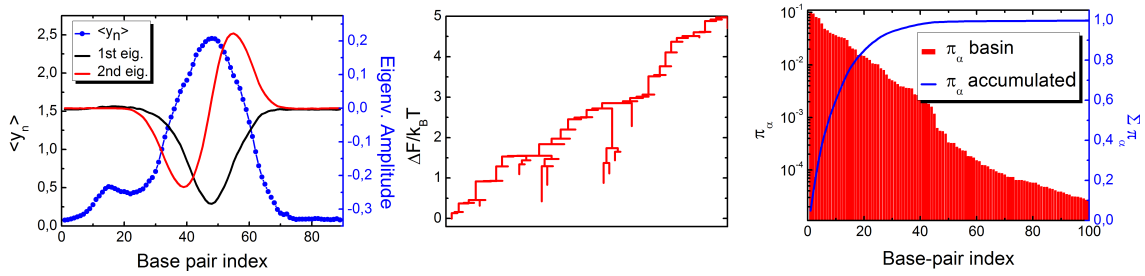


Figure 9.5: **Analysis of the random sequence**: Left panel shows the average opening profile of the chain and the first two eigenvectors. Clearly no structure is seen. Central panel shows the free energy dendrogram, where no prevalent states appear, as the weight is rather distributed through the nodes. Right panel shows the distribution of the weights of the basins and the cumulant weight, where a large number of basins share a significant amount of population. Compared to real sequences, we will see how the features are totally different.

The basin network has 8388 basins (we can compare this size with the ones of the promoter sequences shown later on), and with a structure which differs to the ones we find for biological sequences. The "background" or *nonspecific* basins suppose a 6% of the total network weight, while it can be seen that quite a large number of basins retain a significant fraction of the trajectory. This means that the network structure is quite distributed onto a lot of states, which have in turn a low free energy with respect to the *nonspecific* state. In this sense, basins identified by our algorithm on a random sequence correspond to very "weak" binding sites.

This random sequence could be conceived as the "weakest" promoter, provided that it lacks of any biological information. When comparing the structure found for each of the promoters analyzed in this chapter, the differences will be clearly identified.

9.5.2 Analysis of Three Promoter Sequences

We show the results for the three studied promoter sequences, the strong promoters viral P5 (given by the 69-bp sequence shown in [116, 118]) and human collagen type I $\alpha 2$ chain (given by the 80-bp sequence shown in [126]) and the weak promoter lac operon regulatory region, given by the 129-bp sequence taken from [137].

Figure 9.6 shows a detail of the low energy region of the free energy dendrograms for the three promoters (upper panel). Lower panel shows a representation of the physical state of different relevant basins identified by our method. These states are related to excitations in biologically relevant regions such as the TSS or the TATA box.

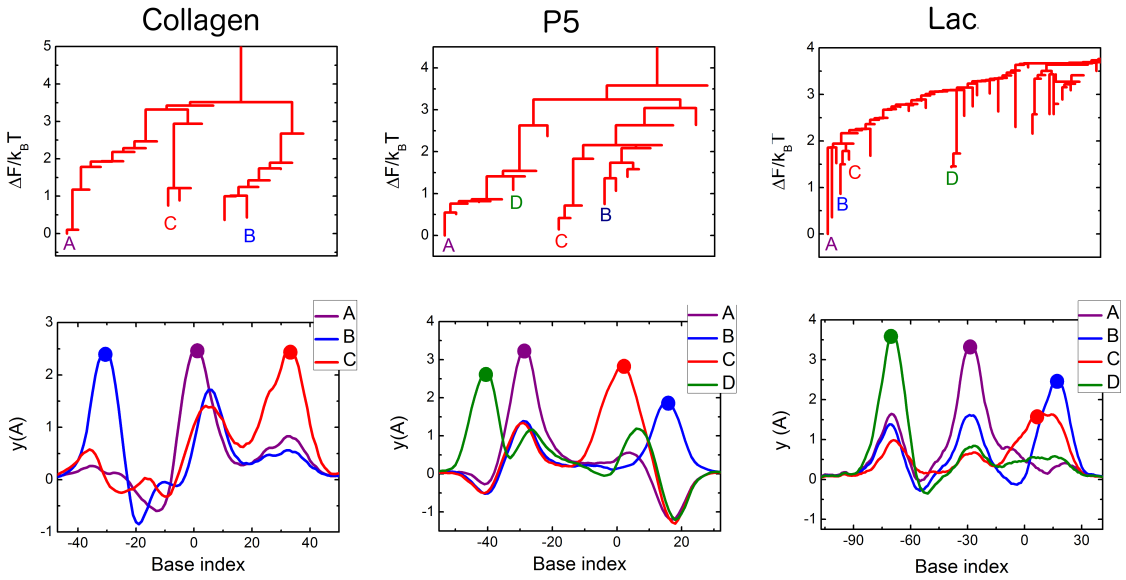


Figure 9.6: **Free energy dendrograms and representative states for the three studied sequences**: Upper three graphs show the free energy dendrograms for the collagen, P5 and Lac sequences, and lower three ones the most relevant states, as highlighted in the landscapes. We can see that in any case they correspond to bubbles in a particular spot in the sequence, with the particle located on it. Also, the topology of the dendrograms is rather different comparing P5 and Collagen promoters with Lac one.

For the collagen sequence, state A, the most populated one, corresponds to the TSS, showing a large bubble around base pair +1, and with the particle located there. States B and C are linked to excitations in other regions such as the TATA box at −35 position (state B). These states match those reported previously in [126].

In the same way, P5 sequence shows two major groups of states, each with two different particular states (A-D and B-C), in coincidence with the findings showed in Chapter 8. State C corresponds to the TSS , while state A to the -35 regions. Lac operon promoter identifies the TSS as basin C, with another state (B) close to it. Nevertheless, the overall behavior of this promoter differs from the first two, as we shall discuss later.

The free energy dendrograms show a radically different structure when comparing the P5 and collagen promoters with the lac operon one. The collagen dendrogram is structured into three main branches, associated with the three physical states depicted in Fig. 9.6. The P5 dendrogram shows an analogous structure, with two main branches, one of them divided in two more, corresponding to states B and C, which are kinetically close (barrier of $\sim 2k_BT$ separating them). The remaining states correspond to very similar configurations, where the conformation of the base pairs or the position of the particle differs slightly.

The dendrogram from the lac promoter is arranged into a completely different fashion. While the two strong promoters showed few prominent basins ($\sim 10-20$) and a large number of low populated ones (not shown in the dendrograms),

the weight in the lac promoter is more distributed, with several basins showing intermediate occupancies. In this sense, the free energy dendrogram is made up of several basins with similar weights.

In order to visualize this difference more clearly, Fig. 9.7 shows the basin weight distribution (red bars, ordered from major to minor) and the cumulate weight (blue line). Remarkably, in the collagen sequence, few basins (25 our of 1661) accumulate 99.4% of the network's weight. The P5 network shows a similar structure, with few basins (23 out of 529) accounting for most 99.4% of the information in the dynamics. These basins are the *specific* ones we defined previously, and are the ones plotted in the dendrograms of Fig. 9.6. Lac promoter shows the more distributed structure we anticipated with the free energy dendrogram, with 88 *specific* basins accounting for the 96.9% of the trajectory, showing a distribution which is closer to that of the random sequence.

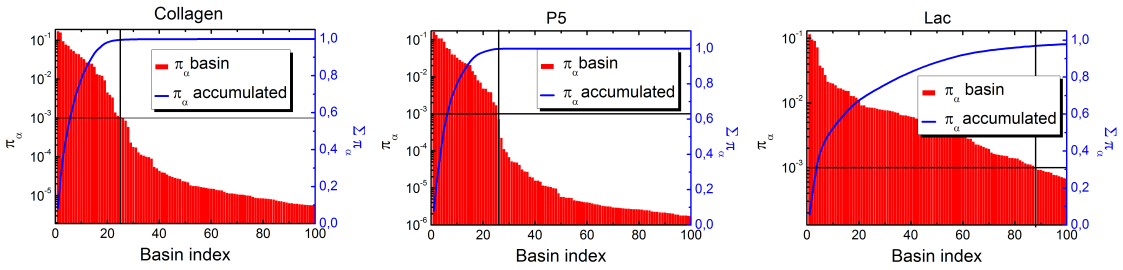


Figure 9.7: **Basin occupancy and cumulant occupancy for the three analyzed sequences** : The weight distribution of the free energy basins in the three sequences is rather different. Collagen and P5 sequences gather the majority of the trajectory in few basins, and the gap is clear. Lac promoter has a wider weight distribution, where more basins share similar weights.

The biological differences between the two kinds of promoters are elucidated by simply checking the structure of the networks obtained by applying our analysis method. Both strong promoters show a small fraction of basins which gather most of the trajectory. These basins correspond to physical states with a significative occupancy along the dynamics. On the contrary, in the lac promoter, several states coexist with relatively similar weights (around 20 states with weight between $10^{-1} - 10^{-2}$ and over 60 with weight between $10^{-2} - 10^{-3}$). This difference in the behavior is clearly seen in the way the cumulative function saturates.

To finish, we characterize quantitatively the three promoters by computing physical quantities regarding the highlighted states in each case. The three employed magnitudes are the weight, entropy and free energy difference with respect to the *nonspecific* state (named as NS), all defined previously. The free energy difference can be intuitively understood as the "depth" of the state (picturing a hill-valley imaginary representation of the free energy landscape), while the entropy the width of the valley (state).

The difference between the strong promoters and the weak one is numerically shown here. In the three cases the TSS and the TATA box are found as one of the most prominent basins in the network, revealing a relationship between the states populated by our model and the biologically active sites of the sequence. Nevertheless, the free energy differences comparing the strong and weak promoters are off by a factor three ($\sim 3k_BT$ for strong ones while $\sim 1k_BT$ for the weak one).

Table 9.1: **Statistical properties for the three studied promoters**: The TSSs are identified, together with additional biologically relevant states. Their population is shown, together with the entropy and the free energy difference with respect to the non-specific state.

Promoter	State	P_α	S/k_B	$-\Delta F/k_BT$
Collagen	A (TSS)	0.169	1.365	3.305
	B (TATA)	0.157	0.1380	3.232
	C	0.086	0.652	2.519
	NS	0.006	0.085	0.000
P5	A (TATA)	0.135	1.051	3.130
	B	0.107	0.913	2.898
	C (TSS)	0.086	0.684	2.681
	D	0.059	0.494	2.301
	NS	0.006	0.027	0.000
lac	A (TATA)	0.115	0.970	1.311
	B	0.095	0.891	1.120
	C (TSS)	0.090	0.775	1.066
	D	0.038	0.373	0.204
	NS	0.031	0.390	0.000

These numbers qualitatively match the distinction between strong and weak promoters in terms of RNA production. Considering our particle as a DNA-interacting protein, P5 and collagen promoters show "stronger" binding sites in what regards to the found free energy difference. In addition, the fraction of *nonspecific* trajectory is much higher in the lac promoter, revealing that *specific* interactions are scarcer and weaker. This results are in good agreement with the literature, as they account successfully for the energy ratios between weak and strong promoters [138]. Unfortunately we cannot compare directly free energy values of bound proteins considering our model.

9.6 Discussion

To conclude, we state briefly some of the achievements we consider important for the model we have described and used through the present chapter. The proposed model can be understood in a two fold manner. First, it is a protein-DNA interaction model, where the protein diffuses one-dimensionally along the DNA chain, and is coupled to the bubbles that form in the sequence. In this regard, we do not claim any generality in our proposal. The model intends to gain insight on the one-dimensional searching stage of DNA-interacting proteins, and on the class of proteins which are influenced by local openings in the DNA double helix. For example, this could serve well for the RNA polymerase and some transcription factors such as the TATA-box binding protein Obviously the actual mechanism of protein-DNA binding is much more complex and many other effects could be taken into account.

On the other hand, this model, together with the analysis procedure, can be un-

derstood as a physical method for analyzing DNA sequences—particularly promoters—in order to identify potential binding sites and also to provide quantitative information about them. We have proved how this latter application successfully differentiates between strong and weak promoters, as the free energy landscapes of both kinds of sequences are structurally different in the context of our model. Clearly, our model and method does not intend to compete with high-throughput methods for identifying TSSs or binding sites, based on bioinformatics algorithms. These sort of techniques are way more efficient but generally rely on statistical and data mining tools, laking of a clear physical or biochemical inspiration. Our mode,l on the contrary, focuses on the ability to yield valuable physical information about the sites, which can be qualitatively compared with experimental data. Chapter 10 goes in more depth on this point.

Chapter 10

Analysis of Cyanobacterial Promoters: Finding and Characterizing the TSS

This Chapter presents a careful analysis of nine cyanobacterial promoters from *Anabaena* PCC 7120 with the model and method presented in previous Chapter, as published in [59] We focus on the identification of the TSS, where the RNA polymerase binds prior to start the transcription. We identify and give quantitative information about the TSSs of the nine analyzed promoters. Furthermore, some of the chosen promoters have more than a single TSS, which allows a comparison between the strength of such sites.

10.1 Motivation: Why Cyanobacterial Promoters?

In previous Chapter we probed the protein-DNA model on three promoters already studied through PBD model. In this Chapter we take a step forward and analyze extensively nine promoter sequences from the same organism in order to validate the model and method. In particular, we choose nine promoters from Cyanobacterium *Anabaena* PCC 7120, and restrict our discussion to the location of the TSSs. We work with a simple prokaryote organism, as they show simple regulatory interactions, so the assumptions of the model are more appropriate. Also, we restrain our analysis to the TSSs of the chosen promoters. The binding site of the RNA polymerase is a common feature of any promoter, and this protein must form a bubble in the DNA molecule in order to read and transcribe its genetic meaning to RNA. Particularly, many studies suggest the relationship between this site and the propensity to form bubbles [97, 121, 126, 127] or even by studying flexibility profiles of DNA sequences [98].

Cyanobacteria are the only prokaryotes able to perform oxygenic photosynthesis, being key contributors to CO_2 fixation. Their interest resides in the ability of some strains to fix atmospheric nitrogen or to form harmful blooms by toxigenic species, among other properties [139]. This ecological relevance adds to their interest as a model for the study of multicellularity in prokaryotes [140], and as potential sources for novel drugs derived from their secondary metabolites [141].

The genome of *Anabaena* PCC 7120 contains 7,211,789 base pairs and 6,223 genes organized in a 6,413,771 base pair chromosome and 6 plasmids [142]. *Anabaena* PCC 7120 has been used for long time as a model for the study of prokaryotic cell differentiation and nitrogen fixation [143]. More recently, the experimental definition of a genome wide map of TSSs of *Anabaena* together with the analysis of transcriptome variations resulting from the adaptation to nitrogen stress have provided a holistic picture of this complex process [144].

We consider this system as a suitable one to probe our method, particularly due to its solid characterization and the wide body of knowledge it exists about it. The nine promoter sequences we have chosen for this study meet some requirements focussed in improving the amount of information we can extract from them. Particularly, four of these promoter exhibit multiple TSSs within the same promoter sequences. This feature allows us for a direct comparison between the strength of these sites within a same sequence, and thus to extract useful conclusions which might be compared with the existing knowledge about them.

10.2 Methods

We employ the same simulation and analysis protocol as in Chapter 9. See Sections 9.3 and 9.4 for further details.

10.3 Results

We analyze nine promoter sequences from cyanobacterium *Anabaena* PCC 7120 [59], which exhibit different regulatory features. All nine promoters have been well characterized from a biochemical perspective. Moreover, four of them have more than a single TSS. These features allow to a direct comparison between our findings and the experimental evidence.

10.3.1 Analysis of Complete Genes

Most works concerning the PBD model limit themselves to the study of short promoter sequences. In principle, this can be justified with various reasons. First, promoters span typically up to few hundreds of base pairs, which is the typical size over which the PBD model makes sense. Second, they are the regulatory regions in genes, and thus would seem to show a richer behavior in what regards to physical properties. Nevertheless, it would be advisable to check how do coding regions behave when compared to promoters, in order to compare their properties within the context of this model, justifying thus the restriction to promoter sequences alone.

In order to cover this gap, we simulate as a preliminary step three complete genes from *Anabaena* PCC 7120. We use for simplicity the PBD model alone, without the inclusion of the interacting particle. Our aim is to check which regions from the whole gene tend to form bubbles with more ease, showing large amplitude motions. The results allow us to compare the occurrence and intensities of the fluctuations detected in the promoter and coding regions, respectively, validating our further analysis, which will be restricted to the promoter regions.

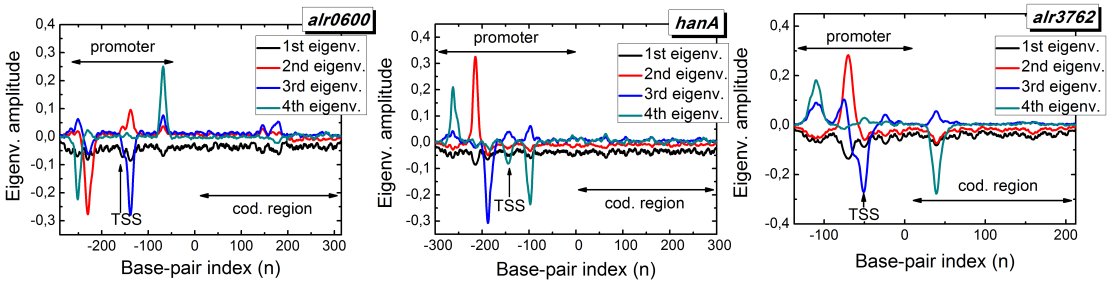


Figure 10.1: **Fourth first eigenvectors calculated for three different complete genes**: The promoter region—with the TSS highlighted—and the codifying region are pointed out. Most of the large amplitude motions appear localized in the promoter region, meaning that bubbles tend to form mostly there. This feature manifests the different mechanical behavior of the promoter and codifying regions, suggesting its key role in the DNA-protein interaction.

We simulate three complete genes, *alr0600*, *hanA* and *alr3762*, with the same *upward* structure of promoter + coding region. Figure 10.1 shows the first three PCA eigenvectors, with the promoter and coding regions highlighted. As discussed in previous chapters, very localized eigenvectors indicate strong fluctuations in the region of maximal amplitude. In the three cases, the first eigenvector is delocalized, with an almost constant profile which accounts for the overall fluctuations of the whole chain. The second, third and fourth ones show highly localized, large amplitude contributions, which indeed tend to concentrate in the promoter region, whereas the coding region shows little large amplitude modes.

This behavior means that, when considering the specific contributions to the overall fluctuations from the whole gene sequences, the largest part comes from the promoter regions. In other words, bubbles open more easily in this part of the sequence, while the codifying region remains on average closed. This fact supports the vision that physical properties of the DNA sequence might play a role within the whole gene, and the influence of the dynamic behavior with DNA-protein interaction problems. In this sense, this observation backs up the idea that some binding sites in promoter sequences can be characterized as regions where bubbles form easily, enhancing protein interactions.

10.3.2 TSS Finding and Base-Pair Opening

We analyze nine promoter sequences from cyanobacterium *Anabaena* PCC7120, of suitable length for the model and analysis method, between 100 and 300 base pairs. Five of these promoters have a single characterized TSS (*alr0750*, *argC*, *conR*, *furA* and *nifB*), while the remaining four exhibit multiple TSSs (*furB*, *ntcA*, *petF* and *petH*) [145–152].

Figure 10.2 shows the base pair opening profile for each promoter sequence, with the TSSs highlighted (upper panels), and the particle trajectory histogram (lower panels). In any case, a peak appears close to each TSS, meaning that bubbles form with high probability in these regions, while the particle dwells frequently at these sites. While these sites are spots which tend to open more easily, the particle is attracted to them, stabilizing the bubbles.

We highlight the fact that the opening probability is not strictly related with the A-T content of the local sequence. Although long stretches of A-T base pairs form “soft” regions where bubbles will form with a very high probability, this simple

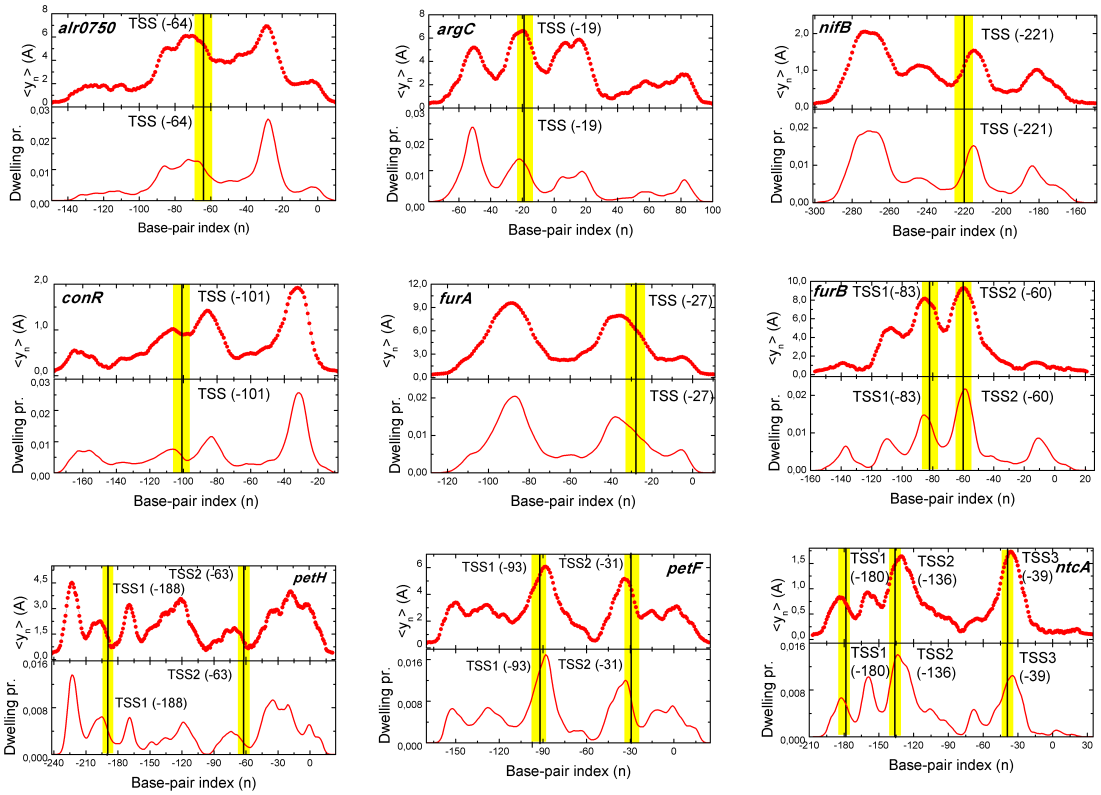


Figure 10.2: **DNA opening versus protein position**: Base pair mean opening (upper panels) and particle occupation histogram (lower panels) calculated for each of the nine promoters we study in this work. The horizontal axis represents the base positions counted from the coding starting point ATG (+1). We use this criterion to label the binding sites of the simulated promoters. The experimentally identified TSSs are shaded in yellow with the exact location marked with a solid bar. In every case, a clear peak appears around these sites, meaning that they are "softer" regions, and thus, bubbles form more likely. This fact supports their key role in the regulatory processes. This openings are not just related with the raw A-T content in the chain. The total A-T content of *Anabaena* PCC 7120 genome is around 58%. The A-T content of each analyzed sequence is: *alr0750* (61%); *argC* (64%); *nifB* (68%); *conR* (57%); *furA* (66%); *furB* (65%); *petH* (62%); *petF* (63%); *ntcA* (65%).

fact does not necessarily applies. The total A-T content of each sequence is written in the caption of Fig. 10.2. The interplay between the sequence and the dynamics is complex, mainly due to the nonlinear coupling between the base pairs. The long-range cooperativity of the model and the disorder of the sequence reveled in its heterogeneity affects both the equilibrium and kinetic properties of the DNA molecule, as it has been pointed out in previous studies [97, 121, 126].

In addition to the peaks centered on the TSSs, each sequence shows additional regions which open easily. Indeed, many of these peaks correspond to typical regulation sites for bacterial organisms, such as the -10 or the -35 regions. Although we focus our conclusions merely on the TSSs, these other regions appear as candidates for possible binding sites of other TFs, which are known to be influenced by the local properties of the DNA molecule [121, 126].

10.3.3 Free Energy Landscape Analysis

In this Section we apply the Markov state model-based analysis method, in order to provide a significative vision of the Free Energy Landscape of the system. This allows us to define systematically the relevant states visited in the dynamics and to calculate quantitative information about them. In Section 10.3.2, we checked how regions with larger average opening and where the particle dwelled with higher probability correlated with the positions of the TSSs. Nonetheless, we studied that in a qualitative way, as we were not able to infer quantitative information from the opening profiles of the promoters. We are interested in giving measures about the strength of the TSSs in the promoter sequences, in particular to make comparisons in those with more than a single TSS.

We present together the data extracted from the simulation and analysis method in Table 10.1. For each sequence, we have selected the TSSs—previously identified through biochemical assays—and some prominent other states which appear in the analysis. They are characterized by the weight, free energy difference with respect to the non-specific state and the entropy. These magnitudes were already presented and defined in Chapter 9. Most populated states determine the most stable states, giving rise to high free energy differences. The entropy informs us about the multiplicity of such macrostates, as low entropies mean few very populated basins, while high entropies, a composition of many low populated microstates. Physically, we can relate this quantity to how localized are the states, or the overall width of the bubble.

To illustrate the free energy landscape of the system, we represent it as a disconnectivity graph or free energy dedrogram. Figure 10.3 shows them for the nine analyzed promoters For the sake of clarity, focus on the high population region, not showing the high energy states corresponding to non-specific basin ($\pi_\alpha < 10^{-3}$).

Intuitively, the vertical arrangement of the states in the dendrogram informs about their stability, while the hierarchical arrangement about the barriers needed to jump over them. In this sense, we represent, not only of the population of each state, but also about their dynamic relation.

In order to make a proper definition of the physical macrostates, we coarse-grain the basins networks, gathering those macrostates separated by barriers lower than k_BT, as they can be considered to be kinetically very close, with low transition times. Indeed, employing a larger lag time would likely merge them together. We highlight with a color circle states associated with a excitation in the TSS, showing their relative weight and a graphical representation of the state they represent. Such states are typically a large bubble located in the TSS with the particle (black ball) located on top.

These nine promoters have been chosen in order to make the most of our model, keeping in mind its limitations. The genome of *Anabaena* PCC 7120 is well-known, and the positions of the TSSs have been defined under different metabolic conditions [153]. Remarkably, these TSSs coincide with relevant states in the dynamics of the model, which are described as heavy free energy basins. Of particular relevance for discussion, are the promoters which exhibit more than a single TSSs within the same promoter sequence, as they allow for a relative comparison between the different found sites [146, 147, 150, 151, 154–156].

ntcA promoter is perhaps of remarkable relevance. In its 230 base-pair sequence, it shows three different TSSs of different nature [157], as seen clearly in Fig. 10.2,

Table 10.1: **Thermo-statistical properties of studied promoters**: Occupancy probabilities and thermo-statistical magnitudes of the TSS and other relevant sites of the promoter sequences. NS stands for nonspecific sites defined in the discussion section. As already stated, each site is labelled starting from the ATG position on the gene (+1)

Sequence	State	π_i	$\Delta F[kT]$	S/k
alr0705	TSS (-64)	0.219	1.42	0.77
	+28	0.288	1.66	0.85
	NS	0.0545	–	–
argC	TSS (-19)	0.220	2.10	0.70
	+50	0.329	2.50	0.59
	NS	0.027	-	-
nifB	TSS (-221)	0.315	3.47	0.39
	-270	0.444	3.81	0.86
	NS	0.010	-	-
conR	TSS (-101)	0.151	1.97	0.58
	-30	0.349	2.80	0.91
	NS	0.021	-	-
furA	TSS (-27)	0.449	3.45	1.35
	-87	0.39	3.32	1.16
	NS	0.014	-	-
furB	TSS1 (-83)	0.302	2.39	0.86
	TSS2 (-60)	0.276	2.30	0.79
	-10	0.149	1.68	0.28
	NS	0.028	-	-
petH	TSS1 (-188)	0.199	3.01	0.74
	TSS2 (-63)	0.117	2.48	0.33
	-220	0.166	2.83	0.40
	NS	0.010	-	-
petF	TSS1 (-93)	0.198	3.03	0.58
	TSS2 (-31)	0.268	3.33	0.67
	+1	0.101	2.35	0.33
	NS	0.010	-	-
ntcA	TSS1 (-180)	0.098	0.96	0.029
	TSS2 (-136)	0.205	1.69	0.73
	TSS3 (-39)	0.292	2.05	0.85
	NS	0.038	-	-

where three large bubbles stand at the indicated positions. Table 10.1 agrees also on this point, as the three TSSs show large stability, although the particular values for each of them is rather different. We can relate this feature with the occurrence and behavior of the three TSSs experimentally determined [157, 158]. First, TSS2, at position -136, produces a constitutive transcript regardless of the culture conditions, while TSS1. at position -180 is only used in the absence of nitrogen. Finally, TSS3, position -49, is also active under all conditions, but is highly induced under nitrogen deprivation. Table 10.1 displays a remarkably low free energy value for

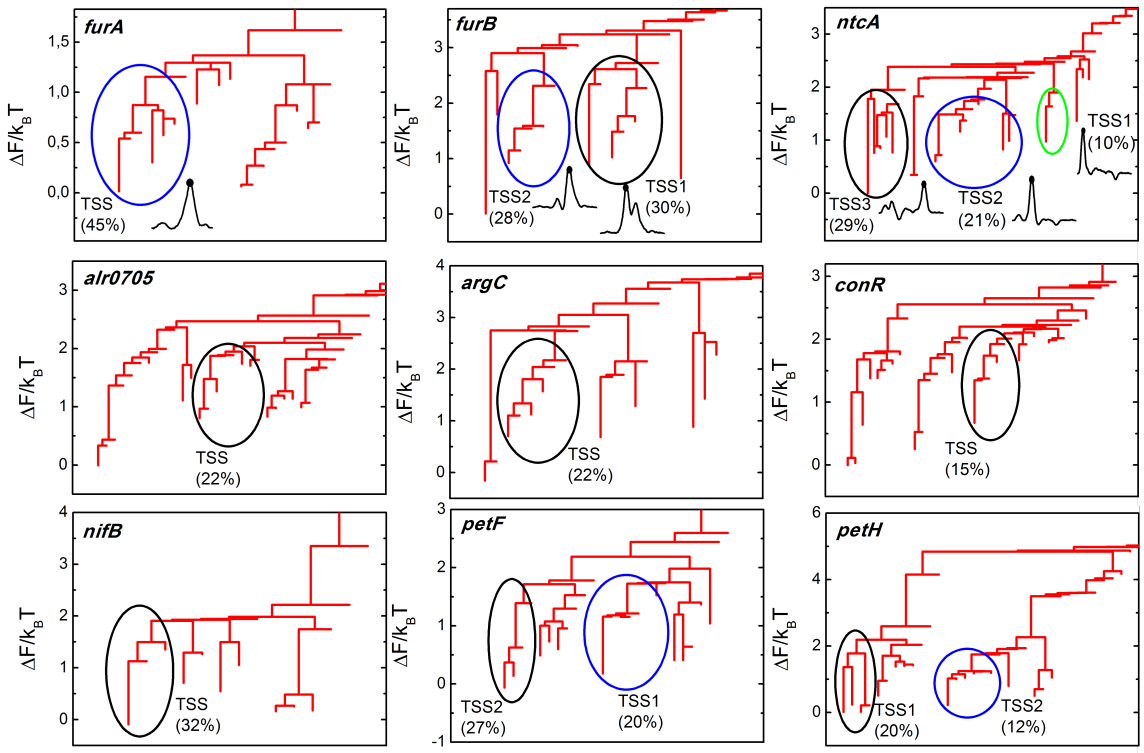


Figure 10.3: **Free energy dendrograms for the nine analyzed promoters**: Basins of attraction corresponding to the TSSs are highlighted.Their weight is indicated in the plot together with a representation of the physical state they represent, typically the particle located in a certain site where a bubble opens. In every case they are a low free energy branch on the disconnectivity graph.

TSS1, indication that its stability in the overall dynamics is rather limited. This suggests that its expression might be enhanced under more restrictive conditions. On the other hand, TSS2 and TSS3 appear as strong binding sites, covering both a large fraction of the total dynamics. These values are in good agreement with the *ntcA* transcription level at these sites under the correspondent conditions of nitrogen availability.

On the other hand, *furB*, *petF* and *petH* show also consistent results. The TSSs of the three promoters are clearly identified, coinciding with the experimental positions [149, 152, 153]. Determination of TSSs for *furB* promoter using the primer extension technique unravels two TSSs at positions -83 and -60 from the ATG, both with similar intensities [149]. Our analysis is in good concordance with such conclusions, as we find two major macrostates with similar weight (0.28 and 0.30), representing each of these TSSs. The resulting profiles for promoters *petF* and *petH* display also several relevant macrostates. Primer extension assays revealed a single TSS for the *petF* gene located at 100 bp upstream the translation start site. More recently, high throughput analysis showed two TSSs for *petF* at -93 and -31, which is in better agreement with our prediction. Transcription of *petH*, which encodes ferredoxin-NADP+ reductase, takes place from a constitutive promoter at -188 base-pairs from the ATC and a NtcA activated promoter, at -63 position. According to the proposed model, both TSSs are found as relevant macrostates in the free energy landscape of our model, although not as high peaks in Fig. 10.2. Indeed, the constitutive TSS at -188 exhibits a higher probability than the non-

constitutive one (see Table 10.1), indicating that the model is consistent with the experimental observations.

Regarding the remaining five promoters, high peaks are found around their single TSS, coinciding with the most—or one of the most—populated macrostrates, as defined. The case of *conR* is perhaps the one where the model shows a worse agreement with the phenomenology, as a significantly more relevant state appears in the dynamics. The experimental conditions under which the TSSs have been determined must also be taken into account. Usually, measures are done under standard culture conditions or under nitrogen deprivation, and thus, the existence of additional TSSs under different conditions, which are imposible to account explicitly with the current version of the model, cannot be discarded. In addition, it must be noted that the model is not considering particularly the DNA-RNA polymerase interaction, but the influence of DNA bubble formation on protein binding. In such sense, we remark that additional binding sites for other proteins influenced by the mechanical conformation of the DNA molecule might also be accounted here.

10.4 Discussion and Conclusions

Through this chapter, we have shown an application of the protein-DNA interaction model, employing it to analyze nine promoter sequences from a particular organism, *Anabaena* PCC7120 [59]. We focus on the TSSs, binding site of RNA polymerase, based on the hypothesis that such proteins couple their binding to DNA bubble dynamics. Applying a suitable analysis method, presented in Chapter 9, we represent the free energy landscape of the promoter sequences, as interpreted within the context of our model. Thanks to this tool, we define in an unbiased way the relevant macrostates of the system and relate them with biologically relevant sites, namely the TSSs, represented as bubbles in the DNA chain at these positions, and the particle bound in the region.

Upon genome analysis and TSSs detection, high-throughput approaches, such as proteomics, are commonly used, resulting in an enormous amount of data in a relatively short period of time. However, analysis of raw data to end up in genome annotation or TSSs mapping is a demanding, time-consuming task, necessary for taking advantage of this information that may delay a more detailed analysis of specific issues. Among the large variety of these methods, [159, 160], a great amount of valuable information is obtained, resulting in highly efficient analysis of genome that, nonetheless, generally lacks a a base on the physical mechanism of protein-DNA interaction. Our model and analysis method adopt a different strategy, not willing to compete in time performance with statistical-based techniques, but allowing a deeper understanding on the driving processes of protein binding. As a consequence of that, we have been able, not only to identify the TSSs, but also to characterize them in terms of physical magnitudes, allowing valuable discussions about the strength of each site.

Besides finding the TSSs in our free energy analysis, out method identifies additional relevant regions of the promoters that have not been experimentally probed yet. For example, we can mention the cases of promoters *furA*, *conR* or *nifB* (see Fig. 10.2 or Table 10.1), where very populated macrostates appear aside from the already discussed TSSs. We do not exclude the possibility of false positives, but these macrostates, given the general character of our model, may be related with

unknown regulatory regions. In this sense, our results suggest further experiments to search possible new relevant activity regions. Additional TSSs can appear if studied under different culture conditions, revealing the complexity of transcriptome profiles even in the case of simple organisms such as bacteria. Moreover, due to the general features in which our model is rooted, some macrostates identified with our method might indicate the existence of binding sites for further regulatory proteins which participate in transcriptome processes of *Anabaena* PCC 7120.

To conclude, we have chosen a particular prokaryotic organism such as *Anabaena* PCC7120 to probe our numerical method. This is done for different reasons. First, it is a well studied and controlled organism, allowing to contrast our results with experimental knowledge. Also, being a prokaryote, exhibits simpler regulatory mechanism, allowing our simple approach to work with more likelihood. Nevertheless, this model and method might be applied to the study of promoter sequences in many other organisms. Being the identification of protein binding sites in promoter sequences a key problem to understand and control regulation in biochemical and biotechnological processes, our method appears as a powerful complementary tool in this scientific endeavor.

Part III

Analysis of Force Spectroscopy Experiments and Simulations: from Forces to Free Energies

Be always drunken. Nothing else matters: that is the only question.
If you would not feel the horrible burden of Time weighing on your shoulders
and crushing you to the earth, be drunken continually.
Drunken with what? With wine, with poetry, or with virtue, as you will. But be drunken.
And if sometimes, on the stairs of a palace, or on the green side of a ditch,
or in the dreary solitude of your own room, you should awaken
and the drunkenness be half or wholly slipped away from you,
ask of the wind, or of the wave, or of the star, or of the bird,
or of the clock, of whatever flies, or sighs, or rocks, or sings,
or speaks, ask what hour it is; and the wind, wave, star,
bird, clock, will answer you: "It is the hour to be drunken!
Be drunken, if you would not be martyred slaves of Time;
be drunken continually! With wine, with poetry, or with virtue, as you will.

CHARLES BAUDELAIRE, *Enivrez-vous (Paris Spleen, 1864)*

Chapter 11

Single-Molecule Techniques and Single-Molecule Force Spectroscopy

This chapter serves as a succinct review of single-molecule techniques, in particular of force-spectroscopy methods, where individual molecules are probed by applying forces in the pN range. Also, we discuss briefly the three main force spectroscopy techniques highlighting their operation range and some achievements attained in the last few years.

11.1 Introduction: Single Molecule Experiments

Single-molecule methods involve the manipulation of individual molecule in order to study their properties. This opens a new and exciting field, which is directly contrasted to conventional experimental measurements. The main difference between single-molecule and traditional biochemical assays is the kind of average done when measuring a some molecular property. Single-molecule methods allow to sample directly the distribution of an observable. This allows, for example, the identification of rare subpopulations, to monitore directly the kinetic pathways or the or to recognize molecular intermediates. All this information is typically hidden in bulk experiments, which rely on ensemble averages done over populations of thermodynamical size $N \sim 10^{23}$

Since the burst of single-molecule techniques, not much more than 20 years ago, they soon gained a lot of popularity in many areas of science, from biology to chemistry, physics or material science. On the one hand, they open a new field which allows to investigate in new properties not possible to measure in the past. Also, they motivated the birth of many new theoretical developments, which were now possible to test directly in the lab. We focus on the application of these techniques to biomolecular systems.

Single-molecule methods are currently central tools for biological physics research. They offer a complementary and totally new approach to test molecular processes. Inside the cell, biomolecular processes occur at an individual scale, where thermal fluctuations are very significant, and molecular motion is fundamental to life. Thanks to single molecule-techniques, processes such as transport of cargo through the cell [161, 162], muscle contraction [163] or cell motility [164] have been

monitored at an individual level.

Also, the reduction of the observation scale increases the importance of statistical fluctuations. Their relative magnitude goes with $1/\sqrt{N}$ and thus it is meaningless in thermodynamic populations, but not with individual molecules. This opens a new and exciting field with new properties which in many cases seem to defy the laws of thermodynamics. The extension of the thermodynamic laws to small systems (see Chapter 14 and [165]) involves the necessity of new theoretical approaches that can now be probed thanks to these techniques.

Up to date, a large number of different single-molecule manipulation techniques have appeared, spanning six orders of magnitude in length (10^{-10} to 10^{-4} m) and force (10^{-14} to 10^{-8} N). These methods can be divided in two broad classes. In the first class fall those techniques which track molecular motion by labelling molecules without applying significant external forces. Single-Molecule fluorescence techniques or Fluorescence Resonance Energy Transfer (FRET) [166, 167] are examples of this class. The second class is devoted to the study of individual molecules through the application of external mechanical loads. Here fall methods such as optical tweezers, magnetic tweezers or Atomic Force Microscope (AFM) [168–170]. We focus on this latter group.

11.2 Single-Molecule Force Spectroscopy

Force plays a fundamental role in many biological processes. Biological motion—from cellular motility, to transport of cargo or DNA replication—is driven by forces at the molecular scale. These forces are in the range of few pN, given the characteristic molecular size and the magnitude of thermal fluctuations $\approx 4k_BT$.

In particular, many biological molecules have a well defined mechanical function. For example, they might have a certain mechanical stability, opposing a resistance to unfold under an external load. Giant protein titin is perhaps one of the most popular examples. Titin is the protein responsible for the passive elasticity in the skeletal and cardiac muscle sarcomere, and presents a huge resistance upon force [171]. Fibronectin and tenascin are components of the extracelllar matrix, and must extend and contract to facilitate certain cellular functions, such as cell migration or adhesion [172]. Other examples demonstrate the importance of learning about the unfolding of macromolecules under force. For example, some nucleic acid structures must break to permit translocation by enzymes such as RNA or DNA polymerases or by RNA helicases. Similarly, some proteins perform an enzymatic activity based on unfolding of molecules, like proteasomes [173, 174] or chaperonines [175, 176], which consume chemical energy to actively unfold or fold proteins.

Additionally, the application of forces to biomolecular systems allows to gain insight about their energy landscape. Force affects the thermodynamics and kinetics of reactions and transitions, perturbing the topology of the original landscape. The heights of the free energy barriers are changed and so is the relativity stability of the minima, enhancing transitions which would not occur in the absence of force. By applying suitable analysis techniques, information about the original landscape can be recovered from the output of the perturbed system.

In this regard, the application of external forces to individual molecules is an appealing approach. Application of external forces to individual molecules determines an interesting approach to probe molecular proceses. They enhance molecular tran-

sitions, but also monitor processes which involve molecular motion, such as translocation of polymers, or tasks performed by molecular motors. In general, techniques which employ an external force to probe a molecule are known as single-molecule force spectroscopy or Dynamic force spectroscopy (DFS from now on). The overall set-up of DFS experiments is very similar. Given the molecule or molecular system we wish to study, one end is attached to a surface in order to immobilize it while the free end is attached to a probe, through which force is applied. This attachment and immobilization are important parts of the set-up, as they should be ideally able to support infinite loads (at least higher than the applied forces) and should not affect the properties of the probed molecule.

There are four different DFS modalities, attending to the way in which force is applied: (a) constant force, where a constant load is exerted and the fluctuations in the extension recorded; (b) constant position, the probe is held at a constant position and we measure fluctuations in the molecular extension and force; (c) force ramp or force extension, where the probe is moved at constant velocity so the force is ramped, measuring the relation between force and extension and (d) force jump, where the force is changed abruptly between different values, recording the molecular extension. In the first two cases, we keep the system in equilibrium, and fluctuations are measured. In the other two cases, a non-equilibrium transition forces the molecule to stretch. The natural reaction coordinate in every case is the molecular extension, which changes in the direction of the pulling force.

The probe that applies the force has different origins, like an optical or magnetic trap or an AFM cantilever. Because of the scale at which we operate, thermal fluctuations impose fundamental limits in the length, force and time resolution of the experiment. In any case, the force is exerted through a linear spring of stiffness κ, determined by the stiffness of the probe alone, or the combination with some molecular linker.

These techniques rely on a proper determination of the force and the extension of the molecule. The precision and accuracy of these measurements depend critically on the ability to measure the position of the probe, and thus on the resolution of the experiment. The thermal environment and the sampling techniques are the limiting resolution sources. The spatial resolution is given by

$$\delta x = \sqrt{\frac{k_B T}{\kappa}}, \tag{11.1}$$

where δx gives the magnitude of the fluctuations in position. Thus, the force resolution is $\delta F = \sqrt{\kappa k_B T}$. In practice, the resolution is enhanced by filtering the position data, which is only sampled at frequencies below the characteristic viscous damping frequency for the molecular motion. In general, maximal resolution is achieved by minimizing the hydrodynamic drag on the probe [168, 169].

To apply the force in a controlled way, the load must first be calibrated. There are different approaches to this, usually relying on the Brownian motion of the probe. The stiffness κ is determined through the equipartition theorem or by analyzing the thermal spectrum [177]. Also, the response of the probe to a known force (like the viscous drag) can be measured. This stiffness depends on the particular experimental technique, which affects the force resolution and also its applicability range. In next sections we review briefly the three main DFS techniques. Table 11.1 gathers some properties of the three most popular DFS techniques. More detailed revisions can

Table 11.1: **Comparison of DFS techniques** (adapted from [168–170])

Feature	Optical Tweezers	Magnetic Tweezers	AFM
Spatial resolution (nm)	0.1-2	5-10	0.5-1
Temporal resolution (s)	10^{-4}	$10^{-1} - 10^{-2}$	10^{-3}
Stiffness (pN/nm)	0.005-1	$10^{-3} - 10^{-6}$	$10 - 10^4$
Force range (pN)	0.1-100	$10^{-3} - 10^2$	$10 - 10^4$
Displacement range(μm)	$0.1 - 10^5$	$5 - 10^4$	$0.5 - 10^4$

be consulted in [168, 169].

11.2.1 Optical Tweezers

Optical tweezers rely on the creation of optical traps due to the pressure of light radiation on small objects (beads) made of polystyrene, latex or silica. When we illuminate a bead by a laser beam, two forces appear: one is proportional to the gradient of the intensity of light, while the other is the scattering force due to the light reflected on the bead surface. When both forces are equilibrated, an optical trap is formed. This trap is an harmonic well to a very good approximation, so the forces acting on the bead follow Hooke's law $F = -\kappa x$, where κ is the stiffness of the trap and x the distance of the bead to the center of the trap.

The molecule we study is attached from one end to the bead, while the other is fixed to a surface or to another bead in a second optical trap (see Fig. 11.1). The force can be modulated by adjusting the intensity of the light or altering the position of the bead with respect to the trap center. The extension of the molecule is monitored using a CCD video camera relying on the interference between light scattered by the bead and the unscattered light light. The force can be held constant by employing a feedback loop which clamps it. Usually, infrared light is used to trap the molecule in order to avoid damaging of the molecule, as biomolecules are nearly transparent in this region of the spectrum.

Optical tweezers are probably the most versatile single-molecule techniques. The can exert forces up to 100 pN with resolution of 0.1 pN. The stiffness of the optical traps is lower than in other methods, $0.01 - 1$ pN/nm, which allows the control of such low forces.

The versatility of optical tweezers has allowed a vast array of measurements. For example, molecular motors have been widely studied through this technique, allowing direct observation of kinesins along fixed microtubules [162], transcription of RNA polymerase [178] or translocation mechanisms. Additionally, optical trapping showed the ability of viral packaging motors under large external loads [179]. Other class of studies involve the study of the mechanical properties of biomolecules. In particular, unfolding of RNA and DNA hairpins have been widely studied, due to the low forces optical tweezers allow to apply [180]. Measures of the mechanical unfolding of hairpin loops have allowed unveiling details of the folding free energy landscape [84, 180–182]. Proteins are also subject to this kind of DFS studies with optical tweezers [183–185].

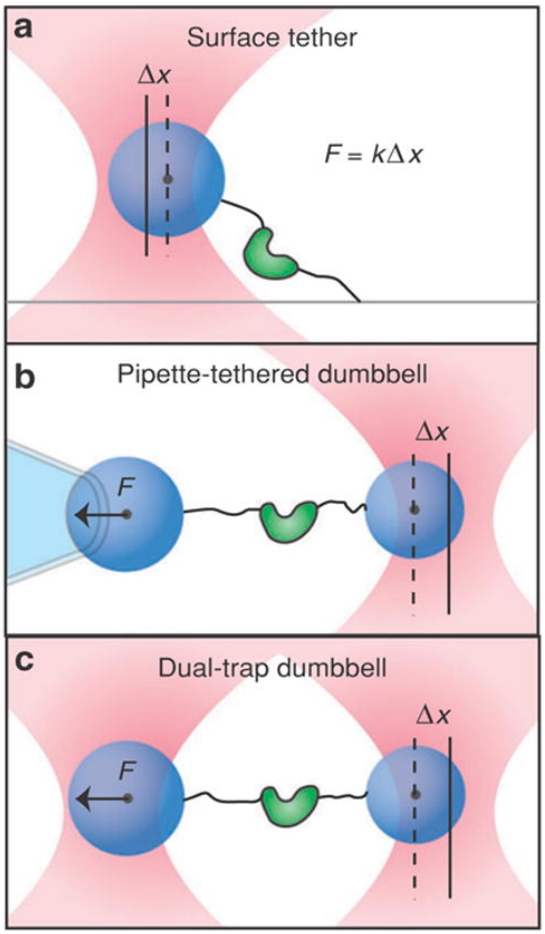


Figure 11.1: **Sketch of an optical tweezers experiment with three different assays**: An optical trap is created by focusing two laser beams. The bead is trapped by it and the molecule attached to it. Force is recorded by measuring the displacement of the bead from the center of the trap. We show three different assays, case (a) shows the surface-based assay, (b) dumbbell-based assay using one optical trap and a micropipette and (c) dumbbell-based assay with two optical traps (picture taken from [168])

11.2.2 Magnetic Tweezers

Magnetic tweezers have a similar philosophy to optical tweezers, but here molecules are manipulated through magnetic forces by attaching them to small superparamagnetic beads (see Fig. 11.2). The bead experiences a force which is proportional to the magnetic field gradient, which is approximately constant, given the scale separation between the molecular movements and the characteristic length of the field. In this sense, magnetic tweezers are intrinsically force-clamped, which is an advantage, given that electronic force-clamp have time resolution limits. The molecules are held between a magnetic bead and a glass surface. They can be pulled by moving the stage that supports the magnets, changing the magnetic field. The position of the beads is tracked from interference measures between unscattered light and the scattered light from the bead.

Magnetic tweezers have several advantages. The first one is their sensitivity, which allows tracking very low forces, from 10^{-2} to 10 pN, where the maximum force depends on the size of the bead. This is due to the very low stiffness of the traps, which is around $\kappa \approx 10^{-4}$pN/nm. Also, they allow to twist molecules by rotating

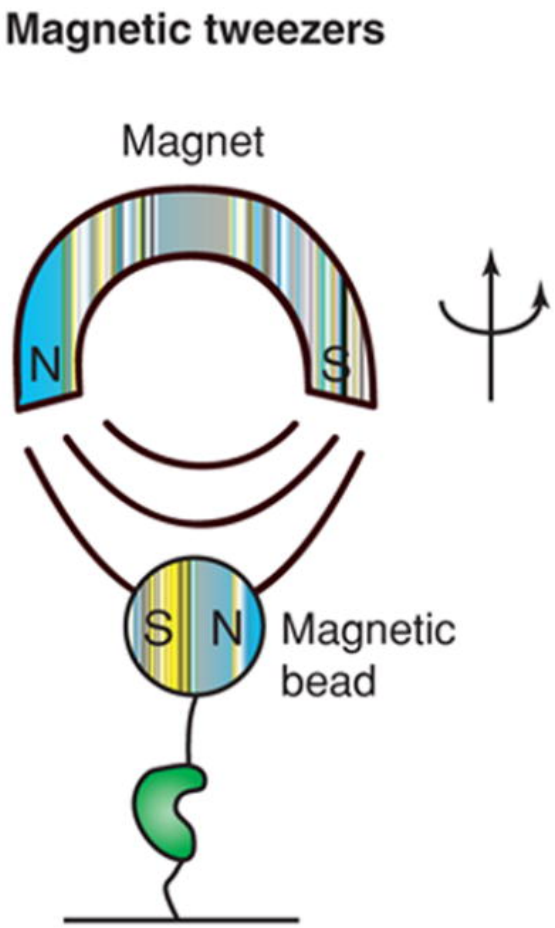


Figure 11.2: **Sketch of a magnetic tweezers set-up**: A superparamagnetic bead is trapped by an inhomogeneous magnetic field which applies a load to a single molecule. Control of the field can apply force and also torque (taken from [168])

the magnets, this is, to apply torque. As magnetic beads act as dipoles, they have a preferred orientation within the magnetic field. Additionally they allow a large parallelization, so several beads can be trapped and monitored simultaneously.

Magnetic tweezers have been extensively employed to investigate the properties of DNA molecule under torque [186, 187], or the mechanism of proteins such as topoisomerases [188, 189]. Also, their intrinsic stability, allows to perform very long force-clamp measurements, recording extensively long equilibrium trajectories for unfolding-refolding of proteins [190].

11.2.3 Atomic Force Microscope (AFM)

The AFM is perhaps the most familiar of the three techniques, given the very straightforward concept in which it is based. The AFM is a version of the scanning prove microscope, which allows to map a surface at sub-nanometer resolution. It is a very useful technique for imaging, but it also allows the measurement of interaction forces with pN resolution. Here, instead of sampling a particular surface, the AFM moves vertically, perpendicular to the plane.

The AFM uses a cantilever to apply force to a single molecule bound by one end to it and by its other end to a surface, which is typically moveable (see Fig.

11.3). This cantilever is in practice a linear spring, with a relatively high stiffness of $10 - 100\,\mathrm{pN/nm}$. By monitoring the deflection of the cantilever with the reflection of a laser beam, the force can be recovered by simply applying Hooke's law. Force is modulated with precision by moving the surface employing piezoelectric actuators. This surface can be retracted at a constant velocity, or in a force-clamp mode, where a feedback loop moves it to set constant deflection (force) in the cantilever. The displacement is also monitored by the piezo-stage.

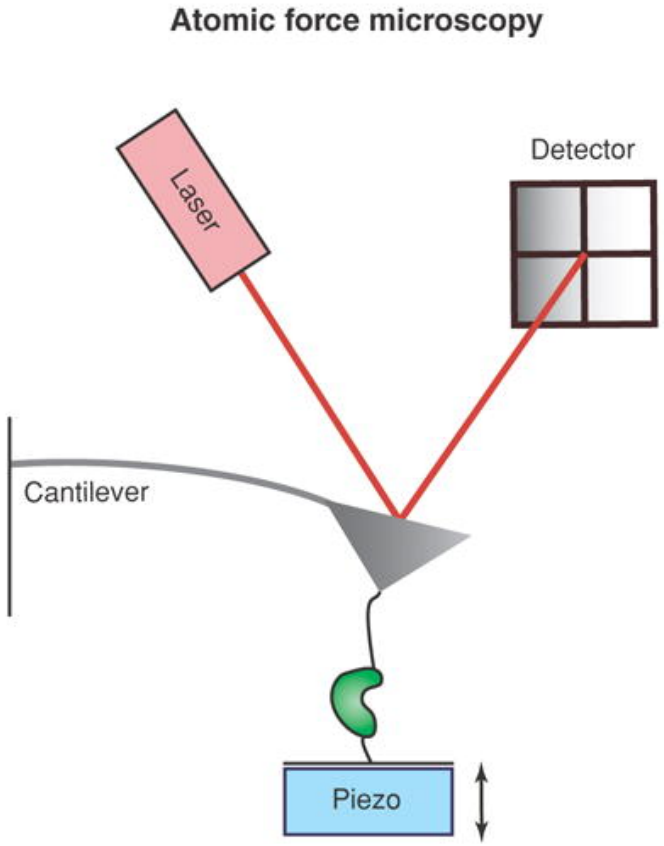


Figure 11.3: **Sketch of an AFM experiment**: A cantilever applies a mechanical force on a molecule of interest which is attached to a tip. The motion is recorded by recording the deflection of the cantilever. The force can be modulated by adjusting the position of the sample with a piezoelectric in the surface (taken from [168]).

In many cases, the molecule is attached to the cantilever tip by nonspecific adsorption, placing the tip in contact with the molecule and applying a large force "pushing" the molecule. This attachment can withstand typically large forces of $100 - 1000\,\mathrm{pN}$ but they are rather unknown and have an uncontrolled geometry. In this regard, many specific attachments have been developed, by modifying chemically the molecule and functionalizing the tips. For example, biotin-avidin bonds are widely used to hold the molecules. Other strategies can involve gold functionalization (useful when cysteine residues appear in the molecule) or more sophisticated techniques [191].

AFM has been applied with large success for studying unfolding of proteins [192–194], by measuring changes in the extension of the molecules as they denature. Protein titin is probably one of the most studied specimens, given their particular tandem structure and their high resistance to external loads [171, 195]. This has

allowed to unveil properties about their folding landscape, and also to reconstruct its role in muscle elasticity. Proteins which do not have a tandem-like structure may be studied also by forming recombinant chimeras between the proteins of interest, isolating them between tandem-like repeats which serve as molecular handles [196–198]. The particular sawtooth patterns in such studies serve as a valuable fingerprint which ensures the validity of the individual measures.

Chapter 12

Free Energy Recovery from Single-Molecule Experiments

This chapter aims to offer a review on two different theoretical frameworks which have proved their usefulness for analyzing force spectroscopy single molecule experiments. In particular, both methods focus on recovering free energy magnitudes from the force-response of an individual molecular system. On the one hand, force spectroscopy theory is concerned about kinetic properties, as it proposes expressions to estimate free energy barriers from escapes on one-dimensional profiles subject to a mechanical force. On the other hand, Jarzysnki equality determines one of the paradigms of the so called thermodynamics of small systems, which allows to compute equilibrium free energies by analyzing the non-equilibrium fluctuations of a system subject to an external perturbation.

12.1 Introduction

The evaluation of single molecule pulling experiments defines an important challenge. Molecular systems are intrinsically stochastic and far from the thermodynamic limit. Therefore, thermal fluctuations play a relevant role, and out-of-the-average rare events are significant. Furthermore, transitions under an external load are usually irreversible, occurring far away from equilibrium. The revolution of single molecule techniques has brought itself the necessity of developing new tools to deal with small out-of-equilibrium systems, or to recover properties from an unperturbed system by analyzing the perturbed response.

Usually, the problem is formulated in the following way. The molecular system is represented as a one-dimensional free energy profile along the pulling direction, which should be be a proper reaction coordinate [18] (although this is is not always the case [199, 200]). This choice is often determined by the experimental limitations, which allow to measure just changes in the molecular extension. The external force perturbs the system, effectively tilting the free energy profile. The perturbation is applied through a pulling device, usually the experimental probe (AFM cantilever or Optical Tweezers trap, for instance), connected in series with some molecular linker such as a polymer or some DNA handles. We consider here two main protocols (although they are not the only ones [84]) according to the way we perturb the system, namely (a) the constant-force mode, where fluctuations in extension are recorded as the load is held constant, (b) the constant-rate mode, where the

molecular extension is recorded as the force is ramped by moving the pulling device at constant velocity. The first case records equilibrium fluctuations, while the later is an out of equilibrium process.

In principle, the unperturbed free energy profile is the target information we wish to obtain. The output from the experiments is typically the force response of the system, namely a waiting time or escape rate (in constant force mode) or a force-extension curve (in constant-rate mode). Reconstructing the whole free energy profile of the molecular system is an extremely challenging problem, which depends critically on the resolution of the experiment and also on the particular molecular system [84].

Nevertheless, we can divide this broad problem into two more specific questions, which are the determination of the kinetic and of the thermodynamic properties of the system. The first question relies on the problem of jump over a free energy barrier, intimately linked to Kramers reaction-rate theory [201], but with the difference that here the free energy profile changes dynamically due to the action of the force. In the second problem, we wish to obtain equilibrium information about a system, by subjecting it to a nonequilibrium transition [33, 202].

We review here two different theoretical frameworks which focus on answering the two previous problems. In the first case, starting from Kramers theory, we consider how to recover the original free energy barrier (kinetic properties) from the distribution of rupture forces [203, 204]. In the second case, we review nonequilibrium free energy methods, specially Jaryznki equality, which is able to relate nonequilibrium work measurements with equilibrium free energy differences between two states [205].

12.2 Kramers Theory

The problem of thermal escape from metastable states is ubiquitous in many scientific areas, from electrical transport theory, to diffusion in solids or chemical kinetics. In 1940 Kramers contributed in this field by proposing an expression for the thermal escape of a Brownian particle from a metastable well [206].

Kramers defines the problem as a brownian particle in a one dimensional profile. The particle is confined in a potential well and must must surmount an energetic barrier in order to reach another, more stable, well (see Fig. 12.1). If the temperature is low—compared with the barrier height—the particle spends most of the time in the potential minimum, so reaching the top of the barrier is a rare event. Once there, it can fall back to the original minimum or reach the target state.

This kind of situations are very common as, for example, it occurs in chemical reactions from a reactant state A to the product state C, by surmounting a energetic barrier where the transition state B is located along some reaction coordinate X. It is also a fundamental problem in biophysics, for example considering an unfolded protein which reaches the folded state in a two-state picture.

Kramers problem takes some assumptions. The equilibration time τ_{eq} within one minimum—this is, the time after which an ensemble of systems has the Maxwell-Boltzmann equilibrium distribution corresponding to an infinite barrier—must be much smaller than the escape time τ_{es} from state A to state C. Also, both time scales must be much larger than the fast degrees of freedom, not considered explicitly in

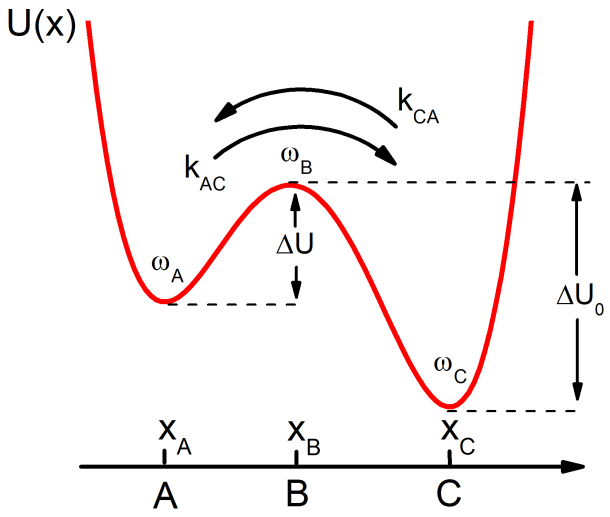


Figure 12.1: **Schematic bistable potential to illustrate Kramers problem**: Metastable states are A and C with characteristic frequencies ω_A and ω_C respectively. Relative energy difference is ΔU_0. The energetic barrier that separates both states is ΔU and is characterized by a top frequency of ω_B.

the model. In turn, this assumption can be written in terms of the energetic scales as

$$k_B T \ll \Delta U < \Delta U_0, \tag{12.1}$$

where ΔU is the height of the barrier from well A and ΔU^0 the energy difference between states C and A, $\Delta U^0 = U_C - U_A$ (see Fig. 12.1).

Another set of competing time scales are related with the coupling to the thermal bath. The damping γ (of inverse time units) determines this scale. If the particle is confined in one of the two wells, it performs oscillations along states A or C with typical frequencies given by

$$\omega_A = \sqrt{\frac{U''(x_A)}{m}}; \qquad \omega_C = \sqrt{\frac{U''(x_C)}{m}}, \tag{12.2}$$

where m is the mass of the particle and $U''(x_A)$ is the second derivative of the energy profile around x_A, and same for x_C. When the particle has an energy larger than the barrier, there is a time scale for the exchange between kinetic and potential energy during the barrier crossing, given by

$$\omega_B = \omega_C = \sqrt{\frac{U''(x_C)}{m}}. \tag{12.3}$$

This allows us to distinguish between two regimes according to the friction γ,

1. Strong friction $\gamma \gg \omega_B$.
2. Weak friction $\gamma \ll \omega_B$.

In the high friction limit, Kramers proved that the transition rate from A to C can be written as [206]

$$k_{AC} = \frac{\omega_A}{\gamma}\frac{\omega_B}{2\pi}e^{-\Delta U/k_BT}. \tag{12.4}$$

While in the moderate-to-strong friction limit, this expression is corrected as:

$$k_{AC} = \frac{1}{\omega_B}\left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + \omega_b^2}\right)\frac{\omega_A}{2\pi}e^{-\Delta U/k_BT}. \tag{12.5}$$

This theory serves as a theoretical starting point for analyzing force spectroscopy experiments, as in such problem a molecular transition occurs between two states separated by some free energy barrier.

Mean First Passage Time

Another important concept for problems of thermal activation is the Mean-First-Passage-Times (MFPT). This term is tightly related to Kramers problems and tries to answer how long does it take a random walker to reach a given target point. In other words, we have an ensemble of dynamic systems at some point (or state) A and we want to know what is the mean time $\langle t\rangle$ needed to reach a point B where they are absorbed. In the overdamped limit, considering the the system is subject to a potential $U(x)$, the MFPT can be exactly computed as [201]:

$$\langle t\rangle = \frac{m\gamma}{k_BT}\int_A^B dy e^{U(y)/k_BT}\int_{-\infty}^{y} dz e^{-U(z)/k_BT}. \tag{12.6}$$

There is a large body of literature studying the relation between Kramers rate and MFPTs. Usually it is considered that [201, 207]

$$k = \frac{1}{\langle t\rangle} \tag{12.7}$$

12.3 Force Spectroscopy Theory

We focus here on a problem of particular interest for analyzing single-molecule experiments, where we apply a pulling force to induce a molecular transition. One of the goals to understand such problem is to extract reliable information about the kinetics of the process in the absence of external forces.

Physically, the formulation of the problem is similar to that of Kramers. The molecular transition is modeled as a thermal escape event over a free-energy barrier. The difference is that the thermally activated escape is done over a free energy barrier that is perturbed by an external force.

We assume that the molecule moves on a free energy profile along the pulling direction x, which comes from the combination of the original profile $U_0(x)$ and the influence of an external force F, exerted by a pulling device of effective stiffness κ (see Fig. 12.2). Usually, this spring is considered to be *soft* compared to the "molecular stiffness"—effective stiffness of the initial equilibrium well. We discuss about this point later on.

The combined free energy profile for the whole system is $U(x) = U_0(x) + V_P(x)$, where $V_p(x)$ is the potential due to the pulling device. In a constant force mode, up to first approximation, $V_p(x) = -Fx$, where F the applied force. In a constant

rate mode, assuming the soft spring approximation, $V_p(x) \approx -\kappa V t x$, where V is the pulling velocity, so similarly $U(x) = U_0(x) - F(t)x$, being $F(t)$ a force which increases with time, tilting the original profile. This profile is assumed to have a single well at $x = 0$ and a barrier of height $\Delta G^\dagger$ at $x = x^\dagger$. The external force F makes the barrier decrease, so $\Delta U(F)$. In the constant force model, F is constant, so the combined free energy profile $U(x)$ is static. In such case the output of the experiment is the waiting time for the system to escape. In the constant rate mode, the pulling device moves at a constant velocity, so the force changes with time $dF/dt = \kappa V$. Here, the output of the experiment is the rupture force. In any case, the escape process is stochastic, so we obtain a force dependent escape time distribution $p(t; F)$ and a velocity-dependent rupture force distribution $p(F; V)$.

Now, the question we want to answer is, how can we obtain the free energy barrier—or the kinetic rates—at zero force from the rupture time or force distributions?

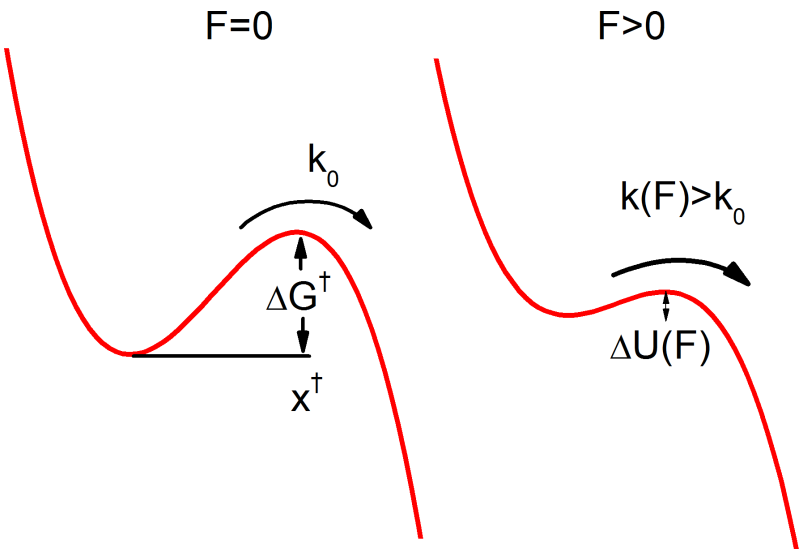


Figure 12.2: **Schematic picture for the force-spectroscopy problem**: In the absence of force, the particle is confined in a one dimensional profile, with one equilibrium well and a single barrier of height $\Delta G^\dagger$ at $x^\dagger$ and an intrinsic rate constant k_0, which are the three magnitudes to determine. The external force tilts the profile, decreasing the barrier height $\Delta U(F)$.

Bell-Evans Phenomenological Theory

One of the first attempts to deal with force spectroscopy problems comes with Bell's pioneering work [208] and the subsequent extension by Evans [209]. At a first approximation, rate of rupture $k(F)$ scales with the exponential of the force F. According to this, in the constant pulling rate mode, the mean rupture force grows proportionally to the logarithm of the pulling speed [203, 210]. Particularly, the expressions for both cases :

$$k(F) = k_0 e^{F x^\dagger / k_B T} \tag{12.8}$$

$$f^* = \frac{k_B T}{x^\dagger} \log \frac{x^\dagger r_f}{k_0 k_B T}, \tag{12.9}$$

where f^* is the most probable rupture force, and $r_f = df/dt$ the loading rate of the experiment. This model is known as the phenomenological model or Bell-Evans formula. It is widely used to extract the intrinsic rate coefficient k_0 and the position of the transition state $x^\dagger$. Nevertheless, the derivation assumes that the transition state $x^\dagger$ does not change with the force and that the barrier $\Delta G^\dagger$ decreases linearly with the force. These assumptions are not true for most shapes of the free energy profile, so Bell-Evans is only applicable under certain conditions, such as mechanically brittle molecules or when the applied tension is sufficiently small not to shift the position of the transition state $x^\dagger$.

Dudko-Hummer-Szabo Theory

Some years ago, Hummer and Szabo [211] and Dudko *et. al.* [212] proposed almost simultaneously an expression that related the pulling velocity with the average rupture force, allowing to recover not only k_0 and $x^\dagger$, but also the height of the free energy barrier $\Delta G^\dagger$. In the first case [211], by applying Kramers theory to a parabolic cusp potential tilted by an external force, they showed that for intermediate pulling speeds $\langle F \rangle \sim (\log V)^{1/2}$. In the second case [212] combination of Kramers theory with certain scaling laws obtained by Garg [213] predicted that $\langle F \rangle \sim (\log V)^{2/3}$.

These two theories disagreed on the results but also on the physical approach. Years later, this discrepancy was resolved in a joint work [204] where a common framework was set and a unified formalism proposed for recovering k_0, $x^\dagger$ and $\Delta G^\dagger$. There, a general dependence was obtained $\langle F \rangle \sim (\log V)^\nu$, where ν was a parameter which depends on the shape of the free energy profile, being $\nu = 2/3$ for a linear-cubic profile, and $\nu = 1/2$ for a parabolic-cusp one. Additionally, values $\nu = 1$ or $\Delta G^\dagger \to \infty$ for any ν recovered Bell-Evans expression.

In particular, for a constant force protocol,

$$k(F) = k_0 \left(1 - \frac{\nu F x^\dagger k_B T}{\Delta G^\dagger}\right)^{1/\nu - 1} \exp\left\{\frac{\Delta G^\dagger}{k_B T}[1 - (1 - \nu F x^\dagger \frac{k_B T}{\Delta G^\dagger})^{1/\nu}]\right\}, \quad (12.10)$$

and the average rupture force at constant pulling velocity

$$\langle f \rangle = \frac{\Delta G^\dagger}{\nu x^\dagger}\left\{1 - \left[\frac{k_B T}{\Delta G^\dagger} \log \frac{k_0 e^{\Delta G^\dagger / k_B T + \gamma}}{x^\dagger \kappa V}\right]\right\}, \quad (12.11)$$

where $\gamma = 0.577...$ is the Euler-Mascheroni constant. This formalism considers some assumptions, namely a soft spring pulling device, high barriers—in order to apply Kramers theory—and high forces.

This formalism seems rather model-dependent, as a particular analytical shape for the underlying molecular free energy-profile is assumed. Nevertheless, it is less model-dependent that it appears, because under sufficiently high forces, any analytical profile can be well represented by a linear-cubic potential [204].

Effect of the Pulling Device

Up to here, we have not discussed about the effects of the device we use to perturb the free energy profile. Nevertheless, in practice, is a relevant point given that single molecule techniques use effective springs of different stiffnesses to probe molecules

(see Chapter 11, and Table 11.1). Also, in many cases polymer handles are used as linkers, setting a complex pulling device whose effective stiffness changes with the pulling force.

We focus in the constant pulling rate mode (or force-extension mode) where the spring is pulled at constant velocity V. In constant force mode (or force-clamp protocol) the role of the spring is limited, given that a constant force is maintained, so it just affects in its fluctuations. The combined free energy profile for a harmonic pulling device is:

$$U(x,t) = U_0(x) + \frac{1}{2}\kappa(Vt - x)^2. \tag{12.12}$$

With the soft spring approximation, the harmonic potential written as $U(x,t) \approx U_0 - \kappa V t x$, so that the perturbing force is $\langle F(t)\rangle = \kappa V t$. As derived in [214] for an arbitrary spring stiffness one should write

$$\langle F(t)\rangle = \kappa V t/\chi \equiv F(t), \tag{12.13}$$

where $\chi > 1$ is $\chi = 1 + \kappa/K_U$, being K_U the effective stiffness of the free energy profile at the equilibrium well, approximated as $U_0(x) \approx K_U x^2/2$. In this sense, χ measures the departure from the soft spring approximating, recovering Dudko-Hummer-Szabo theory when $\kappa \ll K_U$, as $\chi = 1$.

Then, for an arbitrary stiffness [214]

$$\langle F\rangle \approx \frac{\Delta G^\dagger}{\nu x^\dagger}\chi\left\{1 - \left[1 - \frac{k_B T}{\Delta G^\dagger \chi^3}\log(1 + e^{-\gamma}/qX)\right]\right\}, \tag{12.14}$$

where γ is the Euler-Mascheroni constant, $q \approx e^{-\kappa(x^\dagger)^2/2k_B T}$ and $X = k_0 k_B T/\kappa V x^\dagger$. This expression introduces a fourth parameter, χ. Although κ is known, K_U is difficult to estimate, so χ is another parameter to fit, which can be used as a criterion for checking if the soft-spring approximation is enough or not.

In many practical cases one does not pull with a regular linear spring, but rather with an arrange of molecular linkers and springs, which determine a complex pulling device, whose stiffness changes as we pull. For example, if we have a polymer linker and we pull with an AFM cantilever, the effective stiffness is dominated by one part or the other depending on the instantaneous pulling force. For low forces, the stiffness of the polymer is very low, so the effective stiffness of the pulling device is ruled by the polymer. At higher forces, the stiffness of the polymer changes dramatically to very high values, becoming the linear spring the dominating part, given the series connection.

Further details on the influence of complex pulling devices can be seen in [215]. Nevertheless, for our practical case, the soft-spring approximation is enough, as we discuss and check in the analysis of experiments and simulations (Chapter 13).

12.4 Non-Equilibrium Methods for Equilibrium Free Energy Calculations

In this Section, we review how we can use *nonequilibrium* methods to calculate *equilibrium* free energies. This might seem a contradictory statement at first, but as shown by Jarzynski in the late 90s [205, 216], nonequilibrium perturbations can be

used to obtain free energies in a formally exact way. Jazysnki identity is not just an expression of theoretical interest, but provides a quantitative basis for the analysis of experiments where single molecules are manipulated mechanically.

12.4.1 Thermodynamics of Small Systems

Before dealing formally into nonequilibrium relations, it is worth to devote some words to the field in which these contributions are set. In this section, we deal with systems which are driven away from an initial state of thermal equilibrium. In principle, these results are valid a general way, but they are mainly relevant for microscopic systems. This broad field is often referred to as *Thermodynamics of small systems* [165].

Here, the central question is how to apply, or until which extent are applicable, the well known laws of thermodynamics, originally formulated for macroscopic systems. While one is able to understand well systems such as steam engines, the behavior of microscopic but equivalent systems, such as molecular machines, presents a challenge. The main difference here is that the $\sim 10^{23}$ involved degrees of freedom basically rule out any possible deviations from the mean behavior. Nevertheless, in smaller systems, which have a characteristic energy scale of $\sim k_B T$, the statistical fluctuations become more prominent, so thermal fluctuations become an active ingredient which leads to rather unexpected properties. A first effect of this scale, is that thermodynamical laws, which are known as equalities, must be rewritten as inequalities or in terms of distributions [202, 217].

There are several examples for this kind of systems, such as magnetic domains in ferromagnets, atomic clusters, or biological macromolecules, like molecular motors, which operate away from equilibrium and dissipate energy continuously. Despite their inherent scientific interest, until the early 90s no experimental method was available to investigate the properties of small systems. The advent of single molecule techniques was a natural boost for this field, allowing an increasing interest and fast development [218].

The field of thermodynamics of small systems is very broad, and itinvolves several topics [219]. For our purpose here, we are interested in exploring the relationship between the work performed and the free energy changes in nonequilibrium thermodynamic processes. In order to begin with that, it is useful to start with a simple and intuitive example, to define some basic concepts and understand the role of fluctuations at such scale.

Thermodynamics Example: Stretching a Rubber Band

We start with an ordinary rubber band, attached to a fixed wall from one end and the other to an ideal spring [202]. We denote with z the length of the rubber band while λ is the distance from the wall to the end of the spring. Importantly, λ is a degree of freedom we can control directly by moving the end of the spring. λ is thus a *work parameter* or *control parameter*, in opposition to z.

We can subject the system to a nonequilibrium process, starting from a well defined thermal equilibrium state, with the control parameter at some initial fixed value $\lambda = A$. The rubber band is now stretched by changing the control parameter to $\lambda = B > A$. If we do so very rapidly, the rubber band heats up, being driven away from equilibrium with the surrounding air, the *thermal enviroment*, and we

perform *work* W on the system. Through this nonequilibrium transition from one state A to another B, we carry out an irreversible process.

The second law of thermodynamics states that the work we have performed on the system is greater or equal to the change in free energy ΔF between the two equilibrium states A and B,

$$W \geq \Delta F = F_B - F_A. \tag{12.15}$$

If we would have changed λ from A to B very slowly, in a *quasistatic* way, the equality $W = \Delta F$ would hold, as the process would have been reversible.

Now, instead of having a rubber band, we consider a biomolecule, such as a DNA hairpin, and the spring is an optical trap or AFM cantilever [220]. Our thermal environment is now an aqueous solution at room temperature. The difference is that the thermal fluctuations are of the relevant energy scale, unlike what happens with a macroscopic rubber band. We start at a value $\lambda = A$ and stretch our biomolecule to $\lambda = B$ following some *nonequilibrium protocol*, allowing the system to relax back to equilibrium in the final state.

If we perform this experiment several times following the same protocol, the work performed in each trajectory is different due to the statistical fluctuations of the thermal bath. Then, we have to understand the laws of thermodynamics in an statistical way. Instead of having a work value performed on the system, we have some work distribution $p(W)$, which depicts the distribution of work values observed over many realizations of our nonequilibrium protocol. Equation (12.15) should be reinterpreted as,

$$\langle W \rangle \geq \Delta F, \tag{12.16}$$

but there would be a significant spread of work values around this average. This statistical reformulation of the second law of thermodynamics, allows single realizations for which $W_i < \Delta F$. This events are apparent "violations of the second law", although there is actually no such violation, as we are not in a thermodynamic system. In such events, random thermal fluctuations interfere constructively in order to facilitate the process, or to "exert work" on the system. Such events are actually quite relevant for determining equilibrium free energies from nonequilibrium trajectories.

12.4.2 Jarzysnki Equality

Jazynski introduced his now famous nonequilibrium work relation, the Jarzynski equality, in 1997, proving a novel treatment of dissipative processes in nonequilibrium systems [205, 216]. Jarzysnki equality provides a practical way to compute free energy differences, and has now been proven with success in many experimental and computational systems [220, 221]. Jarzysnki equality states that,

$$e^{-\Delta F/k_B T} = \langle e^{-W/k_B T} \rangle. \tag{12.17}$$

where ΔF is the free energy difference between two equilibrium states and W is the work performed over a nonequilibrium protocol through such states.

Equation (12.17) deserves further explanation in order to a correct applicability. In the same fashion as with the toy example given in previous section, Jarzynski

equality considers a system kept in contact with a thermal bath at temperature T, and takes a control parameter λ which controls the equilibrium state. For example this parameter might be the the pulling distance of a biomolecular polymer (not its extension, which would be an stochastic coordinate). The nonequilibrium transition starts at some fixed value $\lambda = A$ where the system is in *thermal equilibrium*. The nonequilibrium protocol or path $\lambda(t)$ is a transition from $\lambda = A$ to $\lambda = B$, where $\Delta F = F_B - F_A$ is the difference in free energy between such states, and the work performed over such transition as:

$$W = \int_{\lambda=A}^{\lambda=B} dW = \int_{\lambda=A}^{\lambda=B} F d\lambda. \tag{12.18}$$

The average $\langle \cdots \rangle$ in Eq. (12.17) is the ensemble average over nonequilibrium trajectories on the fixed protocol. It is a combination of an ensemble average over initial conditions, chosen according to the equilibrium Boltzmann probability in state $\lambda = A$, and a path average over individual realizations. If we had deterministic dynamics, only a single trajectory exists for any given initial condition, but for stochastic dynamics, as in small systems, the path average is over realizations of noise. Recall that we do not need the final state $\lambda = B$ to be in thermal equilibrium, as no work is performed on the relaxation to equilibrium

Jarzynski equality is often rewritten as $\langle e^{-D/k_B T} \rangle = 1$, where $D = \langle W \rangle - \Delta F$ is the work dissipated along the given trajectory. Due to Jensen equality $\langle e^{-x} \rangle \geq e^{-\langle x \rangle}$, second law $\langle W \rangle \geq \Delta G$ immediately follows from Eq. 12.17. As mentioned before, Jarzynski equality only holds if there exists nonequilibrium trajectories where $D \leq 0$, the "violations of the second law". These trajectories ensure that the microscopic equations of motion are time-reversal [33].

12.4.3 Forward and Reverse Processes: Crooks Fluctuation Theorem

Equation (12.17) considers nonequilibrium processes where λ is changed from A to B, being thus a *forward* process. In the same way, we can perform the *reverse* process, where λ is varied from B to A. This reverse protocol is the time reversal of the one used in the forward process. The whole cycle starts from the equilibrated state $\lambda = A$, then a transition to B, let the system *reequilibrate* and move it back to $\lambda = A$. We can denote now W_F the work performed on the forward process and W_R on the reverse one. For a thermodynamic system, the work performed through the complete cycle $\lambda : A \to B \to A$, satisfies,

$$W_F + W_R > 0, \tag{12.19}$$

which is essentially a "no free lunch theorem", ruling out perpetual motion machine, which allows to extract net energy from the thermal environment. For a microscopic system, this inequality is as

$$\langle W \rangle_F + \langle W \rangle_R > 0, \tag{12.20}$$

where the ensembles averages are over the forward $p_F(W)$ and backwards $p_R(W)$ distributions respectively. This is, on average we have "no free lunch", but occasionally, we might recover work for some cycles. The distributions satisfy a symmetry relation as proved by Crooks [222]:

$$\frac{p_F(+W)}{p_R(-W)} = e^{(W-\Delta F)/k_B T}. \tag{12.21}$$

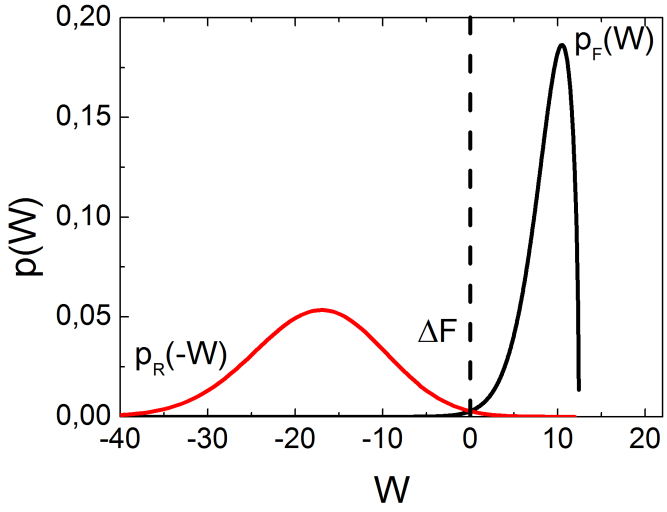


Figure 12.3: **Examples of forward and reverse distributions satisfying Crooks relation**: The reverse distribution is given by the forward distribution multiplied by a monotonically increasing function of W. Both distributions intersect at ΔF.

This result is closely related to various *fluctuation theorems* derived for entropy production in out-of-equilibrium systems [223]. Equation (12.21) has a number of implications. First, it tells us a way to compute the forward distribution by multiplying the reverse one by a monotonically increasing function of W. This implies that the mean of $p_F(W)$ is located to the right of the mean of $p_R(-W)$, which is what Eq. (12.20) says. Moreover, Eq. (12.21) implies that the forward and reverse distributions intersect at $W = \Delta F$. Finally, we can recover Jarzysnki equality as a corollary or Eq. (12.21), by multiplying both sides by $p_R(-W)e^{-W/k_B T}$ and integrating over all values of W.

12.4.4 Computing Equilibrium Free Energies from Nonequilibrium Work Measurements: Practical Issues

In practice, the work distribution is sampled with a limited number of realizations. In such way, although Jarzysnki equality is exact, we compute the Jarzynski estimator:

$$\Delta F_J = -k_B T \log \frac{1}{N} \sum_{i=1}^{N} e^{-W_i/k_B T}, \tag{12.22}$$

where N is the number of nonequilibrium realizations available to sample the work distribution. We define the bias of our estimator as $B_N = \Delta F_J - \Delta F$, where ΔF is the target free energy difference as obtained from infinite sampling. The bias is a statistical measure of the systematic error due to finite sampling, and Jarzynksi equality provides an *unbiased* estimator, this is $B_N \to 0$ for $N \to \infty$. Nevertheless, this does not mean that the free energy value can be easily obtained from it. One of

the biggest challenges in using Jarzynski equality is the exponential average, which suffers from poor convergence.

If the work distribution is broad with respect to k_BT, only a few trajectories at the low work tail contribute significantly to the weighted average, while the remaining events have and exponentially small weight. In physical terms, when the transformation is conducted rapidly, most trajectories do not sample relevant regions of phase space, resulting in broadened work distributions that reflects an increasing relevance of dissipation.

The errors associated to finite sampling in the Jarzysnki estimator, have been a matter of interest since the advent of the technique [224]. For example, it is now known that Jarzysnki bias starts at $B_1 = \langle D \rangle$ and decreases monotonically as N increases, approaching to zero in the limit of infinite sampling [225]. Also, some work distributions allow a direct analytic treatment and thus an exact derivation of the bias. Gaussian work distributions are of particular interest, as they appear in the near-equilibrium regime, when the change $\lambda : A \to B$ is done sufficiently slowly, but can appear under different situations [226]. Here,

$$\Delta F = \langle W \rangle - \sigma^2 / 2k_BT, \tag{12.23}$$

where σ^2 is the variance of the work distribution.

When the reverse process is available, one can combine forward and reverse trajectories to compute an *optimal* free energy estimator. Starting from Crooks relation (Eq. (12.21), we can rewrite it as an average

$$\int f(W;\Delta F) e^{-W/k_BT} p_F(W) dW = \int f(W;\Delta F) e^{-\Delta F/k_BT} p_R(W) dW, \tag{12.24}$$

by multiplying both sides by an arbitrary function $f(W;\Delta F)$. This equation becomes an implicit equation for ΔF. As Bennet proved, the function that minimizes the average squared error of the estimated free energy is $f = [e^{-(W-\Delta F)/k_BT}/N_f + 1/N_R]^{-1}$ where N_F and N_R are the number of forward and backward realizations [227]. This result can be also obtained by using a maximum-likelihood approach [228]. Then, the value ΔF satisfies the relation

$$\sum_{i=1}^{N_F} \frac{1}{1 + \frac{N_F}{N_R} \exp[(W_i - \Delta F)/k_BT]} = \sum_{j=1}^{N_R} \frac{1}{1 + \frac{N_R}{N_F} \exp[-(W_j - \Delta F)/k_BT]}. \tag{12.25}$$

This is now known as Bennet free energy estimator and it can be understood as the maximum likelihood estimator of the free energy given a set of forward and reverse non-equilibrium work mesurements, starting from Crooks fluctuation theorem [228]. It is the minimum variance estimator of all asymptotically unbiased estimators.

Chapter 13

Experimental Analysis of DFS Experiments on Mechanical Unbinding of FNR:Fd and FNR:Fld

This chapter focuses on the experimental analysis developed in [229]. We propose an analysis method for extracting meaningful free energy magnitudes from DFS experiments for the unbinding of biological complexes. The method is applied to the unbinding of two protein:protein complexes via AFM experiments. We present in this Chapter the analysis procedure and its application to the particular biological complexes. In order to get a full understanding of the results, we need to propose a new shape for the free energy profile governing the process. This task in undertaken in next chapter.

13.1 Motivation

In the work developed in the present and next chapters we present a detailed analysis and discussion of DFS experiments for mechanical dissociation of biological complexes [229]. Here, a protein:protein or protein:ligand complex is forced to dissociate by applying a mechanical bias through some single-molecule technique.

Our principal objective is to obtain meaningful physical insight about the studied complexes. In particular, we focus on recovering information about the free energy landscape governing the process by analyzing the force response of the system. This is possible thanks to the two theoretical frameworks reviewed in Chapter 12.

In principle, a combined application of both theoretical frameworks would provide a global picture of the kinetic and equilibrium characteristics of the system, given the joint recovery of the the free energy barrier $\Delta G^{\dagger}$ and the dissociation free energy ΔG^{0}. Nevertheless, in order to understand properly these magnitudes, they should fit together within a suitable shape of the underlying free energy profile.

A bare analysis of the DFS experiments for the two protein:protein complexes renders $\Delta G^{\dagger}$ and ΔG^{0} values which are hard to understand with the conventional shapes for the free energy profile. As we discuss through this Chapter, this finding is not unique for our two particular complexes, but it seems rather ubiquitous in biological complexes. This motivates us to propose a new shape for the free en-

ergy profile governing this kind of processes which satisfies the discrepancy we meet. Based on this profile, a phenomenological model for mechanical unbinding of biological complexes is proposed, which helps us to validate the vision of our system and the analysis protocol.

In this regard, we organize these two Chapters as follows. We devote the present Chapter to the analysis of the experiments. The biological system is properly introduced and so is the experimental set-up. Next, we review carefully our analysis procedure, involving first the identification of curves with an unbinding events, and the physical interpretation of the experimental output. Next, we apply it to the experiments, rendering values for $\Delta G^{\dagger}$ and ΔG^{0}, discussing the problem we find. Chapter 16 focuses on the proposal of a physical model to understand the experiments, including a new free energy profile to govern the mechanical unbinding events. We analyze numerical simulations on this model by means of the same protocol we followed with the experiments, validating its robustness. Finally, the joint results are discussed in the frame of this free energy profile, whose implications are central for the success of the analysis protocol, matching also the biological consequences on the complexes.

13.2 The Biological System

13.2.1 Force spectroscopy Experiments on Biological Complexes

The particular biophysical problem is mechanical unbinding of biological complexes with DFS experiments. They are formed by the association of two molecules, typically some ligand or small protein docked in the binding pocket of a larger one. This large protein is immobilized in a substrate, while the pulling device is functionalized with the other one. An unbinding experiment has two well differentiated stages (see Section 13.4). First, both molecules are approached to form the stable complex. Then, they are pulled from each other, producing the mechanical dissociation.

In principle, the process of mechanical unbinding is governed by a free energy landscape represented along the pulling direction, the reaction coordinate of the process. The initial state is the bound complex, and the final one the unbound complex, characterized by an absence of interaction. In this sense, the system is characterized by two energy magnitudes, a free energy barrier $\Delta G^{\dagger}$—which controls the kinetic properties—and the dissociation free energy ΔG^{0}—free energy difference between the unbound and bound complex, which controls the thermodynamic behavior. In principle these are the magnitudes we aim to recover by measuring rupture forces through DFS experiments.

Mechanical unbinding experiments contrast with mechanical unfolding of biomolecules (nucleic acids or proteins) in a basic feature. When pulling a biomolecule, the information about the folding landscape of proteins or RNA/DNA hairpins is obtained by forcing a transition between the folded structure and a fully stretched structure [18, 84]. The main difference is that in mechanical unfolding experiments there is always an underlying stretching of a polymer which affects the final state. In mechanical dissociation experiments once the complex is unbound the interaction is lost.

In other words, mechanical unfolding is a completely different process from thermally-driven unfolding or through any other denaturant. This is because the final state has a low entropy, as we are stretching the molecule. In thermal denaturation, the unfolded configuration is an ensemble of random coiled molecules, with large conformational entropy. In this sense, a direct comparison between the mechanical unfolding landscape and the thermal unfolding landscape is a hard issue.

Bridging the Gap between Single Molecule and Bulk Experiments

In mechanical unbinding events, the initial state is the bound state, and the final state is the unbound state, the same as for thermal spontaneous dissociation. For example, in a Isothermal Tritiation Calorimetry (ITC) experiment [230], species A is kept on a cell, while molecule B is injected. Complexes AB form, and the interchanged heat can be analyzed to obtain the binding free energy $\Delta G_C = G_B - G_U$, where G_B stands for the bound complex and G_U for the unbound one. If the complex is stable, $\Delta G_C < 0$.

In a DFS experiment, our initial state would be characterized by G_B and the final by G_U, so the free energy difference between both states is $\Delta G^0 = \Delta G_U - \Delta G_B = -\Delta G_C$, which should coincide with the calorimetry value given that it is an equilibrium value, and thus independent of the path employed.

The only difference between both experimental procedures comes from the conformational entropy contribution of the thethering [18]. While in calorimetry experiments the molecules involved have complete conformational freedom, in the DFS experiments the complexes are restrained. In this sense we could argue $\Delta G^0 - (-\Delta G_C) = S_c$, where S_c is some conformational entropy contribution. Nevertheless, if we attend to the involved degrees of freedom in every case, this should not be larger that $1 k_B T$, so not too significative. This implies that, if we are able to recover ΔG^0 through DFS measurements, the obtained value would be comparable to those obtained from thermodynamic bulk assays.

13.2.2 FNR:Fd and FNR:Fld two Binding Partners for a Common Substrate

The systems we study herein consist of the complex form by the flavoenzyme ferredoxin-NADP$^+$ reductase (FNR; being NADP$^+$ the nicotinamide adenine dinucleotide phosphate), which contains a flavin adenine dinucleotide (FAD) group and its two different binding partners, ferredoxin (Fd) with a [2Fe-2S] cluster, and flavodoxin (Fld), with a flavin mononucleotide (FMN) group, from the cyanobacterium *Anabaena* PCC7119 [231]. Two Fd or Fld molecules interact sequentially with FNR for the step-wise transfer of two electrons. Finally, reduced FAD from FNR is used to convert NADP$^+$ into NADPH. Both the enzyme and its redox partner form a transient complex to transfer electrons in the photosynthetic electron-transfer chain.

This redox system is of particular interest as two proteins of different nature (Fd and Fld) interact at the same site of FNR [232]. In this sense, it has been revealed that this system can be considered as a paradigm for investigating which are the key issues determining the complex formation and the electron transfer process [230, 231, 233].

The choice of these two protein:protein complexes for our study is of particular

interest, due to the common features both share, and also the differences. First, both complexes show a similar thermodynamic affinity [230], meaning that their dissociation free energy ΔG^0 is rather similar. Nevertheless, they are known to display different interaction mechanisms [230, 231, 233–237], given their biological role. In particular, Fld replaces Fd under iron deficient conditions [230, 234], which is able to bind to the same site of the enzyme, but in a more non-specific and promiscuous manner. This less strongly and durably bond is known to decrease the efficiency in transferring electrons. The difference in size of the interacting surfaces and of the key residues involved in the complex stabilization should lead to a different kinetic behavior under the presence of an external load.

In this sense, they determine a remarkable model for our purpose, which is to recover jointly the kinetic and thermodynamic properties of the complexes. We should be able to distinguish the differences in the kinetic behavior, while obtaining similar thermodynamic properties.

13.3 Experimental Set-Up

We analyze DFS experiments realized by Dr. Carlos Marcuello and Dra. Anabel García Lostao [238, 239] for mechanical unbinding of the two protein:protein complexes FNR:Fd and FNR:Fld. Experiments were carried out using the force spectroscopy mode in a Cervantes Fullmode SPM system (Nanotec Electrónica S.L. Spain), in the Advanced Microscopy Laboratory (LMA, INA). Figure 13.1 shows a schematic representation of the experimental set up. The involved elements, the protein:protein complex, a PEG polymer linker and the AFM cantilever, which is responsible of exerting the pulling force.

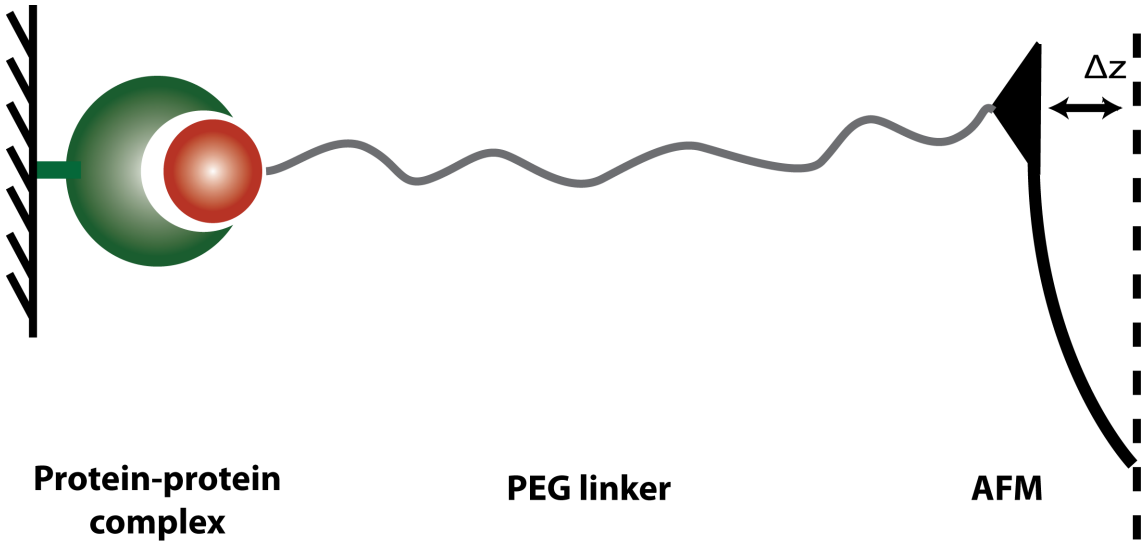


Figure 13.1: **Schematic picture of the experimental set up for DFS unbinding experiments**: FNR molecules are immobilized and oriented in a mica surface. AFM cantilever is functionalized with PEG polymer linkers and Fld/Fd proteins.

FNR molecules were labelled, separated and immobilized on mica surfaces, as described previously [240]. Maleimide-terminated flexible polyethylene glycol (PEG) linker silicon nitride AFM cantilevers with nominal spring constant of 20pN/nm

(Novascan Technologies Inc, Ames, USA) were used. PEG polymer has a nominal stretched length of 20nm (PEG MW 3400) and persistence length of $0.37nm$. 42M thiolated-Fld/Fd, labeled and purified as reported in [238, 240] were incubated on the maleimide-PEG-cantilevers in PBS, EDTA, pH 7.0 for 1 hour and washed extensively with the same buffer. Labelling and subsequent immobilization steps were performed to orient the interaction surfaces of both proteins one towards each other, which optimizes the recognition ability and the collection of successful unbinding events in DFS scans.

We register several hundred force-distance cycles for Fd and Fld-cantilever/FNR-mica approaches at different loading rates, ranging between $2 - 80 \times 10^3 pN/s$. The protocol for selecting the appropriate unbinding force-curves is detailed in Section 13.4. Negative control experiments were also carried out by blocking the available FNR sites by incubating the samples with 0.70nM Fld. This is further detailed in [238, 240] and Section 13.4.

13.4 Analyzing DFS Experiments

We carry out DFS-AFM experiments in the force-extension modality. In this sense, the output of the experiments is a force versus extension curve which should contain information about the studied complex. Nevertheless, single-molecule experiments have typically a low success rate. This is, most of the individual experiments fail in achieving an unbinding event. Thus, those curves were an unbinding event is identified must first be selected, in order not to introduce false data into our final analysis. This should be done following some careful criterion, where unbinding events are identified through some fingerprint they lay out.

Force-Extension Curves

A complete force-extension curve is a cycle made up of two different stages, the approach and the retraction. First, the AFM tip coated with one of the molecules approaches at constant speed towards the substrate. Second, the tip is retracted at the same velocity to reach the initial position[1].

Figure 13.2 shows an sketch of a complete force-extension cycle. The cycle starts at point A, with the functionalized tip far enough from the substrate so that no interaction is measured. Then the tip approaches to the substrate, until is close enough so that the two molecules can interact. This occurs at point (B), where the cantilever deflects towards the sample. Through this step, the two molecules can interact, providing that the orientation is adequate. From (B) to (C) the pushing of the tip is maintained, and a higher deflection measured, due to the repulsion forces from substrate and tip. The pushing stops at a certain contact foce F_c, low enough to avoid damage of the sample.

In the second stage of the cycle, the tip is retracted from the sample at some constant velocity V. During the retraction, adhesion forces are measured as a hysteresis in the curve, from (D) to (E). At some point the spring force is higher than the interaction forces, and the cantilever pulls off sharply, going back to the original position (F). This jump from (E) to (F) provides the measure of the unbinding force.

[1] As explained in Chapter 11, the surface is actually what is retracted thanks to a piezoelectric actuator, not the tip

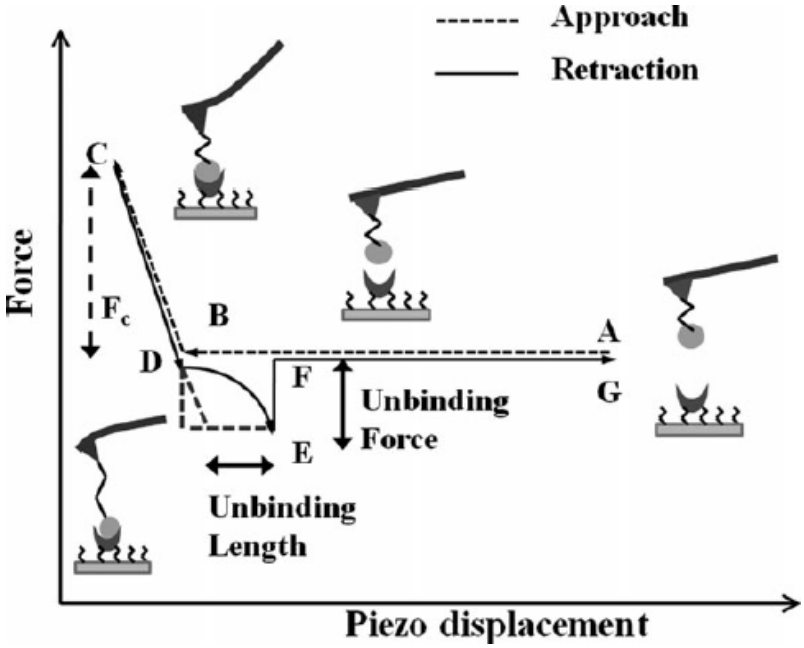


Figure 13.2: **Sketch of a complete aproach-retraction curve for a ligand:receptor rupture experiment**: The approach stage comprises a phase with no interaction (from (A) to (B)), and then the contact phase (B) to (C). In the retraction stage, once the contact region is overcome ((C) to (D)), the unbinding event occurs in the (D) to (E) to (F) phase, where a force response takes place until the jump-off ((E) from (F)) indicates that the complex has been mechanically unbound. (Picture taken from [241]).

In principle, this approach-retraction process does not necessarily forms a successful bond between the two biomolecules, and nonspecific interaction between the tip and the substrate can occur. These kind of events do not carry any significant information about the biological system and must be ruled out of the final analysis. In next section we review the practical way we employ to identify specific rupture events.

Selecting Curves with Specific Events

Every approach-retraction trajectory can give rise to different curves, hinging upon the kind of molecular interaction which has occurred. Nevertheless, only those containing a binding-unbinding event between the two molecules of interest have physical relevance. In this way, it is useful to classify the possible curves that may arise, in order to have clear criterion to select the curves.

Figure 13.3 shows six possible real AFM curves showing different cases which might be observed. Curve 1 shows no detectable event, as the approach and retraction curves are totally superimposed. Obviously it should be discarded. Curve 2 shows a jump-off event but due to some non-specific interaction of the tip with the surface. This can be identified in the slope of the curve, which remains the same during the retraction in the contact region. It contains no biological information, so is also to be discarded.

Curves 3 and 4 are examples of specific events. The slope of the curves changes in the retraction process. This means that, when retraction starts, the system becomes stretched. Particularly, the influence of the linker polymer is observed, as a WLC-model should be adequately superimposed to the part of the curve in the contact

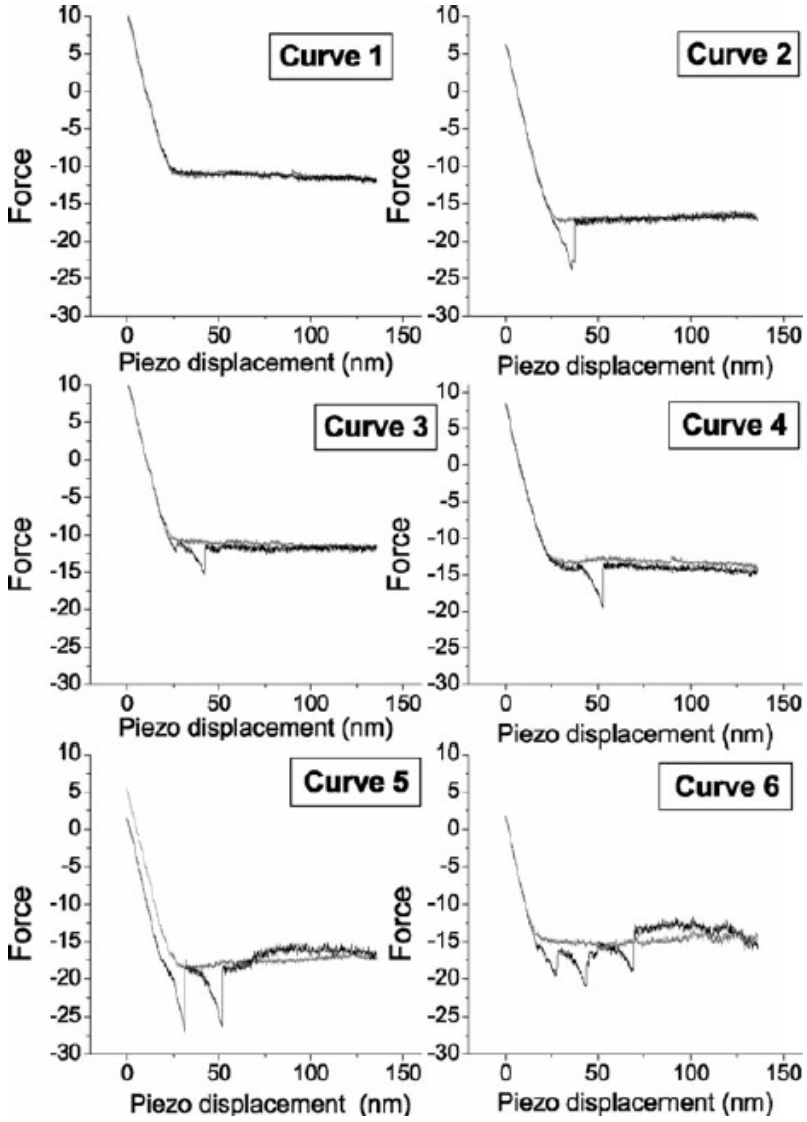


Figure 13.3: **Examples of possible individual curves in DFS-AFM experiments**: Curves (1) and (2) are to be discarded, as (1) shows no interaction and (2) a non-specific interaction with the surface. Curves (3) and (4) show single specific events, while (5) and (6) specific multievents. Nevertheless, curve (6) should be also discarded as it is difficult to identify the actual unbinding events. (Picture taken from [241]).

region. This is typically taken as the fingerprint of the specific event curve. Now, the rupture force is taken as the jump-off force in the retraction curve.

Finally, curves 5 and 6 shows examples of several specific ruptures. This is due to a variety of reasons. The AFM cantilevers are not functionalized with a single polymer linker and protein, but rather several [242]. In this sense, more than a single rupture event can occur at once, giving rise to a multipeaked curve. This is observed in curve 5. Curve 6 must be discarded, as several specific-nonspecific events are superimposed. As a common strategy, if there is doubt it is always better to discard the curve.

The rate of success in single-molecule experiments is usually very low, specially when random strategies are used. For improving the success rate functionalization and orientation efforts are taken, allowing to reach success levels of over the 50% [238, 240] (see Fig. 13.4).

Obtention of Rupture Force Distributions

We build rupture force histograms analyzing the subsets of curves with clear specific unbinding events. Typically (see Fig. 13.4, left), these histograms have long tails in the high force region. This is due to multievents, which overestimate the rupture forces. It is often assumed [241], that two rupture events result in a rupture force double to the single event and so on. In this sense, a possible strategy is to deconvolute the experimental distribution into individual distributions accounting for single, double, triple... events. This is usually done by fitting to successive gaussians, such that $\langle f \rangle_2 = 2\langle f \rangle_1$, (where the subindex indicates the number of expected rupture events). Nevertheless, we think that this not an optimal strategy, given that the theoretical distributions are not gaussian functions, not with Dudko-Hummer-Szabo theory, nor with Bell-Evans [203, 204]. The analysis protocol we propose (see Section 13.5) is largely unaffected by such multievents, as we prefer to take the most probable rupture force (peak of the distribution) rather that the average value calculated from the individual gaussian distributions. This reduces largely the error, as seen when comparing [229] and [238]. Also, the position of such maximum is minimally influenced by the underlying multievents distributions.

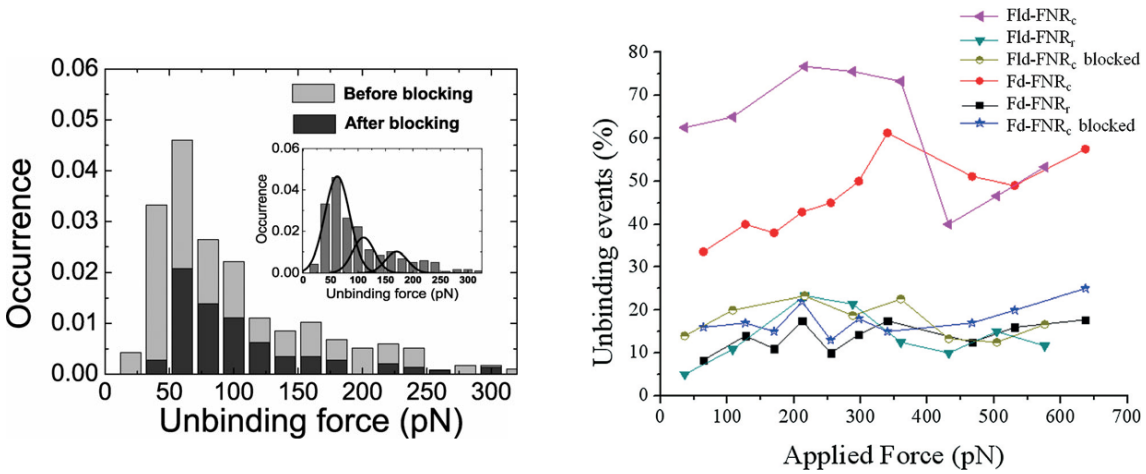


Figure 13.4: **Rupture force histogram and binding success for different experimental strategies**]: (Left) The rupture force histogram is built prior selection of the curves containing specific events. Inset shows a possible identification of multievents, not optimal in our opinion. Rupture force histogram under blocking conditions is compared, with a clear decrease in the specific events rate. (taken from [241]) (Right) Clearly functionalization strategies are a key element in DFS experiments (taken from [238]).

Usually, in order to validate the experimental protocol, blocking experiments are performed. Here, the ligand is not only in the functionalized tips, but also previously diluted in the solvent cell, so that some complexes should be already be formed, prior to the DFS experiments. This should results in a greatly smaller chance of unbinding events, and thus force distributions with reduced occurrence, as seen in Fig. 13.4.

13.5 Analysis Protocol: Free Energy Barriers and Dissociation Free Energies from Force Measurements

We propose a joint analysis protocol to recover the free energy magnitudes $\Delta G^{\dagger}$ and ΔG^{0} from DFS experiments. We extract both values in an independent way by combining two theoretical approaches, described in Chapter 14.

Extracting the free energy barrier

Force-spectroscopy theory, allows estimating the escape barrier and the position of the transition state from the dependence of the average -or most probable- rupture force as a function of the loading rate (see Section 14.3). As developed by Dudko *et. al.* [204]

$$f^{*}(r_f) = \frac{\Delta G^{\dagger}}{\nu x^{\dagger}} \left[1 - \left(-\frac{k_B T}{\Delta G^{\dagger}} \log \frac{r_f x^{\dagger}}{k_0 k_B T} \right)^{\nu} \right]. \tag{13.1}$$

Here f^{*} is the most probable rupture force, and $r_f = V\kappa$ is the loading rate, where V is the pulling velocity and κ the effective spring constant of the pulling device, while $\Delta G^{\dagger}$ is the free energy barrier height, $x^{\dagger}$ the position of the transition state (maximum of the barrier) and k_0 the intrinsic rate constant, all in the absence of force.

Parameter ν determines the particular shape of the profile. With $\nu = 1$ we recover Bell-Evans expression [203]. Expression Eq. (13.1) was originally developed for two particular cases [204], $\nu = 1/2$ and $\nu = 2/3$ which assume respectively a parabolic cusp potential and a cubic potential. Years later, the generality of Eq. (13.1) was proved for any polynomial potential of order n with $\nu = (n-1)/n$ [243]. Nevertheless, we assume here the cubic approach, as any analytical potential can be approximated by a cubic potential when expanded around the inflection point in the vicinity of the escape force [204].

We highlight the fact that equation Eq. (13.1) expresses the dependence of the *typical or most probable* rupture force with the pulling rate, in opposition to Eq. (12.11) in Section 12.3, which expressed the dependence of the average rupture force. As the shape of the rupture force distribution is known, it is easy to change from one expression to the other. We prefer to work with the *mode* of the distribution rather than the average, as explained in Section 13.4.

Calculating Equilibrium Free Energy Differences

The dissociation free energy ΔG^{0}, difference in free energy between the bound and unbound states, is an equilibrium magnitude. This allows us to use Jarzysnki equality to estimate it from non-equilibrium work measurements, which is the case of constant-rate trajectories. Given a number of N pulling trajectories, we calculate the Jarzysnki estimator ΔG^{0}_{J} of the *actual* dissociation free energy ΔG^{0} as

$$\Delta G^{0}_{J} = -k_B T \log \frac{1}{N} \sum_{i=1}^{N} e^{-W_i / k_B T}, \tag{13.2}$$

where W_i is the non-equilibrium work performed over the i-th unbinding trajectory. As stressed in previous chapters, Jarzynsnki equality requires a proper definition of the non-equilibrium protocol we are following. We must have some control parameter λ to switch the Hamiltonian of the system from an initial *equilibrium* state $\lambda = A$ to another defined one $\lambda = B$. The work W performed on the system is the integral along the λ curve.

In our case the non-equilibrium protocol is properly defined, as we have the control parameter $\lambda = Vt$ which we switch from $\lambda = 0$ (at t=0) to some $\lambda^\dagger$ where the system is unbound. In our particular case, the final value of $\lambda^\dagger$ is not critical as long as the rupture event have taken place. Dragging the unbound molecule would not contribute significantly to the work but for the dragging force $F_d = \gamma V$, which is not significant. Also, it must be stressed that the work is properly defined when integrating the force-extension curve as a function of the *control parameter* λ rather than the stochastic variable γ [244]. This is a common mistake which might lead to misestimations of the actual free energies [245].

In our case, given a rupture trace $f(\lambda)$, the work is calculated as

$$W_i = \int_0^{\lambda^\dagger} f(\lambda)\mathrm{d}\lambda = \int_0^{\gamma^\dagger} W_{WLC}\mathrm{d}\gamma + \frac{1}{2}\frac{(f^\dagger)^2}{\kappa}, \tag{13.3}$$

where F_{WLC} is the force-extension curve of a Worm-Like-Chain (WLC) model[2], κ the spring constant of the AFM, λ is the control parameter (position of the cantilever), and γ the distance to the tip of the cantilever (see Section 14.1 and Fig. 14.2 for a more careful definition of these coordinates). This expression is computed easily by considering the change of variables from λ to γ, $\lambda = \gamma + f/\kappa$ (assuming negligible change in the molecular coordinate x_p) and equilibrium at the tip of the AFM. Equation (13.3) is equivalent to the work accumulated by the whole pulling device over the unbinding process, polymer linker and linear spring together.

In principle, Jarzynski equality is exact, and thus independent of the pulling rate. Nevertheless, we have already discussed about the poor convergence problems the exponential average causes. At very fast pulling rates, we are very far away from equilibrium so a very large number of experiments would be necessary to get a reasonable estimate[3]. In this regard, we calculate ΔG^0_J as a function of the pulling rate, expecting convergence to ΔG^0 as the rate decreases.

Recall that Eq. (13.3) depends only on the rupture force $f^\dagger$, as $\gamma^\dagger$ is determined by the WLC model expression by numerical inversion. This gives a very robust way to estimate ΔG^0_J, as we do not rely on the shape of the rupture force curve, but just on the peak. Given a rupture force distribution $p(f^*)$, we map it directly to a work distribution $p(W)$ by applying Eq. (13.3) and calculate the Jarzynski estimator $\Delta G^0_J(r_f)$ as

$$\Delta G^0_J(r_f) = -k_B T \log \int p(W) e^{-W/k_B T} dW \approx -k_B T \log \sum_{i=1}^{N} e^{-W_i/k_B T} p(W_i)\Delta W, \tag{13.4}$$

[2]The Worm-Like-Chain model is one of the most used models for polymers. Its force-extension curve is $FL_p/k_BT = 0.25\,(1 - x/L)^{-2} + x/L - 0.25$, where L_p is the persistence length and L the contour length of the polymer.

[3]The necessary number of experiments for convergence N scales with e^D, where D is the dissipated work $D = \langle W \rangle - \Delta G$ [246]

given a histogram representation of the distribution built by N bins.

13.6 Results

We analyze DFS measurements for mechanical dissociation of protein:protein complexes formed by flavoenzyme Ferredoxin-NADP$^+$ reductase (FNR) with Ferredoxin (Fd) and Flavodoxin (Fld) from cyanobacterium *Anabaena* PCC7119. For each loading rate, the individual curves are analyzed, selecting those with a clear unbinding event following the criteria explained in Section 13.4. We build rate dependent force histograms $p(f^\dagger|r_f)$. These histograms are analyzed with the protocol explained in Section 13.5 to calculate the free energy magnitudes ΔG^0 and $\Delta G^\dagger$.

Free energy barriers for FNR:Fd and FNR:Fld complexes

From each rupture force distribution, we define the most probable rupture force f^* as the bin with largest accumulated number of events $f^* : \max p(f^\dagger)$ We plot f^* as a function of r_f and fit it to Eq. (13.1), obtaining $\Delta G^\dagger$, $x^\dagger$ and k_0.

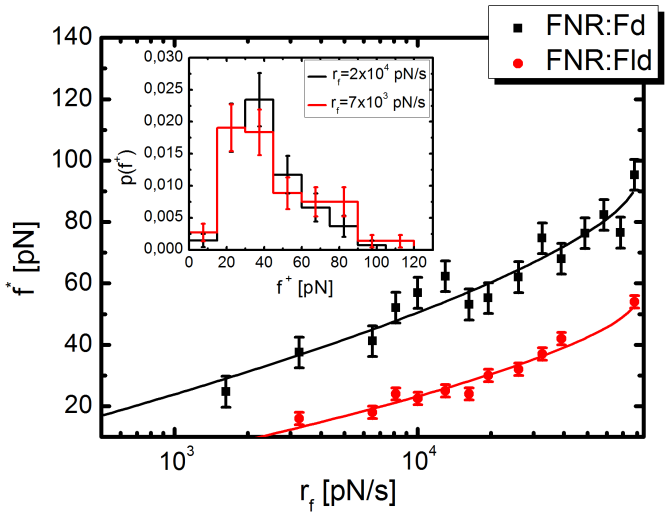


Figure 13.5: **Typical rupture force f^* as a function of the pulling rate r_f**: Solid lines are minimum square fits to Eq. (13.1), showing excellent agreement. Inset shows two rupture force distributions for different rates.

Figure 13.5 shows the typical rupture force f^* as a function of the loading rate r_f for protein:protein complexes FNR:Fd and FNR:Fld. Solid curves are the best fit obtained by minimum squares to Eq. (13.5). Inset shows two examples of rupture force distributions, for different pulling rates. The agreement of the fitting is excellent[4], yielding, for **FNR:Fd**: $\Delta G^\dagger = 6.85 \pm 0.47 k_B T$, $x^\dagger = 0.46 \pm 0.02 nm$ and $k_0 = (8.60 \pm 0.45) \times 10^{-3} s^{-1}$; while for **FNR:Fld**: $\Delta G^\dagger = 4.85 \pm 0.40 k_B T$, $x^\dagger = 0.56 \pm 0.03 nm$ and $k_0 = (1.10 \pm 0.06) \times 10^{-2} s^{-1}$.

This is the first time that kinetic unbinding properties are measured for these complexes, so they cannot be compared to any known result. As we already mentioned, that thermodynamical properties for both complexes are rather similar,

[4] $\chi^2 = 1.2$ for FNR:Fd and $\chi^2 = 1.3$ for FNR:Fld

having similar stability. Nevertheless, we have seen how the free energy barriers are rather different. This is already evident by simple inspection of Fig. 13.5, as FNR:Fld needs lower forces to unbind than its partner FNR:Fd. Also, the free energy barriers seem to locate at a very close distance, $x^{\dagger} \approx 5\,\text{Å}$. We go back to this point on Chapter 14, when discussing the results within a suitable free energy profile shape.

Obtention of the Dissociation Free Energy

We calculate the rate dependent work distributions $p(W|r_f)$ from each rupture force distribution $p(f^{\dagger})$. We apply Jarzynski equality to each work distribution, obtaining a rate dependent Jarzynski estimator ΔG^0_J.

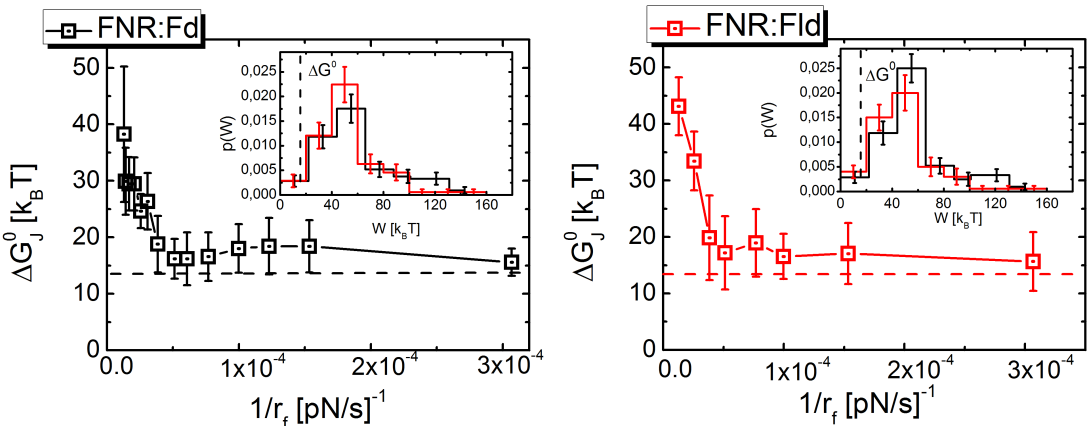


Figure 13.6: **Jarzinski estimator as a function of the inverse rate**: As the pulling rate decreases, the estimator converges to the calorimetry determined free energies (dashed lines). Inset shows two work distributions for different pulling rates.

Figure 13.6 shows Jarzynski estimator plotted as a function of the inverse rate $1/r_f$, where the error bars are calculated with Jackknife resampling method [247]. Inset shows the work distributions for two particular pulling rate values. Dashed lines are the binding free energies obtained from calorimetry experiments for each of the complexes ($\Delta G^0_{FNR:Fd} = 13.5 k_B T$ and $\Delta G^0_{FNR:Fld} = 12.8 k_B T$) [230]. Plotting against $1/r_f$ is due to visualization reasons, as convergence is observed in a more clear way.

We check how the estimator ΔG^0_J converges as we decrease the pulling rate to a particular value, very close to the calorimetry free energy difference. At high pulling rates, the bias $B_J = \Delta G^0 - \Delta G^0_J$ is large because experiments are driven very far away from equilibrium. As we have around $100 - 200$ samples per rate, convergence is poor at high rates. As we approach to equilibrium, for pulling rates $r_f \sim 3 - 20 \times 10^3 pN/s$, the bias tends to zero, so the Jarzynski estimator ΔG^0_J gives the expected calorimetry value.

This last fact is a remarkable one, given that we are comparing results obtained by analyzing single molecule experiments, with results from bulk experiments. The conditions of both experiments are very different, not just because the size of the populations involved in the measures, but also because of the pathways taken. Here we check the stability (free energy difference) of the complexes by applying an external mechanical force, while in calorimetry, this is measured thermally.

Table 13.1: **Free energy barrier height $\Delta G^{\dagger}$, position $x^{\dagger}$ and dissociation free energy ΔG^{0} for some biomolecular complexes**: Typically $\Delta G^{0} > \Delta G^{\dagger}$ can be observed. (a), (b), (c) and (e) are presented in Refs. [243] , [249] ,[250] and [251] respectively. (d) are obtained after an analysis of data given in [251] (seeSI). (f) is obtained in this work and (g) in [230].

Complex	$\Delta G^{\dagger}[\mathrm{k_B T}]$	$x^{\dagger}$ [nm]	$\Delta G^{0}[\mathrm{k_B T}]$
Biotin:streptavidin	13.56^{a}	0.55^{a}	30.9^{b}
Biotin:avidin	11.74^{a}	0.49^{a}	33.7^{c}
LFA-1:ICAM1	8.57^{d}	0.17^{d}	15.5^{e}
LFA-2:ICAM2	7.55^{d}	0.40^{d}	14.3^{e}
FNR:Fld	4.85^{f}	0.56^{f}	12.8^{g}
FNR:Fd	6.86^{f}	0.46^{f}	13.5^{g}

Figure 13.6 shows an apparent bias of $B \approx 3-4k_BT$, so Jarzynski estimator does not converge exactly to the calorimetry value, but rather to an slightly higher one. There are two different contributions which explain this discrepancy. First, the conformational entropy contribution, as argued earlier, answers for an overestimation of around $1k_BT$. The second source is the polymer linker we use in the experiment. The PEG is known to adopt a helical conformation, which unwinds when applying a pulling force [248]. This work work is performed against the system and it is not inverted in unbinding the complex. This difference is known to be of $3k_BT$ [245, 248], value which is approximately equal to the bias we observe.

Nevertheless, this contribution cannot be observed directly in the AFM individual traces. We cannot determine thus if it applies to every single curve, or just to a fraction of them, so it is difficult to estimate the exact magnitude of the effect it would have in the estimation of the ΔG^{0}_{J}. We prefer to remain cautious at that point, given that the results are satisfactory within error bars.

13.7 Discussion: Relation between Dissociation Free Energies and Free Energy Barriers in Mechanical Unbinding of Biological Complexes

We have calculated in previous section the free energy barrier $\Delta G^{\dagger}$ and the dissociation free energy ΔG^{0} for two different protein:protein complexes. Remarkably, the dissociation free energy values matches in both cases the calorimetry free energies. Nevertheless, the results we obtained are somewhat paradoxical, given that we have in both cases $\Delta G^{\dagger} < \Delta G^{0}$. This surprising feature implies that the obtained free energy values cannot be understood within a conventional molecular profile. This picture considers that the particle is initially in a free energy well, and escapes by surmounting a barrier of height $\Delta G^{\dagger}$, relaxing back to the unbound state, at ΔG^{0}, being $\Delta G^{\dagger} > \Delta G^{0}$.

We have searched in the literature for analogous information for different biological complexes, finding that this is rather a common feature of such kind systems. Table 13.1 shows the values of the free energy barrier and dissociation free energy for a total of six different biological complexes, including the ones we determined here.

The condition $\Delta G^{\dagger} < \Delta G^{0}$ is not a particular feature of our system, but appears as a common one for biological complexes of different nature. For example, biotin:streptavidin or biotin:avidin are protein:ligand complexes, while LFA-1:ICAM1 and LFA-2:ICAM2 are protein:protein complexes, as are FNR:Fld or FNR:Fd.

This finding implies that mechanical unbinding of biological complexes cannot be understood through usual molecular profiles, something which motivates us to propose a new free energy profile shape for governing such systems. This new shape has a twofold purpose. First to provide a suitable framework for understanding the magnitudes in Table 13.1. Second, to serve as basis for a phenomenological model for mechanical unbinding of biological complexes, where we can prove our analysis protocol in order to back up its validity. Indeed, we discuss how the particular shape of the free energy profile is tightly linked with the performance of the analysis procedure we proposed above.

Chapter 14

Phenomenological Model for Mechanical Unbinding of Biological Complexes: from Forces to Free Energies

This Chapter focuses on the numerical simulations done in reference [229], in complementarity to the work exposed in Chapter 13. We propose a suitable free energy profile for the mechanical dissociation of biological complexes, where the obtained values for $\Delta G^{\dagger}$ and ΔG^{0} can be properly understood. Additionally, this profile serves as basis for a model to understand DFS experiments on biological complexes. Tuning the physical values of $\Delta G^{\dagger}$ and ΔG^{0}, we perform numerical experiments, which allow us to apply the same analysis protocol to the simulated curves, and recover the free energy magnitudes we chose. This serves us to further validate the analysis procedure and prove its validity. Indeed, the particular shape of the free energy profile turns to have key consequences on the success of our analysis procedure.

14.1 Mesoscopic Model for Mechanical Unbinding of Biological Complexes

We propose a physical model for force-driven unbinding of biological complexes via force spectroscopy experiments. Considering the set-up for this sort of experiments, our model is made up of two ingredients, a phenomenological potential to represent the biological complex and the pulling device. The election of the potential profile is a central decision in the process, at it must accomplish the condition $\Delta G^{\dagger} < \Delta G^{0}$, which characterizes unbinding of biological complexes.

Phenomenological Potential for Mechanical Unbinding of Biological Complexes

The ligand:receptor complex is represented as a brownian particle subject to a one-dimensional potential, which is the free energy profile of the system along the pulling coordinate. This profile must have some characteristics in order to reproduce faithfully the phenomenology of such systems:

1. An Equilibrium well accounting for the bound state.
2. A Free energy barrier $\Delta G^{\dagger}$ at $x^{\dagger}$ to represent the kinetic properties of the system.
3. The Unbound state as a flat region (no interaction), with a free energy difference with respect to the bound state of ΔG^{0} (dissociation free energy).
4. The free energy barrier and dissociation free energy must hold $\Delta G^{\dagger} < \Delta G^{0}$.

We choose a particular shape for such potential which fulfills the four conditions stated before. Figure 14.1 shows a plot of the free energy profile. Mathematically, we choose the following equation:

$$G(x_p) = D(1 - e^{-ax_p})^2 + Ue^{-(x_p - x^{\dagger})^2/b} + F_0[1 + \tanh w(x_p - s)], \tag{14.1}$$

where x_p is the molecular coordinate. This profile reproduces the three relevant regions in the unbinding process. The first term is a Morse potential which accounts for the equilibrium bound state at $x_p = 0$. The second term is a gaussian barrier of height $\Delta G^{\dagger} \approx D+U$, width b and placed at $x_p = x^{\dagger}$. The third term is tanh function which originates a second slope which leads to the unbound flat state $\Delta G^{0} = 2F_0+D$ within a characteristic length of $1/w + s$.

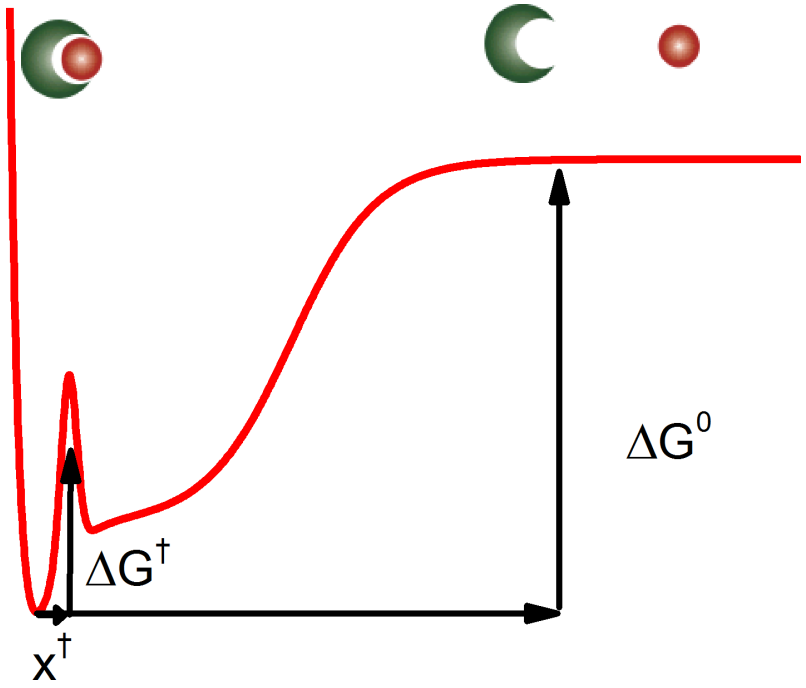


Figure 14.1: **Free energy profile for mechanical unbinding of biological complexes**: The profile is characterized by three regions, first an equilibrium well, accounting for the bound complex, second a steep free energy barrier of height $\Delta G^{\dagger}$, finally a smooth slope leading to the unbound state at ΔG^{0}.

There is an interplay between two different slopes or barriers which appear at two different length scales, the first slope (free energy barrier $\Delta G^{\dagger}$ at $x^{\dagger}$) has a characteristic length defined by $b^{1/2}$, while the latter is controlled by $1/w$. The central feature of our free energy profile proposal is the first *brittle* slope of height

$\Delta G^\dagger$ and a second *smooth* slope which leads to the unbound state. The choice of the particular analytical expression and the exact set of parameters are not crucial for the performance of the potential, as long as this condition is maintained.

Considering the experimental results obtained in Chapter 13 protein:protein complexes FNR:Fd and FNR:Fld, we propose a suitable parameter set which defines a potential with a barrier of height $\Delta G^\dagger = 7.7 k_B T$ at $x^\dagger = 0.5nm$, and a dissociation free energy of $\Delta G^0 = 14.7 k_B T$. We choose $D = 12pNnm$, $a = 3nm^{-1}$, $U = 24pNnm$, $x^\dagger = 0.5nm$, $b = 0.03nm^{-2}$, $F_0 = 24pNnm$, $w = 0.75nm^{-1}$ and $s = 4nm$. This is the parameter set employed for Fig. 14.1.

Modeling the Pulling Device

The pulling device in our experimental set up consists on two different parts, the polymer linker and the AFM cantilever. This is modeled as a nonlinear spring (polymer model) connected in series with a linear spring (AFM cantilever). The polymer is modeled with a Worm-Like-Chain model, whose force-extension response is given by the expression:

$$F_{WLC}(X) = \frac{k_B T}{L_P}\left[\frac{1}{4}\left(1 - \frac{X}{L_0}\right)^{-2} - \frac{1}{4} + \frac{X}{L_0}\right], \tag{14.2}$$

where P is the persistence length of the polymer, L_0 its contour length and X the extension. For the PEG polymer employed in the experiments, values are $L_0 = 20\,\mathrm{nm}$ and $L_P = 0.37\,\mathrm{nm}$. We connect the polymer in series with a linear spring of stiffness $\kappa = 20\,\mathrm{pN/nm}$, considering equilibrium at the contact point, so that the force at the polymer is equal to the one at the spring. This assumption is convenient as prevents us from considering the tip of the cantilever as an additional "particle" in our model—which would force us to give a mass, a damping and other parameters hard to evaluate—and is supported by the scale separation between both systems [215]. The effective stiffness of the whole pulling device is $K_{eff} = (k_{WLC}(F)^{-1} + \kappa^{-1})$, where $k_{WLC}(F)$ is the force-dependent stiffness of the WLC model $k_{WLC}(F) = \mathrm{d}F_{WLC}/\mathrm{d}X$. At low forces, the stiffness of the polymer dominates, as the stiffness is very low for low extension [1]. As force increases, k_{WLC} increases, and κ becomes the dominating stiffness of the system.

Physical Model for Mechanical Unbinding of Biological Complexes

The complete model is just a combination of the potential profile and the model for the pulling device. Pulling experiments are run by retracting the linear spring at a constant velocity, so the polymer applies an increasing force to the brownian particle in the profile. Effectively, this is equivalent to tilting the potential, so the particle is finally able to escape by thermal activation. The distribution of escape forces depends on the pulling velocity, in the same way the experiments did.

Figure 14.2 shows an schematic picture of the experimental set-up and our equivalent model. It is important to notice the three different length coordinates we introduce. Considering the bound state as the reference point, λ is the control parameter, and accounts for the relative position of the cantilever or linear spring. This is the coordinate we change directly, as $\lambda = Vt$. Coordinate γ is an stochastic variable, as

[1] In particular, for the WLC model, at low extension $\kappa_{WLC} \approx 3k_BT/L_P L_0 \approx 1.6\,\mathrm{pN/nm}$

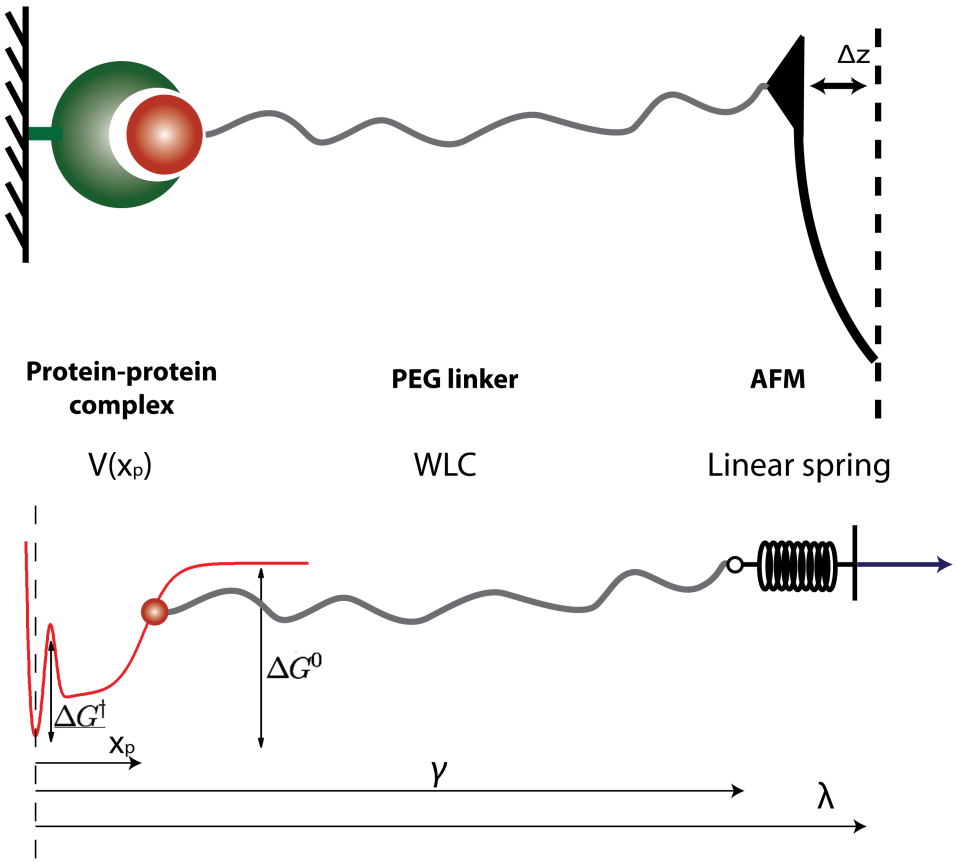


Figure 14.2: **Free energy profile for mechanical unbinding of biological complexes**: The model is determined by two components. The molecular complex is defined by a one-dimensional free energy profile. The pulling device is modeled as a WLC model in series with a linear spring. Unbinding trajectories are simulating by moving the spring at constant velocity until the particle escapes to the unbound region.

it accounts for the position of the tip of the cantilever, or the linking point between the polymer and linear spring. Finally, x_p is the molecular coordinate, setting the position of the brownian particle in the free energy profile. Thus, the extension of the WLC is $X = \gamma - x_p$, and the force equilibrium condition is equivalent to $F_{WLC}(\gamma - x_p) = \kappa(\gamma - Vt)$. Additionally, γ and λ are related by $\lambda = \gamma - f/\kappa$, where f is the force at the tip of the cantilever. This expression neglects the contribution of x_p which is a reasonable assumption, given that $x_p \approx 0$ during the majority of the escaping trajectory.

Simulation Details

Numerical simulations are run in order to mimic the experimental traces. We integrate the Langevin equation of motion for the molecular coordinate x_p on a prefixed protocol. The Langevin equation is,

$$m\ddot{x}_p = -\eta\dot{x}_p - \nabla G(x_p) + F_{WLC}(\gamma - x_p) + \xi(t), \tag{14.3}$$

where m is the mass of the brownian particle (reduced mass of the biological complex), η the viscous damping and $\xi(t)$ the thermal white noise as usual. The stochastic equation is integrated by a Runge-Kutta stochastic fourth order algorithm [81] with a force extension protocol. For each particular experiment at pulling velocity

V, we start at the equilibrium position, where $x_p = 0$ and $\lambda = 0$, increasing parameter $\lambda = Vt$ until $\lambda = 40\,\text{nm}$ to ensure that the rupture event has taken place, as the polymer length is $L_0 = 20\,\text{nm}$. For each time step, the force term in equation Eq. (14.3) is calculated by numerical inversion of the equilibrium condition $F_{WLC}(\gamma - x_p) = \kappa(\gamma - \lambda)$, given a value of λ. We run a total of 10000 realizations for each pulling velocity.

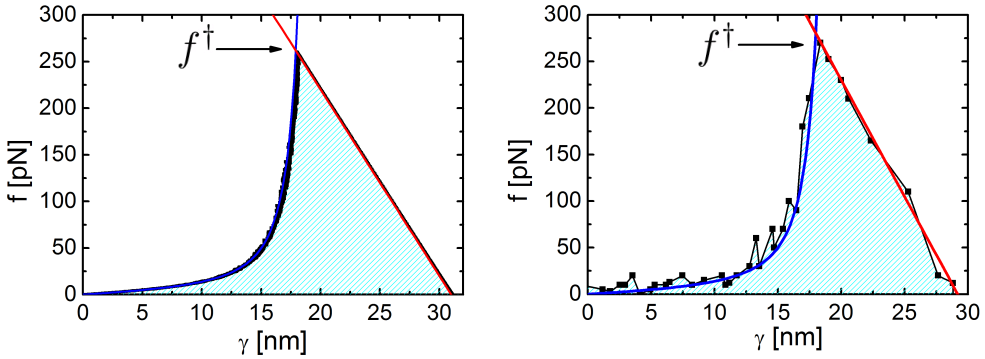


Figure 14.3: **Rupture $f-\gamma$ curve as obtained from numerical integration of the model (left) and as measured in the experiments (right)**: The similarity between the experimental measurements and the numerical calculations is complete. In the $f-\gamma$ representation, force rises with a WLC model (blue solid line) until the rupture force $f^\dagger$ is reached. Then the system relaxes with the stiffness κ (red solid line).

Figure 14.3 shows a comparison between a simulated rupture $f-\gamma$ curve (left) and an experimental one (right). Clearly, numerical simulations on the proposed model are able to reproduce faithfully the observed phenomenology. The different regions in the $f-\gamma$ representation are clear. First, the force increases following a WLC model of $L_0 = 20\,\text{nm}$ and $P = 0.37\,\text{nm}$ (blue solid line in Fig. 14.3). Then a discontinuity occurs when the system escapes (mechanical unbinding event) at $f = f^\dagger$. Then the system relaxes following a linear relation of slope κ (red solid line in Fig 14.3). The area inside (dashed light blue area) is the total work performed on the system.

The units of our simulation are the following. We use pN units for force, nm units for length, m for the mass unit, and time unit of $(m \cdot \text{nm/pN})^{1/2}$. Simulations are carried out at room temperature $T = 4.1\text{pNnm} = k_BT$. The damping in the normalized time units is $\eta = 10$, so we work effectively in the overdamp regime.

14.2 Results on the Numerical Simulations of the Mesoscopic Model

We run numerical simulations on the physical model for mechanical unbinding, setting a free energy profile of known barrier $\Delta G^\dagger$ and dissociation free energy ΔG^0. We apply the same analysis protocol described in Section 13.5 to recover both values in order to prove the validity of the analysis procedure and the suitability of the proposed shape for the free energy profile.

We simulate the numerical experiments by integrating the Langevin equation of motion for a potential profile with $\Delta G^\dagger = 7.7k_BT$, $\Delta G^0 = 14.7k_BT$ and $x^\dagger = 0.5\text{nm}$. These values reproduce approximately those found for the protein:protein

complex FNR:Fd, and are realized with the parameter set shown in previous section. We choose pulling rates so that the most probable rupture forces span over the experimental range $f^* \sim 20 - 100$pN. We run 10^4 realizations for each pulling rate, saving in each case the rupture force $f^\dagger$ as the highest force the system reaches. We build force histograms $p(f^\dagger|r_f)$ and calculate f^* as the mode of the distribution. Force histograms are mapped into the work histograms using Eq. (13.3), in order to apply Jarzynski equality for computing the equilibrium free energy difference.

Recovering the Free Energy Barrier

Figure 14.4 shows the most probable unbinding force as a function of the pulling rate. Black square points are the mode of the rupture force distributions, and the error bars the width of the bins employed to build the histograms. Inset shows two rupture force histograms at two different pulling rates. Red solid line is best least square fit to Eq. (13.1), from which we obtain $\Delta G^\dagger = 7.28 \pm 0.20 k_B T$, $x^\dagger = 0.35 \pm 0.08 nm$ and $k_0 = 10.88 \pm 0.12 t^{-1}$, where t is the adimensional time units. The agreement between the fitting protocol and the simulated data is excellent, as the free energy barrier is recover with great accuracy, so is the position of the transition state.

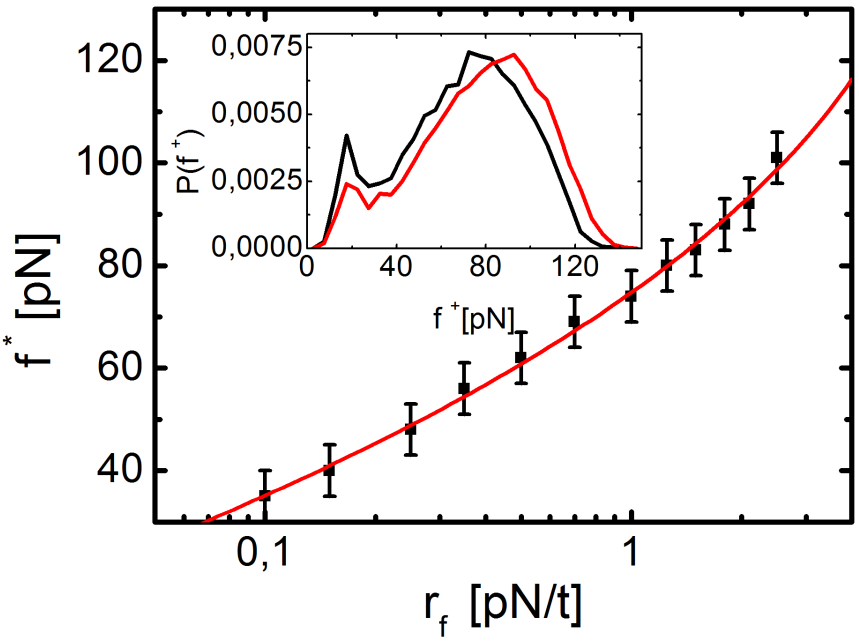


Figure 14.4: **Typical rupture force as a function of the pulling rate for the numerical simulations**: Numerical data is fitted to Eq. (13.1) showing excellent agreement, and recovering successfully the free energy barrier. Inset shows the two rupture force distributions.

Equation (13.1) assumes an escape over a cubic barrier, pulling done with a soft linear spring. Our model is more complex, as the shape of the free energy profile is rather specific, and we pull with a WLC polymer in series with a linear spring, this is, an overall non-linear spring.

The soft spring assumption considers that $\kappa_U \ll \kappa_P$, being κ_U the effective spring constant from the potential profile, calculated from the curvature at the equilibrium well, and κ_P the effective constant of the pulling device. In our particular case,

$\kappa_V \approx 2Da^2 \approx 216pN/nm$, while κ_P depends on the force. At low pulling forces, the WCL model dominates, and $\kappa_P \approx 1.6$pN/nm and the soft spring approximation is satisfied. As we increase the force, the stiffness of the WLC rises greatly, but then, being connected in series, the stiffness of the cantilever dominates, and $\kappa_P \approx \kappa = 20pN/nm$, so the soft spring approximation applies also.

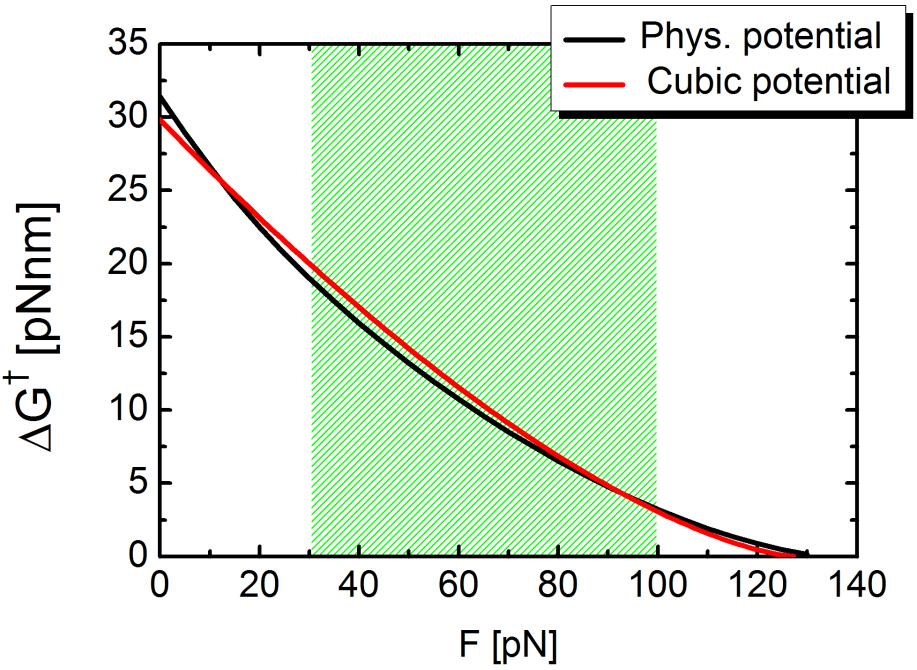


Figure 14.5: **Dependence of the height of the free energy barrier with the pulling force for a cubic potential and for our free energy profile**: The dependence for both profiles is very similar in the unbinding region (green dashed area), justifying using Eq. (13.1) with $\nu = 2/3$.

Regarding the cubic potential approximation, any analytic potential with a barrier, can be approximated as a cubic potential about the inflection point, in the vicinity of the escaping event. We justify this fact for our potential in the range of forces in which the experiments are performed. Figure 14.5 shows the dependence of the heigh of free energy barrier as a function of the applied force. The parameters are those for the barrier employed in this analysis. Red solid line is the dependence for a cubic potential, which can be computed analytically [207], as $\Delta G^\dagger(f) = \Delta G_0^\dagger(1 - f/f_c)^{3/2}$, where $\Delta G_0^\dagger$ is the free energy barrier in the absence of force, and f_c is the critical force, defined as the force at which the barrier vanishes $\Delta G^\dagger(f_c) = 0$. For a cubic potential this is exactly $f_c = \Delta G^\dagger/(\nu x^\dagger)$, where $\nu = 3/2$ [207].

Black solid curve shows the dependence of the free energy barrier height with the force for the potential profile defined by Eq. (14.1). The curve has been computed numerically, as the dependence is not analytical. The critical force is $f_c \approx 130pN$, which gives an effective $\nu = 0.483$, quite close to the behavior of a quadratic potential. The green dashed shows range of forces over which the unbinding takes place. Both dependences are quite similar in the range, explaining why fitting to a cubic potential is a reasonable choice in the physical range of pulling forces.

Recovering the Dissociation Free Energy

Figure 14.6 shows the Jarzynski estimator ΔG_J^0 as a function of the inverse of the pulling rate, where the error bars were calculated with Jackknife resampling method. Blue solid line indicates the unbinding free energy set in the model $\Delta G^0 = 14.7 k_B T$. Clearly, the estimator captures this value, converging as the pulling rate decreases. The convergence is much better than in the experimental case, as we are averaging over 10^4 realization, compared to the 10^2 experimental curves used.

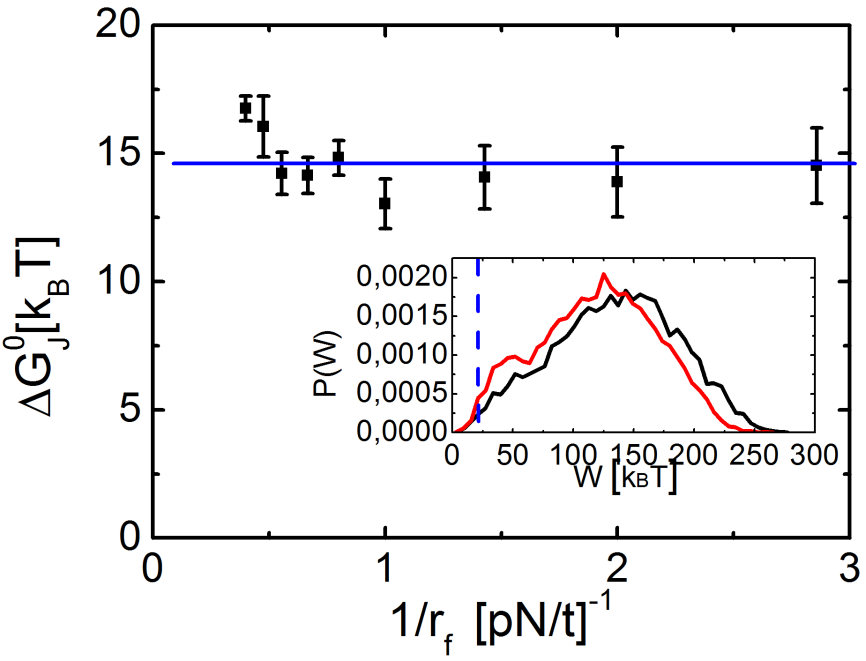


Figure 14.6: **Jarzynksi estimator as a function of the inverse of the pulling rate for the numerical simulations on the phenomenological model** : Jarzynski estimator converges successfully to ΔG^0 (blue dashed line). Inset shows two work distributions for different pulling rates.

Inset shows two work distributions, with the dissociation free energy ΔG^0 value indicated as a vertical blue dashed line. Quasistatic pulling would lead to $p(W) = \delta(W - \Delta G^0)$. As we pull out of equilibrium, $\langle W \rangle > \Delta G^0$, although for some events $W < \Delta G^0$, which allow Jarzynski estimator to converge. These events occur in the low force tail, and have key consequences in the performance of the analysis method, as we discuss in Section 14.3.

Validation of the Analysis Method

We have applied the analysis protocol on a single parameter set, in order to build a free energy profile with $\Delta G^\dagger$ and ΔG^0 similar to the biological values found in for the protein:protein complexes we studied. For the sake of consistency, we probe the analysis protocol on the physical model using four different parameter sets.

Figure 14.7 shows the profiles for each of the four chosen parameter sets. We choose them to have four different ΔG^0 values, but just two different free energy barriers $\Delta G^\dagger$. This choice guaranties that the joint obtention of both free energy magnitudes from the same data is completely independent.

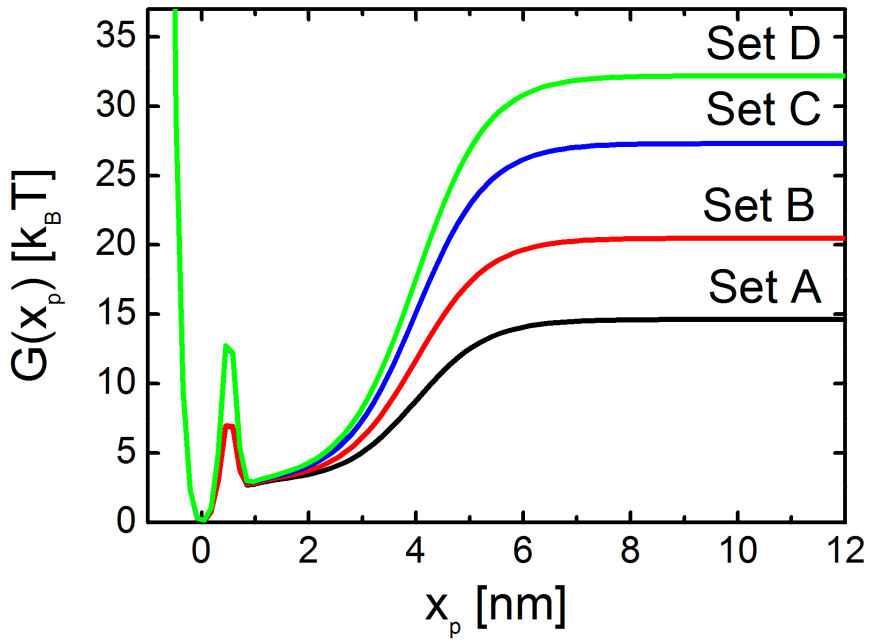


Figure 14.7: **Free energy profiles for the four data sets employed**: We choose the parameters to have four different values of ΔG^0 and only two different free energy barriers $\Delta G^\dagger$. This election allows us to prove that our analysis protocol is able to obtain both free energy magnitudes from the same data in an independent way.

We plot in Fig. 14.8 (left) the typical rupture forces versus the pulling rate for each of the four parameter sets. Clearly, the curves from sets A-B and C-D are respectively superimposed, as they have the same barrier height. The four data sets fit perfectly to Eq. (13.1). Actual barriers and obtained ones are shown in Table 14.1.

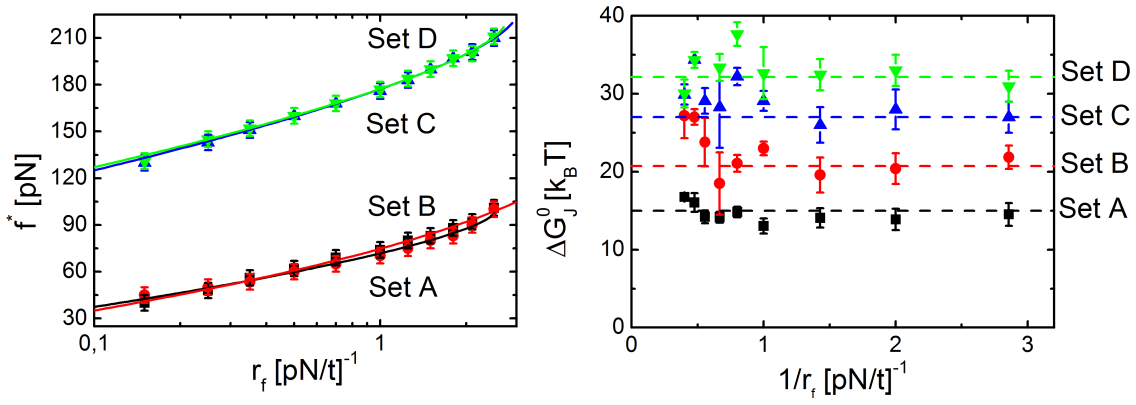


Figure 14.8: **Typical rupture force as a function of the pulling rate (left) and Jarzynski estimator for the four data sets (right)**: Force-rate curves for sets A-B and C-D are respectively superimposed, having both the same free energy barrier. Jarzynski estimator ΔG_J^0 converges successfully to the value ΔG^0 in the potentials employed (dashed lines).

Figure 14.8 (right) shows the Jarzynski estimator ΔG_J^0 as a function of the inverse of the pulling rate r_f^{-1}. Dashed lines indicate the value ΔG^0 for each of

the four parameter sets. The estimator converges in the four cases, revealing that Jarzynski equality is able to recover successfully the dissociation free energy.

Table 14.1: **Free energy magnitudes ΔG^0 and $\Delta G^\dagger$ set for each parameter set and estimation according to our analysis protocol ΔG^0_J from Jarzynski equality and fitted $\Delta G^\dagger_f$.**

Parameter Set	$\Delta G^0(k_BT)$	$\Delta G^\dagger(k_BT)$	$\Delta G^0_J(k_BT)$	$\Delta G^\dagger_f(k_BT)$
A	14.6	7.7	13.93 ± 0.5	7.3 ± 0.3
B	20.5	7.7	20.28 ± 1.0	6.7 ± 0.5
C	27.3	14.1	24.55 ± 0.3	13.2 ± 0.6
D	32.2	14.1	32.66 ± 1.56	12.5 ± 0.4

Table 14.1 gathers $\Delta G^\dagger$ and ΔG^0 as set in the four profiles, together with the estimations obtained through our analysis protocol. ΔG^0_J is the average of the last three values shown in Fig. 14.8 for each parameter set.

14.3 Discussion

In this work, developed through Chapters 15 and 16, we have shown that, by employing a suitable analysis protocol, DFS experiments can be used to obtain both the kinetic and thermodynamic properties of ligand:receptor complexes [229]. Our analysis method relies on a free energy profile which models the mechanical unbinding process. The shape of the profile is motivated by an apparent paradox we find in the analysis of our experiments, which also seems to be ubiquitous for mechanical unbinding of biological complexes (see Table 13.1). The condition $\Delta G^\dagger < \Delta G^0$ is not satisfied by a conventional molecular potential, such as a Lennard Jones or Morse potential.

We propose a free energy profile which satisfies such condition. This profile is characterized by two regions, a steep slope—accounting for the free energy barrier—and a smooth slope—leading to the unbound state. The interplay between this two slopes is of central importance for the performance of our analysis method, but is also motivated by the biological and physical characteristics of the process we are modeling, as we shall discuss in this section.

Implications of the Free Energy Profile on the Analysis Protocol

The analysis protocol we propose requires the distribution of rupture forces as only input for computing two different magnitudes. We calculate the work from the force by Eq. (13.3), so the integration of the force-extension curve is not necessary. In this sense, it might be surprising how two independent magnitudes are recovered from a single experimental output, the peak of the rupture force curve. The underlying reason is the shape of the free energy profile and the information from the force distributions in which each of the two analysis techniques rely.

The free energy profile is characterized by the scale separation between a short range steep barrier and a second smooth slope (or barrier). In the DFS experiments,

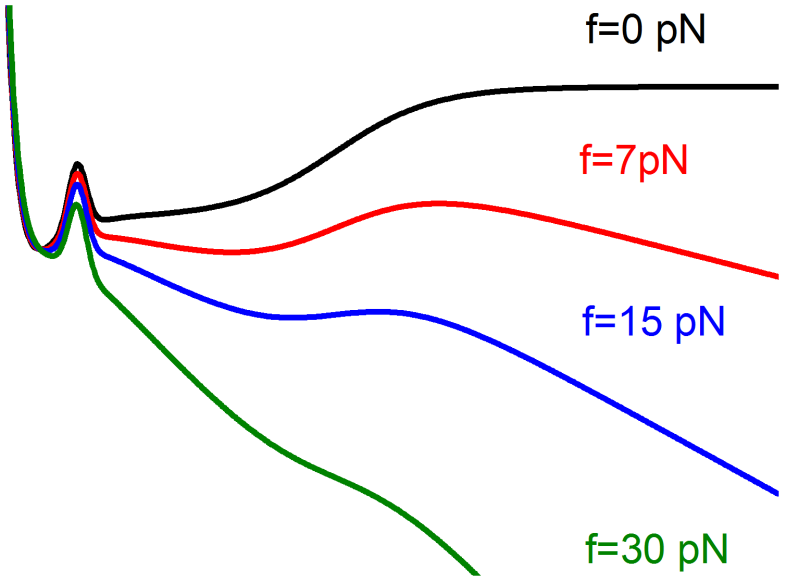


Figure 14.9: **Effect of the pulling force on the free energy profile**: The force tilts the profile, and the second slope vanishes at low forces, so that it becomes effectively a single barrier profile.

the bias exerted by the pulling force tilts the profile with a $-fx_p$ term, where f increases with time. Figure 14.9 shows the effect of an increasing force on our potential profile. The second smooth barrier vanishes at small pulling forces, while the steep one does not. In this sense, for the typical escape forces (over 30 pN), the system hops over a single barrier profile, which is completely equivalent to one where the second region does not exist. Nevertheless, with low probability, the system hops at very low forces, and thus surmounts the second barrier. In this sense, the majority of rupture events carry information just about the first barrier, while those in the low force tail have information about the second slope. This fact is shown in the force distributions in the inset of Fig. 14.5, where a second distribution seems to be superimposed in the low force region.

Our analysis method takes advantage of this uncoupling between $\Delta G^\dagger$ and ΔG^0, which can be obtained independently by applying Eq. (13.1) and Jarzysnki equality to the rupture force distributions. Equation (13.1) accounts just for the mode of the distribution, this is the average contributions. Thus, for the usual pulling rate range, we are recovering the first steep barrier.

Jarzynski equality performs an exponential average of the work distributions, enhancing events in the low force tail. Thus, those low-force escapes weight more in the calculation of the Jarzysnki estimator than average events, reason why we recover ΔG^0 successfully.

Certainly, our free energy profile can be understood as a two barrier profile, where both barriers have different length scales. In this sense, we should be able to recover ΔG^0 with Eq. (13.1) at some point. If we pulled very slowly, so that the average

rupture forces are very low ($f^* \sim 10$), we should be able to do so, as the typical events would be jumps over the two slopes. Nevertheless, this pulling ranges are extremely low, and thus experimentally unfeasible, due to the force resolution of the AFM. In this sense, our analysis protocol takes advantage also of the practicalities of DFS experiments.

Implications of the Free Energy Profile on the Physical Process of Mechanical Unbinding of Biological Complexes

The two uncoupled regions are also motivated by the physical process of mechanical unbinding, answering to different steps in the process. Figure 14.10 shows a schematic picture of the different steps in the mechanical unbinding process along the free energy profile.

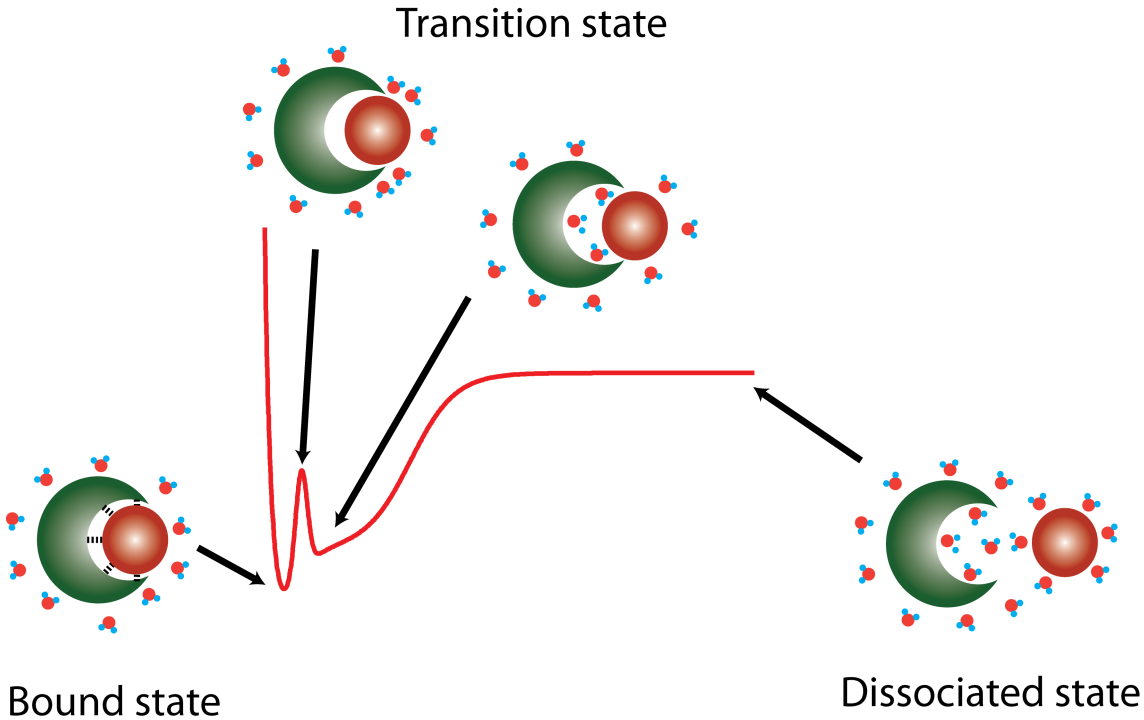


Figure 14.10: **Schematic view for the physical interpretation of the proposed free energy profile for mechanical dissociation of biological complexes.**: We can distinguish different interaction regions upon the mechanical dissociation process. First, the steep inner barrier must be overcome, involving the rupture of the short-range molecular bonds between the interacting surfaces. Then, the first water molecules access the interface region degreasing the free energy of the system. In order to dissociate completely the system, the molecules must be separated within few nanometers, solvating completely the intermolecular space and overcoming the electrostatic interactions and the dipolar moment coupling of the two proteins.

The first steep barrier reflects the local short-ranged molecular interactions between the interface residues which keep the complex in the bound state, like hydrogen bonds or salt bridges. This barrier is located at few Å from the equilibrium state, distance over which the system cannot be unbound. Over the barrier, the first water molecules enter the intermolecular region, solvating partially the interacting surfaces. This is an energetically favorable process, so the free energy profile lowers (see Fig. 14.10).

The second region accounts for the complete dissociation between both molecules, and thus spans over few nanometers. This interaction region is originated by long-range non-specific interactions between the molecules, like electrostatic, or the coupling between the dipolar moments of the molecule. This is modeled as the smooth slope which leads to the plateau where the interaction vanishes. In this sense, the first barrier has a specific origin, caused by the particular interactions between the residues in the intermolecular surfaces. The second barrier is originated by effects at a larger scale, not on the particular details of the molecules, but on their behavior as a "whole".

Mechanical Unbinding of FNR:Fd and FNR:Fld: Discussion about their Free Energy Profile

We have obtained for first time the unbinding free energy barrier and dissociation free energy for protein:protein complexes FNR:Fd and FNR:Fld from single molecule force spectroscopy measurements. The obtained magnitudes are correctly understood within the free energy profile shape we propose to govern such process. Additionally the results match previous knowledge about the complexes, providing also novel information, in particular about their kinetic properties.

The studied protein:protein complexes are of particular interest for our motivation here, as both Fd and Fld are common binding partners of flavoenzyme FNR, docking to the same binding site. FNR catalyzes the transfer of two electrons to reduce $NADP^+$ to NADPH from two independent Fd molecules, while in some algae and cyanobacteria, Fld replaces Fd under iron-deficient conditions [230, 234]. While both share similar binding affinities [230], they are known to display different interaction mechanisms [230, 231, 233, 235, 236], due mainly to the difference in size of the interacting surfaces and the residues involved in the complex stabilization.

Our analysis of the DFS experiments is able to recover the thermodynamic features of both complexes, reflected in moderate affinities which are very similar in both cases $\Delta G^0 \approx 13 k_B T$. The unbinding free energies calculated through Jarzysnki equality agree respectively within error bars with the calorimetric binding free energy reported in [230]. Remarkably, this behavior is similar in both complexes, despite we find significant differences in their kinetic behavior. The rupture forces are different in both cases, so the free energy barrier heights contrast considerably ($\Delta G^\dagger \approx 7 k_B T$ for FNR:Fd and $\Delta G^\dagger \approx 5 k_B T$ for FNR:Fld). These differences are attributed to the particular features concerning the formation of each complex. Showing a larger interacting interface, Fd binding to FNR is more specific, with salt bridges between certain key positive residues on the FNR surface and acidic residues on Fd [233, 235–237], which would contribute in addition to other non-specific interactions such as hydrogen bonds or the hydrophobic effect.

On the contrary, kinetic analysis of side-directed mutants and docking studies on FNR:Fld suggest that Fld can adopt multiple orientations on the FNR surface, and that charged residues are not involved in crucial specific interactions [231, 233, 236]. In this sense, these features agree with our findings. Upon complex rupture, specific short range-interactions contribute decisively to the FNR:Fd interaction, reflected in a higher free energy barrier when compared to FNR:Fld, whose binding mechanism is mainly due to the hydrophobic effect. Thus, larger forces are needed to unbind the former complex. Nevertheless, non-specific interactions, which contribute mainly to

the second barrier in the profile, are very similar in both cases, given the overall shape of the molecules, so the ΔG^0 values are virtually alike for both complexes.

Part IV

Conclusions and Future Work

We chase misprinted lies
We face the path of time
And yet I fight
And yet I fight
This battle all alone
No one to cry to
No place to call home
My gift of self is raped
My privacy is raked
And yet I find
And yet I find
Repeating in my head
If I can't be my own
I'd feel better dead

ALICE IN CHAINS, *Nutshell*

Chapter 15

Concluding Remarks

Molecular simulations provide a powerful way of understanding biological systems. They provide high resolution information, which allows to capture in a detailed way both the thermodynamic and kinetic properties of the system of interest. This allows to study problems of high relevance, such as protein folding, allosteric regulation or even drug discovery. Even more, it is possible to make a direct connection to experimental data, which helps in interpreting the results or in suggesting new experimental routes.

Nevertheless, in order to fully realize their potential, we must two meet two requirements. First, we need a meaningful model of our biological system, which is serves as the input of our simulation. This model should be able to make a direct connection with the real system, in order to render accurate results and allow to make predictions. Second, we require of methods that can provide knowledge and make a quantitative connection between the output of our simulations and the experimental data.

The present Thesis belongs to this field, dealing with three different problems where modeling biomolecular processes and employing suitable analysis methods are the overarching elements. In particular, the work we presented, has in free energy calculations a common guiding thread, providing it as a valuable tool for understanding the static and dynamic properties of biological molecules.

In the next lines, we summarize the most relevant achievements of the Thesis. More detailed discussions on the results are found in the pertinent Chapters.

Part I

In this part we have analyzed the unfolding mechanism of a model protein under the presence of an external force. This computational study is inspired in force clamp single molecule experiments, where a molecule is probed by being subject to a mechanical force. We apply two complementary analysis methods in order to understand its configurational space and to unveil the unfolding pathway(s). First we describe the system by low-dimensional representations of the free energy landscape along relevant order parameters. Second, we describe the system with a Markov state model. This study allows us to arrive to some conclusions.

- We calculate the one dimensional profiles along two suitable reaction coordinates, the end-to-end distance ξ and the fraction of native contacts Q. The first coordinates is motivated by the topology of the system. As we apply a

mechanical force along the pulling direction, this degree of freedom becomes naturally the slowest one, and the reaction coordinate of the system. Indeed, this is the coordinate employed experimentally to describe molecules under the presence of force. The fraction of native contacts has been proved to work successfully to characterize proteins, even when no native-centric assumption is taken. These two profiles agree in their overall description of the system. We identify four major metastable states, the native, the stretched, and two intermediates, an aligned and a half-stretched conformation. These two states are energetically favorable, as the protein aligns its ends in the direction of the pulling force. The half-stretched conformation is the most stable state under the present conditions. Even more, it seems to be the mechanical intermediate the system uses to transition to the fully stretched (unfolded) conformation.

- We apply the Bayesian test to evaluate the quality of ξ and Q as reaction coordinates. Surprisingly, ξ results in a poor choice, given the multi-peaked structure of the Transition-Path Ensemble, which suggest that the projection is non-Markovian. Q reveals itself as a better reaction coordinate, locating the transition state at a very low value of Q. This prevents us to meet any relevant conclusion about the actual conformational space, nor about the unfolding mechanism.

- We use PCA as a tool for finding relevant order parameters. The first PC describes the system in a quite similar way to ξ and Q, meaning that the largest amplitude motions are related with transitions among the native, half-stretched and stretched states. The second PC provides a more detailed vision of the system, with several transitions between states separated by low energy barriers. This suggests that the actual conformational space of this system could be more complex, being several different states projected onto similar Q and ξ values.

- We build a Markov state model of the system. After applying an appropriate analysis protocol, we describe the equilibrium ensemble of the system as a network made up of 13 different macrostates. In a coarse way, the structure agrees with previous findings, as the network is divided into three different regions, a native, stretched and half-stretched. Nevertheless, it provides a more accurate vision of the system, characterized by the different points:

  1. There are two time scales involved. One is related with transitions between the native and half-stretched states within times of $\sim 100\,\text{ns}$. The second one implies the unfolding transition, which occurs within longer time scales of $\sim 10\,\mu\text{s}$.
  2. The native state is actually made of two different states which are separated by a large barrier and have a very different kinetic role. One is related to fast transitions to the half-stretched state, while the other to transitions to different intermediates, promoting the unfolding transition.
  3. There are two proper intermediate states which connect the native ensemble with the stretched ensemble. Physically they are related with a disruption of the hydrophobic core of the protein.

- Applying Transition Path Theory allows to reveal the unfolding pathways of the system. The unfolding transition does not occur through a single well defined route, but rather through an ensemble of Transition Pathways. Coarsely, there are two main routes, driven by the two intermediates previously mentioned.

- Surprisingly, the half-stretched configuration plays little role in the unfolding mechanism, as just a little fraction of the unfolding flux passes through it.

In few words, this part serves us to probe some of the current state-of-art methods for understanding molecular simulations. Low dimensional representations provide an overall correct description, where the main strength is its simplicity and the intuitive picture they provide. Markov state models are more sophisticated analysis methods, able to represent multidimensional free energy landscapes in an understandable way. They allow to identify easily the stability of each state, and also the transition rates between free energy basins. Also, they allow a straightforward calculation of the pathways connecting two subsets of states, and thus of the (un)folding mechanism.

Given the multiplicity of unfolding pathways and the overlapping of conformational states, one-dimensional descriptions failed to explain correctly our system. In this sense, the Markov state model description appear as the most meaningful source of information. This is a surprising feature given that we chose a remarkably simple system, not just given the protein model, but also due to the presence of the pulling force, which set a natural reaction coordinate for the system to evolve through. Election of more sophisticated reaction coordinates, probably by some optimization mechanism, would surely yield a better free energy profile, where the states of the system would be better represented. Nevertheless, the multiplicity of unfolding pathways seem an incompatible ingredient with any one-dimensional representation.

The one-dimensional descriptions are not able to explain correctly our system, given the multiplicity of unfolding pathways and the overlapping of confomational states. Hence, Markov state models present a more meaningful vision, which captures the complexity of the unfolding mechanism. Given the presence of an external force, which simplifies the dynamics of the system, this is a surprising fact. Election of more sophisticated reaction coordinates, probably by some optimization mechanism, would probably yield better free energy profiles, where the states of the system would be better represented. Nevertheless, the multiplicity of unfolding pathways seem an incompatible ingredient with any one-dimensional representation.

Part II

In this part we have studied Peyrard-Bishop-Daouxois DNA model in three different applications. We can enumerate the main achievements in the following way.

1. We modify the PBD model in order to incorporate a barrier in the on-site potential which accounts for the solvent effects. This barrier has allowed to improve the performance of the model in two different ways. First, the melting transition is closer to the experimentally observed one. We achieve a sharper denaturation, as the barrier increases the cooperativity of the model. Second,

the dynamics of bubble formation becomes also more realistic. Bubbles last longer and have a more appropriate size, in agreement with the experimental observations. Additionally, we use PCA to characterize the melting transition in a novel way. The dependence of the PCA eigenvalues with temperature allows to identify the transition a mode which drops to zero.

2. We modify the PBD model to include a diffusing particle -or generic protein- which interacts with DNA bubbles. This model is inspired in the protein-DNA interaction mechanism, by which some proteins couple to physical properties of the DNA molecule. We probe the model on three different promoter sequences, which were previously studied with PBD model. Two of the sequences are strong promoters and the other a weak promoter, in terms of RNA expression. The performance of the model is satisfactory, as the particle couples to the bubble dynamics, which achieves longer bubbles in better agreement with biological time scales. For these three promoter sequences, the most relevant states correlate with openings at significant biological sites, such as the transcription starting site of binding sites of different transcription factors. Additionally, we are able to find relevant differences between the structure of the free energy landscape of the strong and weak promoters. Strong promoters show few states which attract the majority of the dynamics, characterized as deep free energy basins. The weak promoter has a more distributed structure, with several states of intermediate population.

3. We apply the model and the analysis method we proposed previously to nine promoter sequences from a particular organism, cyanobacterium *Anabaena* PCC7120. The relevance of this study is that such promoters were never studied before in the context of PBD model. We focus on the identification of the transcription starting sites. Additionally, some of the these promoters display more than a single transcription starting site. In this sense, our method can be used to give a quantitative comparison between them, and likely correlate it with the biological knowledge about the system. Our analysis gives meaningful information about the nine promoters, identifying in every case the transcription starting sites as a prominent state in the dynamics. The relative importance of these sites in promoters with several transcription starting sites correlates with their biological performance. The model identifies further sites as likely binding sites for transcription factors which in some cases agree with typical binding sites in prokaryotes, such as the -10 or -35 site.

This work exploits one of the most controversial issues regarding the PBD model, the relation between bubble formation and protein binding sites in promoter sequences. Our study moves forward as it includes the active ingredient in this process, which is the diffusing protein. In the first work we choose promoters which were already analyzed in a similar context, so being able to find relevant binding site is not quite an achievement. However, thanks to the improvement of the model and the analysis method we are able to characterize these sites and also the overall free energy landscape of the system. In the case of the cyanobacterial promoters, we apply a physical model for the first time, with successful results. Taking this model and method as a way to analyze promoter sequences, we can conclude that for the studied promoters, the TSS is identified as a region which opens with more

probability, and were the particle has a larger probability to bind. The occurrence of this state seems to be also related to the biological "strength" of the promoter. Obviously we do not mean in any case that bubble formation drives protein binding, but our studies show a certain correlation between both processes.

Part III

In this part we present a collaboration with experimentalists, in which we analyze DFS-AFM experiments for mechanical unbinding of two protein:protein complexes. Our purpose is to recover the free energy barrier and the dissociation free energy. Additionally, we propose a model for such process, which allows us to perform numerical simulations to reproduce experimental data and apply the same analysis procedure. The main achievements in this work are the following:

- We have proposed a systematic and robust method for extracting independently the free energy barrier and the dissociation free energy from DFS experiments. We extract meaningful values for both magnitudes where, remarkably, the dissociation free energy matches the calorimetric value in both complexes.

- We find a discrepancy that free energy barriers are lower than dissociation free energies. This is a problematic feature which cannot be understood with conventional molecular potentials. This problem is a recurrent one for unbinding of biological complexes, as we find several ones with the similar discrepancy.

- We propose a new free energy profile that accomplishes such discrepancy. The free energy profile is characterized by two decoupled regions, a first steep one and a second smooth one.

- Simulations on a phenomenological model based on this free energy profile are able to reproduce the experimental data. Application of the same analysis procedure on the simulations recover successfully the free energy magnitudes.

The central idea of this work is that the mechanical unbinding process of biological complexes is governed by a free energy profile characterized by two regions with separated scales. The first one is very brittle and rules over the average rupture trajectory. The second is smooth and dominates on low force escapes, explaining the success of the analysis method we present. From a phenomenological point of view, this first region is associated with short range specific interactions, while the latter answers for nonspecific interaction between both molecules.

Future Perspectives

The work developed through the Thesis leaves naturally many open topics which suggest new lines for future research. We can enumerate these perspectives also related to each of the three topics.

Part I

The methodology presented here is of wide application, not exclusively to the field of molecular simulation. A natural extension is the study of different systems with similar techniques. A possible topic is the analysis of protein folding for different kinds of systems using all-atom simulations. In particular, an interesting issue regards the comparison between the topology of Markov networks for those proteins which fold in a two-state manner and downhill folding proteins, which fold in the absence of barrier. The differences in their one-dimensional profile are well known, but the global structure of their free energy landscape remains unexplored.

Regarding the particular protein model we explored, we just studied the mechanical unfolding mechanism. We find here a complex behavior, where the apparent mechanical intermediate played a had little role in the unfolding mechanism. An interesting issue is the connection between the mechanical and the thermal unfolding. Single-molecule techniques manipulate molecules by modifying its actual free energy landscape. In this sense, the behavior under force could have little relevance regarding the *in vivo* one. A direct project would be the simulation of the thermal denaturation for the same protein model and an analogous analysis, yielding a comparison between both landscapes. This would allow us to seek for common features, such as the presence of the half-stretched conformation, and its role in the unfolding mechanism.

Part II

PBD model has been extensively studied by different researchers and for different purposes. Our work here presents a valuable tool for analyzing promoter sequences which relies, not on its performance, but rather on its physical insight and the ability to quantify possible binding sites. Naturally, an open issue is the application of the model and analysis method to further promoter sequences, particularly to those well characterize from a biological point of view in order to validate the possible predictions they might render.

The protein-DNA model we proposed consideres a generic protein which interacts with bubbles in the DNA sequence. This model could be as an inspiration for further DNA-interacting proteins which operate at a similar scales. One of the best candidates are helicases, which use the energy of ATP to open a bubble along the DNA molecule. The main ingredient that should be added to our model is the inclusion of the asymmetry in displacement of the protein. This should be related to some energy consumption, transforming the model into a mixed molecular motor one. Sequence effects on the velocity of the motor and its efficiency would be easy to test with this proposal. Currently we are developing this project, aiming to achieve concluding results in the short term.

Part III

We have presented an analysis method and model which should be general for unbinding of biological complexes. In this sense, testing it with any similar system appears is a direct application. This could not jut validate the method and the relevance of the free energy profile, but also to check if this discrepancy is ubiquitous and holds the physical origin we claim.

Another way to test to our proposals is to reconstruct the unfolding profile by means of molecular dynamics simulations. The reconstruction of such profiles is currently a rather straightforward task, given the available enhanced sampling techniques and their efficiency in their implementation in many software packages. An atomic detailed description of this process could provide us great insight about the actual unbinding mechanism, finding the actual origin of these two decoupled regions, if they exist. Additionally, this constitutes an interesting connection between the molecular simulations and the experimental data.

Chapter 16

Conclusiones y Perspectivas

Las simulaciones moleculares proporcionan un método eficaz para comprender sistemas biológicos. La alta resolución que proporcionan, permite obtener de manera muy detallada tanto las características termodinámicas como las cinéticas del sistema de interés. Esto permite estudiar problemas de gran relevancia, tales como la descripción de los mecanismos de plegamiento en proteínas, regulación alostérica,o incluso contribuir al diseño de fármacos. Asimismo, es posible una relación directa de los resultados computacionales con los datos experimentales, lo cual contribuye tanto a la interpretación de resultados como a sugerir nuevos experimentos.

No obstante, para ser capaces de aprovecharnos de todo su potencial, debemos cumplir dos requisitos. Primero, es necesario un modelo adecuado de nuestro sistema biológico, ya que éste será la base de nuestra simulación. Este modelo debe ser capaz de conectar de manera directa con el sistema real, tanto para poder proporcionar resultados precisos, como para permitir la elaboración de predicciones. Por otra parte, es necesario disponer de métodos de análisis que puedan proporcionar un conocimiento directo, así como realizar una conexión cuantitativa entre los resultados computacionales y los experimentales.

La presente Tesis Doctoral se ubica en esta problemática, tratando tres cuestiones diferentes donde el modelado de procesos biomoleculares y el empleo de técnicas adecuadas de análisis son los elementos sustentantes. En particular, el trabajo que hemos presentado tiene en el cálculo de energías libres el hilo conductor, al ser ésta una herramienta de gran valor cuando se pretende comprender las propiedades estáticas y dinámicas de las moléculas biológicas.

En los próximos párrafos, resumimos los logros más relevantes de esta Tesis Doctoral. Discusiones más detalladas de los resultados pueden encontrarse en los capítulos correspondientes.

Parte I

En esta parte hemos analizado el mecanismo de desplegamiento de una proteína modelo bajo la presencia de una fuerza externa. Este estudio computacional se inspira en los experimentos de molécula individual a fuerza constante, donde una molécula se somete a una fuerza mecánica. Hemos aplicado dos métodos de análisis complementarios para comprender el espacio conformacional del sistema y revelar los caminos de desplegado seguidos. Comenzamos describiendo el sistema con representaciones de baja dimensión de su paisaje de energía libre a lo largo de parámetros de orden relevantes. A continuación, construimos el modelo de Markov del sistema,

comparando ambas descripciones. Este estudio nos permite llegar a las siguientes conclusiones:

- Calculamos los perfiles unidimensionales a lo largo de dos coordenadas de reacción apropiadas, la distancia entre extremos ξ y la fracción de contactos nativos Q. El empleo de la primera de estas coordenadas está motivado por la topología del sistema. La fuerza mecánica convierte ξ en la coordenada de reacción natural del sistema. De hecho esta es la coordenada empleada experimentalmente para describir moléculas bajo la acción de una fuerza. La fracción de contactos nativos ha sido empleada de manera en numerosas ocasiones para caracterizar el plegado de proteínas, incluso cuando los contactos nativos no se asumen previamente en el modelo. Estos dos perfiles están de acuerdo en su descripción global del sistema. Identificamos cuatro estados metaestables principales, el nativo, el estirado, y dos intermediarios, el alineado y el medio-estirado. Estas dos últimas conformaciones son energéticamente favorables ya que la proteína alinea sus extremos en la dirección de la fuerza. En particular, la conformación medio-estirada es el estado más estable bajo estas condiciones. Así mismo, parece ser el intermediario mecánico del sistema, apareciendo como una configuración a medio camino entre la nativa y la totalmente estirada.

- Aplicamos el test Bayesiano para evaluar la calidad de ξ y Q como coordenadas de reacción. Sorprendentemente, ξ resulta una elección mediocre, dada la estructura de múltiples picos en el histograma de los caminos de transición, lo cual sugiere que la proyección es no-Markoviana. Q aparece como una coordenada de reacción de mayor valor, al localizar el estado de transición en un valor muy bajo de Q. No obstante no es posible determinar el camino de desnaturalización con esta descripción simple.

- Empleamos PCA como herramienta para encontrar parámetros de orden relevantes. La primera componente principal describe el sistema de manera similar a ξ y Q, lo que implica que los modos de mayor amplitud se relacionan con transiciones entre las configuraciones nativa, medio-estirada y estirada. La segunda componente principal proporciona una versión más detallada del sistema, con numerosas transiciones separadas por barreras de energía libre bajas. Esto sugiere que el verdadero espacio conformacional del sistema podría ser más complejo, al estar más de un estado proyectado en valores similares de Q y ξ.

- Construimos el modelo de Markov del sistema, que representa el conjunto de equilibrio de nuestro sistema como una red con 13 macroestados diferentes. A *grosso modo*, la estructura es la misma que la encontrada anteriormente, ya que la red está dividida en tres regiones diferentes, la nativa, la estirada y la medio estirada. No obstante, esta descripción permite una visión más precisa del sistema, caracterizada por los siguientes puntos:

    1. Existen dos escalas temporales diferentes involucradas en nuestro sistema. La primera está asociada a transiciones entre los estados nativo y medio-estirado con un tiempo característico de $\sim 100\,\mathrm{ns}$. La segunda se asocia a la transición de desnaturalización, con un tiempo característico mayor de $\sim 10\,\mu\mathrm{s}$.

2. El estado nativo está compuesto por dos estados distintos separados por una barrera considerable, lo cual les otorga un papel cinético muy diferente. El primero está relacionado con las transiciones rápidas al estado medio-estirado, mientras que el otro con transiciones entre diversos intermediarios, que conducen al estado desnaturalizado.
3. Encontramos dos estados intermediarios reales, que conectan las configuraciones nativas con la estirada. Físicamente, ambos implican la ruptura del núcleo hidrófobo de la proteína.

- Aplicando *Transition Path Theory* somos capaces de calcular los caminos de desnaturalización que sigue el sistema. Esta transición no ocurre a través de una única ruta bien definida, sino más bien a través de un conjunto de caminos de transición. Existen dos rutas principales, impulsadas por cada uno de los dos intermediarios mencionados anteriormente.

- La configuración medio-estirada tiene un papel despreciable en el mecanismo de desnaturalización, al participar en tan sólo una pequeña fracción del flujo de desnaturalización.

En esta parte hemos aplicado algunos de los métodos de análisis existentes para comprender simulaciones moleculares. Las representaciones de baja dimensión son capaces de proporcionar descripciones globalmente correctas, donde su principal virtud es tanto la simplicidad como su descripción intuitiva. Los modelos de Markov son métodos de análisis más sofisticados, capaces de representar el paisaje de energía libre de una manera fácilmente comprensible. Permiten la identificación sencilla de la estabilidad de los estados, así como de las tasas de transición entre mínimos de energía libre. Asimismo, permiten un cálculo directo de los caminos que conectan dos subconjuntos de estados de la red, y por tanto de los mecanismos de plegado o desplegado.

Las descripciones unidimensionales no son capaces de explicar de manera adecuada nuestro sistema, dada la multiplicidad de estados de desplegamiento y el solapamiento de estados conformacionales. Así, los modelos de Markov presentan una visión más relevante. Dada la presencia de la fuerza externa, que simplifica la dinámica del sistema, este hecho es sorprendente. La elección de coordenadas de reacción más sofisticadas, probablemente mediante algún algoritmo de optimización, proporcionarían probablemente perfiles de energía libre más adecuados, donde los estados que visita el sistema estarían adecuadamente representados. No obstante, la multiplicidad de caminos de desplegado parece ser un ingrediente incompatible con cualquier representación unidimensional.

Parte II

En esta parte, hemos estudiado el modelo de Peyrard-Bishop-Dauxois para la molécula de DNA mediante tres aplicaciones diferentes. Podemos enumerar los principales logros de la siguiente manera:

1. Modificamos el modelo de PBD para incorporar una barrera que tuviese en cuenta los efectos del solvente. Esta barrera mejora los resultados del modelo original de dos maneras diferentes. Por una parte, la transición de desnaturalización es más similar a la experimental. Reproducimos una curva de

desnaturalización más abrupta, ya que la barrera incrementa el grado de cooperatividad del modelo. Por otra parte, la dinámica de la formación de burbujas es más realista. Las burbujas duran más y son más grandes, de acuerdo con las observaciones experimentales. Además, empleamos PCA para caracterizar la transición de desnaturalización de manera novedosa. La dependencia de los autovalores de PCA con la temperatura nos permiten identificar esta transición como un modo que tiende a cero.

2. Proponemos una modificación del modelo original de PBD para introducir una partícula o proteína genérica que interacciona con las burbujas del DNA. Este modelo está inspirado en los mecanismos de interacción proteína-DNA, por los cuales ciertas proteínas se acoplan a las propiedades físicas de la molécula. Aplicamos este modelo a tres promotores diferentes, ya estudiados en otros trabajos con el modelo de PBD. Dos de ellos son promotores fuertes y el otro débil, en términos de expresión de RNA. El modelo se comporta de manera satisfactoria, ya que la dinámica de la partícula se acopla a las de los pares de bases, consiguiendo burbujas más duraderas, en consonancia con las escalas biológicas. En los tres promotores analizados, los sitios más relevantes están correlacionados con regiones de relevancia biológica, tales como el sitio de inicio de la transcripción. Asimismo, encontramos diferencias relevantes entre la estructura del paisaje de energía libre de los promotores fuertes y débiles. Los promotores fuertes tienen unos pocos estados que atraen la mayor parte de la dinámica, caracterizados como cuencas de energía libre profundas. Por otro lado, el promotor débil muestra una estructura más distribuida, con varios estados de población intermedia.

3. Aplicamos el modelo y método de análisis anterior a nueve secuencias de promotores de un organismo en particular, la cianobacteria *Anabaena* PCC7120. La relevancia de este estudio es que dichos promotores no han sido analizados previamente con el modelo de PBD. Nos centramos en la identificación del sitio de inicio de la transcripción. Adicionalmente, algunos de estos nueve promotores tienen más de un sitio de inicio de la transcripción. De esta manera, nuestro método puede usarse para proporcionar una comparación cuantitativa entre ellos, correlacionándola con el conocimiento biológico de que disponemos. Nuestro análisis nos aporta información de valor sobre los nueve promotores, ya que los sitios de inicio de la transcripción son identificados en todos ellos como estados de relevancia en la dinámica del modelo. La importancia relativa entre dichos promotores con varios sitios de inicio de la transcripción muestra una correlación con su papel biológico. El modelo identifica además más sitios probables de unión para factores de transcripción, que en algunos casos aparecen en regiones típicas para organismos procariotas, como la -10 o -35.

Este trabajo explota una de los aspectos más controvertidos de modelo de PBD, la relación entre la formación de burbujas y la presencia de sitios de unión en secuencias de promotores. Nuestro estudio supone un paso más en este punto, al incluir el ingrediente activo de este proceso, la proteína que se difunde a lo largo de la cadena de DNA. En el primer trabajo escogemos promotores ya analizados en un contexto similar, de manera que la identificación de sitios de unión no supone ningún logro

particular. Gracias al método de análisis, somos capaces de caracterizar de una manera global el paisaje de energía libre del sistema. En el caso de los promotores de cianobacteria, aplicamos el modelo por primera vez en ellos, siendo los resultados satisfactorios. Si interpretamos el modelo y método como un procedimento para analizar secuencias de promotores, hemos relacionado el sitio de inicio de la transcripción con regiones que se abren con mayor facilidad. Asimismo, la probabilidad de estos estados se relaciona con la "fuerza" biológica de los promotores. Obviamente, en ningún momento concluimos que la formación de burbujas sea la causa de la unión de proteínas a la molécula de DNA, pero nuestro estudio muestra una cierta correlación entre ambos fenómenos.

Parte III

En esta parte presentamos una colaboración experimental, en la cual analizamos experimentos de DFS-AFM para la disociación mecánica de dos complejos diferentes proteína:proteína. El objetivo es la recuperación de la barrera de energía libre y la energía libre de disociación. Asimismo, proponemos un modelo para dicho proceso, el cual nos permite realizar simulaciones numéricas para reproducir los datos experimentales, sobre los cuales podemos aplicar el mismo método de análisis. Los principales resultados conseguidos son:

- Proponemos un método sistemático y robusto para extraer de manera independiente la barrera de energía libre, así como la energía libre de disociación a partir de experimentos de DFS. Extraemos valores relevantes para ambas magnitudes, de acuerdo conn el valor de calorimetría de la energía libre de disociación conocido para ambos complejos.

- Encontramos que las barreras de energía libre son menores que sus respectivas energías libres de disociación. Este problema aparece de manera recurrente en la disociación de complejos biológicos, de manera que los perfiles de energía libre convencionales no son adecuados para representarlos.

- Proponemos un nuevo perfil de energía libre que tenga en cuenta dicha discrepancia. Este perfil está caracterizado por dos regiones desacopladas, una primera abrupta y una segunda suave.

- Realizamos simulaciones sobre un modelo fenomenológico basado en dicho perfil permiten reproducir los datos experimentales. La aplicación del mismo procedimiento de análisis en las simulaciones recuperan de manera exitosa ambas magnitudes de energía libre.

La idea central de este trabajo es que el perfil de energía libre que gobierna la disociación mecánica de complejos biológicos está caracterizada por dos regiones a diferente escala. La primera es abrupta y determina el comportamiento promedio. La segunda es suave y domina sobre los escapes a baja fuerza, explicando la aplicabilidad del protocolo de análisis empleado. Desde un punto de vista fenomenológico, esta primera región está asociada con interacciones específicas de corto alcance, mientras que la segunda con interacciones no específicas entre ambas moléculas.

Perspectivas futuras

El trabajo realizado a lo largo de esta Tesis Doctoral plantea numerosas preguntas que pueden motivar trabajos futuros. Enumeramos algunas de estas perspectivas, relacionadas con cada uno de los tres bloques tratados:

Parte I

La metodología aquí presentada es de amplia aplicación, no sólo en el campo de la simulación molecular. Una continuación natural de este planteamiento es la aplicación al estudio de distintos sistemas con técnicas similares. Un planteamiento posible es el análisis del plegamiento de distintos tipos de proteínas empleando modelos a todos los átomos. En concreto, sería interesante comparar la topología de las redes de Markov para proteínas que pliegan en dos estados con aquellas que pliegan sin barrera (*downhill folders*). Las diferencias en su perfil de energía libre unidimensionales son bien conocidas, si bien no tanto la estructura de su paisaje de energía libre.

Respecto al modelo de proteína estudiado, hemos explorado solamente su mecanismo de desplegado. Un problema interesante es la conexión entre la desnaturalización mecánica y la térmica. Las técnicas de manipulación de moléculas individuale modifican su paisaje de energía libre. De esta manera, el comportamiento bajo fuerza podría tener poco interés en relación con su comportamiento *in vivo*. Un proyecto inmediato consiste en simular la desnaturalización térmica para la misma proteína modelo y realizar un análisis análogo, que permita una comparación entre ambos paisajes. Esto posibilitaría buscar características comunes, como la prevalencia de la configuración medio-estirada, y su posible papel en el mecanismo de desplegado.

Parte II

El modelo de PBD ha sido estudiado ampliamente por numerosos investigadores con propósitos diversos. Nuestro trabajo muestra una herramienta interesante para el análisis de secuencias de promotor. Su valor no radica en su eficiencia, sino en su modelado del proceso físico y en su habilidad para cuantificar posibles sitios de unión. Naturalmente, una continuación directa es su aplicación a otras secuencias de promotores, particularmente si están bien caracterizadas desde un punto de vista biológico.

El modelo de proteína-DNA propuesto considera una proteína general que interacciona con las burbujas formadas en el DNA. Este planteamiento podría inspirar modelos de interacción para otro tipo de proteínas que operen a escala similar. Una de las principales candidatas son las helicasas, que emplean la energía del ATP para abrir una burbuja y desplazarla a lo largo de la molécula de DNA. El principal ingrediente a incluir en este modelo es la asimetría en el desplazamiento de la proteína. Ésta debería estar relacionada con el consumo de energía, transformando el modelo en uno de motores moleculares a nivel mesoscópico. Los efectos de la secuencia en la velocidad del motor, así como su eficiencia serían sencillos de comprobar con esta propuesta. Actualmente estamos desarrollando este proyecto con objeto de llegar a resultados concluyentes a corto plazo.

Parte III

Hemos presentado un modelo y un método de análisis general para la disociación de complejos biológicos. De esta manera, su aplicación a este tipo de sistemas es de utilidad directa. Esto serviría no sólo para validar el método y la relevancia de este perfil de energía libre, sino también para comprobar si la discrepancia encontrada es general y está fundamentada en el origen físico que argumentamos.

Otra manera para comprobar la forma de este perfil de energía libre es mediante simulaciones de dinámica molecular. La reconstrucción de este tipo de perfiles es hoy en día relativamente directa, dada la disponibilidad de técnicas de *enhanced sampling* así como la eficiencia con la que están implementadas en la mayoría de los paquetes de simulación. Una descripción con resolución atómica de este proceso nos proporcionaría un conocimiento detallado del mecanismo de disociación, permitiéndonos explorar el origen de estas dos regiones desacopladas. Asimismo, supondría una conexión interesante entre las simulaciones de dinámica molecular y los resultados experimentales.

Bibliography

[1] Erwin Schrodinger. *What is Life?* Cambridge University Press, 2012.

[2] David L. Nelson, Albert K. Lehninger, and Michael M. Cox. *Lehninger Principles of Biochemistry*. New York: W. H. Freeman, 2008.

[3] L. Pauling, R. B. Corey, and H. R. Branson. "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain". In: *Proc. Natl. Acad. Sci. USA* 37.4 (1951), pp. 205–211.

[4] C. B. Anfinsen. "Principles that govern the folding of protein chains". In: *Science* 181.4096 (1973), pp. 223–230.

[5] Cyrus Levinthal. "Are there pathways for protein folding?" In: *Journal de Chimie Physique et de Physico-Chimie Biologique* 65 (1968), pp. 44–45.

[6] David J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

[7] Robert Zwanzig, Attila Szabo, and Biman Bagchi. "Levinthal's paradox". In: *Proceedings of the National Academy of Sciences* 89 (1992), pp. 20–22.

[8] K. Dill and H. S. Chan. "From Levinthal to pathways to funnels". In: *Nat. Struct. Biol.* 4.1 (1997), pp. 10–19.

[9] Hugh Nymeyer, Angel E. García, and José Nelson Onuchic. "Folding funnels and frustration in off-lattice minimalist protein landscapes". In: *Proceedings of the National Academy of Sciences* 95.11 (1998), pp. 5921–5928.

[10] F. Noé et al. "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations". In: *Proc. Natl. Acad. Sci. USA* 106.45 (2009), pp. 19011–19016.

[11] Peter G. Wolynes. "Recent successes of the energy landscape theory of protein folding and function". In: *Q Rev Biophys* 38.4 (2005), pp. 405–410.

[12] G. R. Bowman, V. S. Pande, and F. Noé. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer Netherlands, 2014.

[13] Kresten Lindorff-Larsen et al. "How fast-folding proteins fold". In: *Science* 334 (2011), pp. 517–520.

[14] D. van der Spoel et al. "GROMACS: fast, flexible, and free". In: *J. Comput. Chem.* 26.16 (2005), pp. 1701–1718.

[15] B. R. Brooks et al. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". In: *J. Compu. Chem.* 4.2 (1983), pp. 187–217.

[16] J. C. Phillips et al. “Scalable molecular dynamics with NAMD”. In: *J. Comput. Chem.* 26 (2005), pp. 1781–1802.

[17] K. A. Beauchamp et al. “Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements”. In: *J. Chem. Theory Comput.* 8.4 (2012), pp. 1409–1414.

[18] C. Bustamante et al. “Mechanical processes in Biochemistry”. In: *Annu. Rev. Biochem.* 73 (2004), pp. 705–748.

[19] S. Kumar et al. “THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method”. In: *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021.

[20] J. Hsin et al. “Molecular origin of the hierarchical elasticity of titin: Simulation, experiment, and theory”. In: *Annu. Rev. Biophys.* 40 (2011), pp. 187–203.

[21] Matthias Rief et al. “Reversible Unfolding of Individual Titin Immunoglobulin Domains by AFM”. In: *Science* 276 (1997), pp. 1109–1111.

[22] D. Frenkel and B. Smit. *Understanding Molecular Simulation.* Academic Press, 2001.

[23] Alessandro Laio and Michelle Parrinello. “Escaping free-energy minima”. In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566.

[24] Peter G. Bolhuis et al. “TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark”. In: *Annu. Rev. Phys. Chem.* 53 (2002), pp. 291–318.

[25] P. G: Wolynes, J. N. Onuchic, and D. Thirumalai. “Navigating the folding routes”. In: *Science* 267.5204 (1995), pp. 1619–1620.

[26] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. “Theory of protein folding: the energy landscape perspective”. In: *Annu. Rev. Phys. Chem.* 48 (1997), pp. 545–600.

[27] M. Buchanan. “When the going gets tough”. In: *Nature Phys.* 6.4 (2010), p. 235.

[28] D. M. F. van Aalten et al. “A comparison of techniques for calculating protein essential dynamics”. In: *J. Comput. Chem..* 18.2 (1997), pp. 169–181.

[29] A. Amadei, A. B. Linnssen, and H. J. Berendsen. “Essential dynamics of proteins”. In: *Proteins* 17.4 (1993), pp. 412–425.

[30] M. Ringneér. “What is principal component analysis?” In: *Nature Biotechnology* 26.3 (2008), pp. 303–304.

[31] J. D. Chodera et al. “Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics”. In: *J. Chem. Phys.* 126.15, 155101 (2007).

[32] D. Prada-Gracia et al. “Exploring the free energy landscape: From dynamics to networks and back”. In: *PLoS Comp. Biol.* 5.6 (2009).

[33] Christophe Chipot and Andrew Pohorille (Editors. *Free Energy Calculations.* Springer, 2007.

[34] X. Huang et al. “Rapid equilibrium sampling initiated from nonequilibrium data”. In: *Proc. Natl. Acad. USA* 106 (2009), p. 19765.

[35] J. G. Kirkwood. “Statistical Mechanics of Fluid Mixtures”. In: *J. Chem. Phys.* 3 (1935), p. 300.

[36] Kin-Yiu Wong and Darrin M. York. “Exact Relation between Potential of Mean Force and Free-Energy Profile”. In: *J. Chem. Theor. Comput.*. 8 (2002), pp. 3998–4003.

[37] Wouter K. den Otter. “Revisiting the Exact Relation between Potential of Mean Force and Free-Energy Profile”. In: *J. Chem. Theor. Comput.*. 9 (2013), pp. 3861–3865.

[38] R. B. Best and G. Hummer. “Reaction coordinates and rates from transition paths”. In: *Proc. Natl. Acad. Sci. USA* 102.19 (2005), pp. 6732–6737.

[39] G. Hummer. “From transition paths to transition states and rate coefficients”. In: *J. Chem. Phys.* 120.2 (2004), pp. 516–523.

[40] J. D. Chodera and V. S. Pande. “Splitting Probabilities as a Test of Reaction Coordinate Choice in Single-Molecule Experiments”. In: *Phys. Rev. Lett.* 107 (2011), p. 098102.

[41] R. B. Best, G. Hummer, and W. A. Eaton. “Native contacts determine protein folding mechanisms in atomistic simulations”. In: *Proc. Natl. Acad. Sci. USA* 110.44 (2013), pp. 17874–17879.

[42] V. N. Maiorov and G. M. Crippen. “Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins.” In: *J. Mol. Biol.* 235.2 (1994), pp. 625–634.

[43] Sandro Bottaro, Francesco Di Palma, and Giovanni Bussi. “The role of nucleobase interactions in RNA structure and dynamics”. In: *Nucleic Acids Research* (2015).

[44] Eric R. Henry, Robert B. Best, and William A. Eaton. “Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations”. In: *Proceedings of the National Academy of Sciences* 110.44 (2013), pp. 17880–17885.

[45] Samuel S. Cho, Yaakov Levy, and Peter G. Wolynes. “P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes”. In: *Proceedings of the National Academy of Sciences* 103.3 (2005), pp. 586–591.

[46] M. Doi. *Introduction to Polymer Physics*. Clarendon Press, 1996.

[47] Sergei V. Krivov. “On reaction coordinate Optimality”. In: *Journal of Chemical Theory and Computation* 9 (2013), pp. 135–146.

[48] Sergei V. Krivov. “The Free energy landscape analysis of protein (FIP35) folding dynamics”. In: *The Journal of Physical Chemistry B* 115 (2011), pp. 12315–12324.

[49] F. Noé et al. “Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments”. In: *Proc. Natl. Acad. Sci. USA* 108 (2011), pp. 4822–4827.

[50] J. D. Chodera and F. Noé. “Markov state models of biomolecular conformational dynamics”. In: *Curr. Opin. Struct. Biol.* 25 (2014), pp. 135 –144.

[51] J.-H. Prinz et al. “Markov models of molecular kinetics: Generation and validation”. In: *J. Chem. Phys.* 134.17, 174105 (2011).

[52] F. Noé. “Probability Distributions of molecular observables computed from Markov Models”. In: *J. Chem. Phys.* 128 (2008), p. 244103.

[53] S. Dasgupta and P. Long. “Performance guarantees for hierarchical clustering”. In: *J. Comput. Syst. Sci.* 70.4 (2005), pp. 555–569.

[54] M. Senne et al. “Emma - a software package for markov model building and analysis”. In: *J. Chem. Theory and Comput.* 8 (2012), pp. 2223–2238.

[55] Rafael Tapia-Rojo et al. “Mechanical unfolding of a simple model protein goes beyond the reach of one dimensional descriptions”. In: *Journal of Chemical Physics* 141 (2014), p. 135102.

[56] D. Prada Gracia. “Paisajes de Energìa Libre en modelos de biomolèculas”. PhD thesis. Universidad de Zaragoza, 2011.

[57] F. Noé et al. “Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states”. In: *J. Chem. Phys.* 126.15 (2007), p. 155102.

[58] R. Tapia-Rojo et al. “Mesoscopic model for free-energy-landscape analysis of DNA sequences”. In: *Phys. Rev. E* 86.2 (2012).

[59] R. Tapia-Rojo et al. “Mesoscopic Model and Free Energy Landscape for Protein-DNA Binding Sites: Analysis of Cyanobacterial Promoters”. In: *PLoS Comput. Biol.* 10.10 (2014), e1003835.

[60] D. J. Wales. “Energy landscapes: some new horizons”. In: *Curr. Opin. Struct. Biol.* 20 (2010), pp. 3–10.

[61] P. Matzner, C. Schutte, and E. Vanden-Eijnden. “Transition path theory for Markov jump processes”. In: *Multiscale Model Simul.* 7 (2009), pp. 1192–1219.

[62] E. Vanden-Eijnden. “Toward a theory of transition paths”. In: *J. Stat. Phys.* 123 (2006), pp. 503–523.

[63] W. G. Noid. “Perspective: Coarse-grained models for biomolecular systems”. In: *J. Chem. Phys.* 139.9, 090901 (2013), pp. –. DOI: `http://dx.doi.org/10.1063/1.4818908`. URL: `http://scitation.aip.org/content/aip/journal/jcp/139/9/10.1063/1.4818908`.

[64] H. S. Chan et al. “Cooperativity, Local-Nonlocal Coupling, and Nonnative Interactions: Principles of Protein Folding from Coarse-Grained Models”. In: *Annu. Rev. Phys. Chem.* 62 (2011), pp. 310–326.

[65] C. Hyeon and D. Thirumalai. “Capturing the essence of folding and functions of biomolecules using coarse-grained models”. In: *Nature Comm.* 2.487 (2011).

[66] V. Tozzini, W. Rocchia, and J. A. McCammon. “Mapping all-atom models onto one-bead Coarse Grained Models: general properties and applications to a minimal polypeptide model”. In: *J. Chem. Theory Comput.* 2.3 (2006), pp. 667–673.

[67] H. Teketomi, Y. Ueda, and N. Gō. “Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions.” In: *Int. J. Pept. Protein Res.* 7.6 (1975), pp. 445–459.

[68] Monique M. Tirion. “Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis”. In: *Physical Review Letters* 77 (1996), p. 1905.

[69] Konrad Hilsen. “Analysis of domain motions by approximate normal mode calculations”. In: *Proteins* 33.3 (1998), pp. 417–29.

[70] A. R. Atilgan et al. “Anisotropy of fluctuation dynamics of proteins with an elastic network model”. In: *Biophysics Journal* 80.1 (2001), pp. 505–515.

[71] I. Bahar, A. R. Atilgan, and B. Erman. “Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential”. In: *Folding and Design* 2 (1997), pp. 173–181.

[72] T. Haliloglu, I. Bahar, and B. Erman. “Gaussian dynamics of folded proteins”. In: *Physical Review Letters* 79 (1997), pp. 3090–3093.

[73] M. Cotallo-Abán et al. “Allostery of actin filaments: Molecular dynamics simulations and coarse-grained analysis”. In: *Proc.Natl. Acad. Sci. USA* 102.37 (2005), pp. 13111–13116.

[74] M. Cotallo-Abán et al. “Analysis of Apoflavodoxin Folding Behavior with Elastic Network Models”. In: *From Physics to Biology: the interface between experiment and cumputation(AIP Conference Proceedings)* 851 (2006), pp. 135–149.

[75] S Tanaka and H. A. Scheraga. “Model of protein folding: incorporation of a one-dimensional short-range (Ising) model into a three-dimensional model.” In: *Proceedings of the National Academy of Sciences* 74.4 (1977), pp. 1320–1323.

[76] V. N. Maiorov and G. M. Crippen. “Contact potential that recognizes the correct folding of globular proteins”. In: *Journal of Molecular Biology* 227.3 (1992), pp. 876–888.

[77] J. D. Honeycutt and D. Thirumalai. “Metastability of the folded states of globular proteins”. In: *Proc. Natl. Acad. Sci. USA* 87.9 (1990), pp. 3526–3529.

[78] R. S. Berry et al. “Linking topography of its potential surface with the dynamics of folding of a protein?model”. In: *Proc. Natl. Acad. Sci. USA* 94.18 (1997), pp. 9520–9524.

[79] A. Imparato, S. Luccioli, and A. Torcini. “Reconstructing the Free-Energy Landscape of a Mechanically Unfolded Model Protein”. In: *Phys. Rev. Lett.* 99 (16 2007), p. 168101.

[80] S. Luccioli, A. Imparato, and A. Torcini. “Free-energy landscape of mechanically unfolded model proteins: Extended Jarzinsky versus inherent structure reconstruction”. In: *Phys. Rev. E* 78 (2008), p. 031907.

[81] H. S. Greenside and E. Helfand. “Numerical Integration of Stochastic Differential Equations”. In: *Bell Labs Technical Journal* 60.8 (1981), pp. 1927–1981.

[82] David A. Evans and David J. Wales. “Free energy landscapes of model peptides and proteins”. In: *Journal of Chemical Physics* 118.8 (2003), pp. 3891–3897.

[83] H. Chen et al. “Dynamics of Equilibrium Folding and Unfolding Transitions of Titin Immunoglobulin Domain under Constant Forces”. In: *J. Am. Chem. Soc.* 137 (2015), pp. 3540–3546.

[84] M. T. Woodside and S. M. Block. “Reconstructing Folding Energy Landscapes by Single-Molecule Force Spectroscopy”. In: *Annu. Rev. Biophys.* 43 (2014), pp. 19–39.

[85] M. Bastian, S. Heynmann, and M. Jacomy. “Gephi: An Open Source Software for Exploring and Manipulating Networks”. In: *Proceedings of the International AAAI Conference on Weblogs and Social Media, San Jose, California* (2009).

[86] V. D. Blondel et al. “Fast unfolding of communities in large networks”. In: *J. Stat. Mech.* 2008.10 (2008), P10008.

[87] Bruce Alberts et al. *Molecular Biology of the Cell*. New York: Garland Science, 2002.

[88] J. D. Watson and F. H. C. Crick. “A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171 (1953), pp. 737–738.

[89] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson. “Molecular Structure of Deoxypentose Nucleic Acids”. In: *Nature* 171 (1953), pp. 738–740.

[90] R. Franklin and R. G. Gosling. “Molecular Configuration of Deoxypentose Nucleic Acids”. In: *Nature* 171 (1953), pp. 740–741.

[91] J. D. Watson and F. H. C. Crick. “Genetic Implications of the structure of Deoxyribonucleic Acid”. In: *Nature* 171 (1953), pp. 964–967.

[92] R. Franklin and R. areG. Gosling. “Evidence for 2-Chain Helix in Crystalline Structure of Sodium Deoxyribonucleate”. In: *Nature* 172 (1953), pp. 156–157.

[93] F. Crick. “Central Dogma of Molecular Biology”. In: *Nature* 227 (1970), pp. 561–563.

[94] *et. al.* M. Kellis. “Defining functional DNA elements in the human genome”. In: *Proc. Natl. Acad. Sci. USA* 111.17 (2014), pp. 6131–6138.

[95] Chris R. Calladine et al. *Understanding DNA*. Elsevier Academic Press, 2004.

[96] G. A. Patikoglou et al. “TATA element recognition by the TATA box-binding protein has been conserved throughout evolution”. In: *Genes and Development* 13.24 (1999), pp. 3217–3230.

[97] B. S. Alexandrov et al. “DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation”. In: *Nucleic Acids Res.* 38 (2010), pp. 1790–1795.

[98] G. Weber, J. W. Essex, and C. Neylon. “Probing the microscopic flexibility of DNA from melting temperatures”. In: *Nature Physics* 5 (2009), pp. 769–773.

[99] R. B. Inman and R. L. Baldwin. “Helix?random coil transitions in DNA homopolymer pairs”. In: *J. Mol. Biol.* 8 (1964), pp. 452–469.

[100] M. Peyrard. *Nonlinear dynamics and statistical physics of DNA*. 2004.

[101] R. M. Wartell and A. S. Benight. “Thermal denaturation of DNAmolecules: a comparison of theory with experiments”. In: *Phys. Rep.* 126 (1985), p. 67.

[102] F. B. Fuller. “The writhing number of a space curve”. In: *Proc. Natl. Acad. Sci. USA* 68 (1971), pp. 815–819.

[103] M. Amouyal and H. Buc. “Topological unwinding of strong and weak promoters by RNA polymerase: a comparison between the lac wild-type and UV5 sites of *E. coli*”. In: *J. Mol. Biol.* 195 (1987), pp. 795–808.

[104] A. Pérez, F. J. Luque, and M. Orozco. “Frontiers in molecular dynamics simulations of DNA”. In: *Acc. Chem. Res.* 45.2 (2012), pp. 196–205.

[105] N. Theodorakopoulos. *Phase transitions in homogeneous biopolymers: basic concepts and methods. ?* 2008.

[106] M. Peyrard and A. R. Bishop. “Statistical mechanics of a nonlinear model for DNA denaturation”. In: *Phys. Rev. Lett.* 62.23 (1989), pp. 2755–2758.

[107] T. Dauxois, M. Peyrard, and A. R. Bishop. “Dynamics and thermodynamics of a nonlinear model for DNA denaturation.” In: *Phys. Rev. E* 47.1 (1993), pp. 684–695.

[108] T. Dauxois, M. Peyrard, and A. R. Bishop. “Entropy-driven DNA denaturation”. In: *Phys. Rev. E* 1 (), R44(R).

[109] T. S. Van Erp et al. “Can one predict DNA transcription start sites by studying bubbles?” In: *Phys. Rev. Lett.* 95 (2005).

[110] C. H. Choi et al. “Comment on :Can One Predict DNA Transcription Start Sites by Studying Bubbles?””. In: *Phys. Rev. Lett.* 96 (2006), p. 239801.

[111] T. S. van Erp et al. “van Erp et al. Reply:” in: *Phys. Rev. Lett.* 96 (2006), p. 239802.

[112] C. J. Benham and R. R. P. Singh. “Comment on ?Can One Predict DNA Transcription Start Sites by Studying Bubbles?” In: *Phys. Rev. Lett.* 97 (2006), p. 059801.

[113] T. S. van Erp et al. “van Erp et al. Reply:” in: *Phys. Rev. Lett.* 97 (2006), p. 059802.

[114] G. Weber. “Sharp DNA denaturation due to solvent interaction”. In: *Europhys. Lett.* 73.5 (2006), p. 806.

[115] K. Drukker, G. Wu, and G. C. Schatz. “Model simulations of DNA denaturation dynamics”. In: *J. Chem. Phys.* 114.1 (2001), pp. 579–590.

[116] R. Tapia-Rojo, J. J. Mazo, and F. Falo. “Thermal and mechanical properties of a DNA model with solvation barrier”. In: *Phys. Rev. E* 82.3 (2010).

[117] A. Campa and A. Giansanti. “Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains”. In: *Phys. Rev. E* 58.3 (1998), pp. 3585–3588.

[118] G. Kalosakas et al. “Sequence-specific thermal fluctuations identify start sites for DNA transcription”. In: *Europhys. Lett.* 68.1 (2004), p. 127.

[119] M. Peyrard, S. Cuesta-López, and D. Angelov. “Experimental and theoretical studies of sequence effects on the fluctuation and melting of short DNA molecules.” In: *J. of Phys.: Condens. Matter* 21.3 (2009), p. 034103.

[120] E. Giudice, P. Várnai, and R. Lavery. *Base pair opening within B-DNA: Free energy pathways for GC and AT pairs from umbrella sampling simulations.* 2003.

[121] B. S. Alexandrov et al. “A nonlinear dynamic model of DNA with a sequence-dependent stacking term”. In: *Nucleic Acids Res.* 37.7 (2009), pp. 2405–2410.

[122] M Peyrard, S Cuesta-López, and D Angelov. “Experimental and theoretical studies of sequence effects on the fluctuation and melting of short DNA molecules”. In: *Journal of Physics: Condensed Matter* 21.3 (2009), p. 034103.

[123] G. Farge et al. “Protein sliding and DNA denaturation are essential for DNA organization by human mitochondrial transcription factor A”. In: *Nature Comm.* 3 (2009), p. 1013.

[124] R. Rohs et al. “The role of DNA shape in protein-DNA recognition”. In: *Nature* 461 (2009), pp. 1248–1253.

[125] B. D. Starr, B. C. Hoopes, and D. K. Hawley. “DNA bending is an important component of site-specific recognition by the TATA binding protein”. In: *Jour. Mol. Biol.* 250 (1995), pp. 434–446.

[126] B. S. Alexandrov et al. “Toward a detailed description of the thermally induced dynamics of the core promoter”. In: *PLoS Comput. Biol.* 5.3 (2009).

[127] A. B. Alexandrov et al. “Pre-melting dynamics of DNA and its relation to specific functions”. In: *J. Phys. Condens. Matter* 21 (2009), p. 034107.

[128] K. Nowak-Lovato et al. “Binding of nucleoid-associated protein Fis to DNA is regulated by DNA breathing dynamics.” In: *PLoS Comput. Biol.* 9 (2013), e1002881.

[129] B. S. Alexandrov et al. “DNA breathing dynamics distinguishing binding from nonbinding consensus sites for transcription factor YY1 in cells”. In: *Nucleic Acids Res.* 40 (2012), pp. 10115–10123.

[130] P. H. von Hippel and O. G. Berg. “Facilitated target location in biological systems”. In: *J. Biol. Chem.* 264 (1989), p. 675.

[131] S. E. Halford and J. F. Marko. “How do site?specific DNA?binding proteins find their targets?” In: *Nucleic Acids Res.* 32.10 (2004), pp. 3040–3052.

[132] P. Samorí (editor). *Scanning Probe Microscopies Beyond Imaging: Manipulation of Molecules and Nanostructures.* Wiley-VCH, 2006.

[133] R. Schleif. *Genetics and Molecular Biology.* Addison-Wesley, Reading, MA, 1993.

[134] Z. Wunderlich and L. A. Mimy. “Spatial effects on the speed and reliability of protein-DNA search.” In: *Nucleic Acid Res.* 36.11 (2008), pp. 3570–3578.

[135] M. Sheinman et al. “Classes of fast and specific search mechanisms for proteins on DNA”. In: *Rep. Prog. Phys* 75.2 (2012), p. 026601.

[136] G. Kalosakas et al. “Lengthscales and cooperativity in DNA bubble formation”. In: *Eur. Phys. Lett.* 68 (2004), p. 127.

[137] A. Apostolaki and G. Kalosakas. “Targets of DNA-binding proteins in bacterial promoter regions present enhanced probabilities for spontaneous thermal openings.” In: *Phys. Biol.* 8 (2011), p. 026006.

[138] L. Bintu et al. “Transcriptional regulation by the numbers: models”. In: *Curr. Opin. Genet. Dev.* 15 (2005), p. 116.

[139] B. J. Haas et al. “How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?” In: *BMC genomics* 13 (2012), pp. 734–745.

[140] W. R. Hess. “Cyanobacterial genomics for ecology and biotechnology”. In: *Curr. Opin. Microbiol.* 14 (2011), pp. 608–614.

[141] E. Flores and A. Herrero. “Compartmentalized function through cell differentiation in filamentous cyanobacteria”. In: *Nat. Rev. Microbiol.* 8 (2010), pp. 39–50.

[142] C. *Anabaena sp. PCC 7120. Cyanobase*. URL: `http://genome.microbedb.jp/cyanobase/Anabaena`.

[143] A. Herrero, A. M. Muro-Pastor, and E. Flores. “Nitrogen control in cyanobacteria”. In: *J. Bacteriol.* 183 (2001), pp. 411–425.

[144] P. Russo and A. Cesario. “New anticancer drugs from marine cyanobacteria”. In: *Curr. Drug Targets* 13 (2012), pp. 1048–1053.

[145] J. Sjholm, P. Oliveira, and P. Lindblad. “Transcription and regulation of the bidirectional hydrogenase in the cyanobacterium *Nostoc* sp. strain PCC 7120”. In: *Appl. Environ. Microbiol.* 73 (2007), pp. 5453–5446.

[146] B. Floriano, A. Herrero, and E. Flores. “Analysis of expression of the *argC* and *argD* genes in the cyanobacterium *Anabaena* sp. strain PCC7120”. In: *J. Bateriol.* 176 (1994), pp. 6397–6401.

[147] R. A. Mella-Herrera, M. R. Neunuebel, and J. W. Golden. “*Anabaena* sp. strain PCC7120 *conR* contains a LytR-CpsA-Psr domain, is developmentally regulated, and is essential for diazotrophic growth and heterocyst morphogenesis”. In: *Microbiology* 157 (2011), pp. 617–626.

[148] M. E. Mulligan and R. Haselkorn. “Nitrogen fixation (*nif*) genes of the cyanobacterium *Anabaena* species strain PCC 7120. The *nifB-fdxN-nifS-nifU* operon”. In: *J. Biol. Chem.* 26 (1989), pp. 19200–19207.

[149] J. A. Hernández. “Ferruc. Uptake Regulator (Fur) en *Anabaena* Sp. PCC7120: Caracterización bioquímica, análisis de genes regulados y estudio de la regulación del propio represor.” PhD thesis. Universidad de Zaragoza, 2004.

[150] A. Herrero et al. “Cellular differentiation and the NtcA transcription factor in filamentous cyanobacteria”. In: *FEMS Microbiol. Rev.* 28 (2004), pp. 469–487.

[151] A. M. Muro-Pastor et al. “Mutual dependence of the expression of the cell differentiation regulatory protein HetR and the global nitrogen regulator NtcA during heterocyst development”. In: *Mol. Microbiol.* 44 (2002), pp. 1377–1385.

[152] A. Valladares et al. “Constitutive and nitrogen-regulated promoters of the *petH* gene encoding ferredoxin: $NADP^+$ reductase in the heterocyst-forming cyanobacterium *Anabaena* sp.” In: *FEBS Lett.* 23 (1999), pp. 159–164.

[153] B. Floriano et al. “Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120”. In: *Proc. Natl. Acad. Sci. USA* 108 (2011), pp. 20130–20135.

[154] E. Olmedo-Verde et al. “Role of two NtcA-binding sites in the complex ntcA gene promoter of the heterocyst-forming cyanobacterium Anabaena sp. strain PCC7120”. In: *J. Bateriol.* 190 (2008), pp. 7584–7759.

[155] S. López-Gomollón et al. “New insights into the role of Fur proteins: FurB(All2473) from *Anabaena* protects DNA and increases cell survival under oxidative stress”. In: *Biochem. H.* 15 (2009), pp. 201–207.

[156] J. A. Hernández et al. “Interaction of FurA from *Anabaena* sp. PCC7120 with DNA: a reducing enviroment and the presence of $Mn(2^+)$ are positive effectors in the binding to *isib* and *furA* promoters”. In: *Biometals* 19 (2006), pp. 259–268.

[157] A. M. Muro-Pastor et al. “Mutual dependence of the expression of the cell differentiation regulatory protein HetR and the global nitrogen regulator NtcA during heterocyst development”. In: *Mole Microbiol* 44 (2002), pp. 1377–1385.

[158] A. Herrero et al. “Cellular differentiationand the NtcA transcription factor in filamentous cyanobacteria”. In: *FEMS Microbiol. Rev.* 28 (2004), pp. 469–487.

[159] J. R. Goñi et al. “Determining promoter location based on DNA structure first-principles calculations”. In: *Genome Biol.* 8 (2007), R263.

[160] V. B. Bajic et al. “Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment”. In: *Genome Biol.* 7 (2006), S3.

[161] R. Mallik et al. “Cytoplasmic dynein functions as a gear in response to load”. In: *Nature* 427 (2004), pp. 649–652.

[162] A. Yildiz et al. “Kinesin walks hand-over-hand”. In: *Science* 303 (2004), pp. 676–678.

[163] D. Altman, H. L. Sweeney, and J. A. Spudich. “The mechanism of myosin VI translocation and its load-induced anchoring”. In: *Cell* 116 (2004), pp. 737–749.

[164] R. Mallik et al. “Cytoplasmic dynein functions as a gear in response to load”. In: *Nature* 427 (2004), pp. 649–652.

[165] C. Bustamante, J. Liphardt, and F. Ritort. “The Nonequilibrium Thermodynamics of Small Systems”. In: *Physics Today* 58 (2005), pp. 43–48.

[166] S. Weiss. “Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy”. In: *Nat. Struct. Biol.* 7 (2000), pp. 724–729.

[167] T. Ha. “Single-molecule fluorescence resonance energy transfer”. In: *Methods* 25 (2001), pp. 78–86.

[168] W. J. Greenleaf, M. T. Woodside, and S. M. Block. "High-Resolution, Single-Molecule Measurements of Biomolecular Motion". In: *Annu. Rev. Biophys. Biomol. Struct.* 36 (2007), pp. 171–190.

[169] K. C. Neuman and A. Nagy. "Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy". In: *Nat. Methods* 5.6 (2008), pp. 491–505.

[170] F. Ritort. "Single-molecule experiments in biological physics: methods and applications". In: *J. Phys.: Condens. Matter* 18 (2006), R531–R583.

[171] M. Rief et al. "Reversible unfolding of individual titin immunoglobulin domains by AFM". In: *Science* 276 (1997), pp. 1109–1112.

[172] R. O. Hynes. *Fribronectins*. Springer: New York, 1990.

[173] A. Matouschek. "Protein unfolding-an important process in vivo?" In: *Curr. Opin. Struct. Biol.* 13 (2003), pp. 98–109.

[174] J. A. Kenniston et al. "Linkage between ATP consumption and mechanical unfolding during the protein processing reactions of an AAA+ degradation machine". In: *Cell* 114 (2003), pp. 511–520.

[175] H. R. Saibil and N. A. Ranson. "The chaperonin folding machine". In: *Trends. Biochem. Sci.* 27.12 (2002), pp. 627–632.

[176] M. Shtilerman, G. H. Lorimer, and S. W. Englander. "Chaperonin function: folding by forced unfolding". In: *Science* 284.5415 (1999), pp. 822–825.

[177] K. Svoboda and S. M. Block. "Biological applications of optical forces". In: *Annu. Rev. Biophys. Biomol. Struct.* 23 (1994), pp. 247–285.

[178] K. C. Neuman et al. "Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking". In: *Cell* 115 (2003), pp. 437–447.

[179] *et. al.* D. E. Smith. "The bacteriophage phi 29 portal motor can package DNA against a large internal force". In: *Nature* 413 (2001), pp. 748–752.

[180] J. Liphardt et al. "Reversible unfolding of single RNA molecules by mechanical force". In: *Science* 292 (2001), pp. 733–737.

[181] A. Alemany et al. "Experimental free-energy measurements of kinetic molecular states using fluctuation theorems". In: *Nature Physics* 8 (2012), pp. 688–694.

[182] M. T. Woodside et al. "Direct measurement of the full, sequence-dependent folding landscape of nucleic acids". In: *Science* 314 (2006), pp. 1001–1004.

[183] M. S. Z. Kellermayer et al. "Folding-unfolding transitions in single titin molecules characterized with laser tweezers". In: *Science* 276 (1997), pp. 1112–1116.

[184] C. Cecconi et al. "Direct observation of the three-state folding of a single protein molecule". In: *Science* 309 (2005), pp. 2057–2060.

[185] D. B. Ritchie and M. T. Woodside. "Probing the structural dynamics of proteins and nucleic acids with optical tweezers". In: *Curr. Opin. Struct. Biol.* 34 (2015), pp. 43–51.

[186] J. Gore et al. "DNA overwinds when stretched". In: *Nature* 442 (2006), pp. 836–839.

[187] T. R. Strick et al. “The elasticity of a single supercoiled DNA molecule”. In: *Science* 271 (1996), pp. 1835–1837.

[188] D. A. Koster et al. “Friction and torque govern the relaxation of DNA supercoils by eukaryotic topoisomerases IB”. In: *Nature* 434 (2005), pp. 671–674.

[189] T. R. Strick, V. Croquette, and D. Bensimon. “Single-molecule analysis of DNA uncoiling by a type II topoisomerase”. In: *Nature* 404 (2000), pp. 901–904.

[190] H. Chen et al. “Dynamics of Equilibrium Folding and Unfolding Transitions of Titin Immunoglobulin Domain under Constant Forces”. In: *J. Am. Chem. Soc.* 137 (2015), pp. 3540–3546.

[191] I. Popa et al. “Force dependency of biochemical reactions measured by single-molecule force-clamp spectroscopy”. In: *Nat. Prot.* 8.7 (2013), pp. 1261–1276.

[192] G. Lee et al. “Nanospring behavior of ankyrin repeats”. In: *Nature* 440 (2006), pp. 246–249.

[193] R. B. Best et al. “Can nonmechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation”. In: *Biophys. J.* 81 (2001), pp. 2344–2356.

[194] H. Dietz and M. Rief. “Exploring the energy landscape of GFP by single-molecule mechanical experiments”. In: *Proc. Natl. Acad. Sci. USA* 101 (2004), pp. 16192–16197.

[195] J. M. Fernandez and H. B. Li. “Force-clamp spectroscopy monitors the folding trajectory of a single protein”. In: *Science* 303 (2004), pp. 1674–1678.

[196] M. Carrion-Vázquez et al. “The mechanical stability of ubiquitin is linkage dependent”. In: *Nat. Struct. Biol.* 10 (2003), pp. 674–676.

[197] H. Dietz et al. “Anisotropic deformation response of single protein molecules”. In: *Proc Natl. Acad. Sci. USA* 103 (2006), pp. 12724–12728.

[198] J. A. Rivas-Pardo et al. “Identifying Sequential Substrate Binding at the Single-Molecule Level by Enzyme Mechanical Stabilization”. In: *ACS Nano* 9.4 (2015), pp. 3996–4005.

[199] V. Barsegov and D. Thirumalai. “Dynamics of unbinding of cell adhesion molecules: Transition from catch to slip bonds”. In: *Proc. Natl. Acad. USA* 102.6 (2003), pp. 1835–1839.

[200] S. Kirmizialtin, L. Huang, and D. E. Makarov. “Topography of the free-energy landscape probed via mechanical unfolding of proteins”. In: *J. Chem. Phys.* 122.23 (2005), p. 234915.

[201] P. Haenggi, peter Talkner, and Michal Borkovec. “Reaction-rate theory: fifty years after Kramers”. In: *Rev. Mod. Phys.* 62.2 (1990), pp. 251–332.

[202] C. Jarzynski. “Nonequilibrium work relations: foundations and applications”. In: *Eur. Phys. J. B* 64 (2008), pp. 331–340.

[203] E. Evans and K. Ritchie. “Dynamic Strength of Molecular Adhesion Bonds”. In: *Biophys. J.* 72 (1997), pp. 1541–1555.

[204] Olga K. Dudko, Gerhard Hummer, and Attila Szabo. “Intrinsic Rates and Activation Free Energies from Single-Molecule Pulling Experiments”. In: *Phys. Rev. Lett.* 06 (2006), p. 108101.

[205] C. Jarzynski. “Nonequilibrium Equality for Free Energy Differences”. In: *Phys. Rev. Lett.* 78 (1997), p. 2690.

[206] H. A. Kramers. “Brownian motion in a field of force and the diffusion model of chemical reactions”. In: *Physica* 7.4 (1940), pp. 284–304.

[207] J.J. Mazo, O.Y. Fajardo, and D. Zueco. “Thermal activation at moderate-to-high and high damping: finite barrier effects and force spectroscopy”. In: *J. Chem. Phys.* 14.138 (2013), p. 104105.

[208] G.I. Bell. “Models for the specific adhesion of cells to cells”. In: *Science* 200.432 (1978), pp. 618–627.

[209] E. Evans, D. Berk, and A. Leung. “Detachment of agglutinin-bonded red blood cells. I. Forces to rupture molecular-point attachments”. In: *Biophys. J.* 59.4 (1991), pp. 838–848.

[210] S. Izrailev et al. “Molecular Dynamics Study of Unbinding of the Avidin-Biotin Complex”. In: *Biophys. J.* 72 (1997), pp. 1568–1581.

[211] G. Hummer and A. Szabo. “Kinetics from Nonequilibrium Single-Molecule Pulling Experiments”. In: *Biophys. J.* 85 (2003), pp. 5–15.

[212] O. K. Dudko et al. “Beyond the Conventional Description of dynamic force spectroscopy of adhesion bonds”. In: *Biophys. J.* 100.20 (2003), pp. 11378–11381.

[213] A. Garg. “Escape-field distribution for escape from a metastable potential well subject to a steadily increasing bias field”. In: *Phys. Rev. B* 51 (1995), p. 15592.

[214] A. Maitra and G. Arya. “Model Accounting for the Effects of Pulling-Device Stiffness in the Analyses of Single-Molecule Force Measurements”. In: *Phys. Rev. Lett.* 104 (2010), p. 108301.

[215] A. Maitra and G. Arya. “Influence of pulling handles and device stiffness in single-molecule force spectroscopy”. In: *Phys. Chem. Chem. Phys.* 13 (2011), pp. 1836–1842.

[216] C. Jarzynski. “Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach”. In: *Phys. Rev. E* 56 (1997), p. 5018.

[217] C. Jarzysnki. “Equalities and Inequalities: Irreversibility and the Second Law of Thermodynamics at the Nanoscale”. In: *Annu. Rev. Condens. Matt. Phys.* 2 (2011), pp. 329–351.

[218] A. F. Oberhauser (editor). *Single-Molecule Studies of Proteins*. Springer, 2013.

[219] K. Sekimoto. *Stochastic Energetics*. Springer, 2010.

[220] J. Liphardt et al. “Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski’s equality”. In: *Science* 296 (2002), pp. 1832–1835.

[221] G. Hummer and A. Szabo. “Free energy reconstruction from nonequilibrium singlemolecule pulling experiments”. In: *Proc. Natl. Acad. Sci. USA* 98 (2001), pp. 3658–3661.

[222] G. E. Crooks. “Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences”. In: *Phys. Rev. E* 60 (1999), p. 2721.

[223] D. J. Evans and D. J. Searles. “The Fluctuation Theorem”. In: *Adv. Phys.* 51.7 (2002), pp. 1529–1585.

[224] A. Pohorille, C. Jarzysnki, and C. Chipot. “Good practices in free-energy calculations”. In: *J. Phys. Chem. B* 114 (2010), p. 114.

[225] D. M. Zuckerman and T. B. Woolf. “Theory of a Systematic Computational Error in Free Energy Differences”. In: *Phys. Rev. Lett.* 89.180602 (2002), p. 180602.

[226] J. Gore, F. Ritort, and C. Bustamante. “Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements”. In: *Proc. Natl. Acad. Sci. USA* 100 (2003), pp. 12564–12569.

[227] C. H. Bennett. “Efficient Estimation of Free Energy Differences from Monte Carlo Data”. In: *J. Comp. Phys.* 22 (1976), pp. 245–268.

[228] M. R. Shirts et al. “Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods”. In: *Phys. Rev. Lett.* 91 (2003), p. 140601.

[229] R. Tapia-Rojo. “Characterizing mechanical unbinding of biological complexes: from forces to free energies”. In: *Whatever jounal* (2015), p. 666.

[230] M. Martínez-Júlvez, M. Medina, and A. Velázquez-Campoy. “Binding Thermodynamics of Ferredoxin: NADP+ Reductase: Two Different Protein Substrates and One Energetics”. In: *Biophys. J.* 96 (2009), p. 4966.

[231] Milagros Medina and Carlos Gómez-Moreno. “Interaction of Ferredoxin-NADP+ Reductase with Its Substrates: Optimal Interaction for Efficient Electron Transfer”. In: *Photosynthesis Research* 79 (2004), pp. 113–131.

[232] M. Martínez-Júlvez, M. Medina, and C. Gómez-Moreno. “Ferredoxin-NADP(+) reductase uses the same site for the interaction with ferredoxin and flavodoxin”. In: *J. Biol. Inorg. Chem.* 4.5 (1999), pp. 568–578.

[233] Juan Fernandez-Recio et al. “Docking analysis of transitent complexes: Interaction of Ferredoxin-NADP$^+$ Reductase with Ferredoxin and Flavodoxin.” In: *Proteins* 72 (2008), pp. 848–862.

[234] Milagros Medina. “Structural and mechanistic aspects of flavoproteins: photosynthetic electron transfer from photosystem I to NADP$^+$”. In: *FEBS Journal* 276 (2009), pp. 3942–3958.

[235] Guillermina Goñi et al. “Flavodoxin: a compromise between efficiency and versatility in the electron transfer from Photosystem I to Ferredoxin-NADP$^+$ reductase”. In: *Biochem. Biophys. Acta-Bioenergetics* 1787 (2009), pp. 144–154.

[236] Ana Serrano et al. “Flavodoxin-mediated electron transfer from Photosystem I to Ferredoxin-NADP$^+$ reductase in Anabaena: Role of Flavodoxin hydrophobic residues in protein-protein interaction”. In: *Biochemistry* 47 (2008), pp. 1207–1217.

[237] R. Morales et al. “A redox-dependent interaction between two electron-transfer partners involved in photosynthesis”. In: *EMBO report* 1 (2000), pp. 271–276.

[238] C. Marcuello et al. “Mechanostability of the Single-Electron-Transfer Complexes of Anabaena Ferredoxin NADP$^+$ Reductase”. In: *Chem. Phys. Chem.* 16 (2015), pp. 3161–3169.

[239] C. Marcuello Anglés. “Mecanismos Catalíticos en Sistemas Proteicos Estudiados a Nivel de Molécula Única”. PhD thesis. Universidad de Zaragoza, 2014.

[240] C. Marcuello et al. “An efficient method for enzyme immobilization evidenced by atomic force microscopy.” In: *Protein Eng. Des. Sel.* 25 (2012), p. 715.

[241] A. R. Bizarri and S. Cannistraro. “The application of atomic force spectroscopy to the study of biological complexes undergoing a biorecognition process”. In: *Chem. Soc. Rev.* 39 (2010), pp. 734–749.

[242] S. Getfert and P. Reimann. “Hidden Multiple Bond Effects in Dynamic Force Spectroscopy”. In: *Biophys. J.* 102 (2012), pp. 1184–1193.

[243] C. Hyeon and D. Thirumalai. “Multiple Barriers in Forced Rupture of Protein Complexes”. In: *J. Chem. Phys.* 137 (2012), p. 055103.

[244] A. Mossa et al. “Measurement of work in single-molecule pulling experiments”. In: *J. Chem. Phys.* 130 (2009), p. 234116.

[245] Anna Rita Bizzarri and Salvatore Cannistraro. “Free energy evaluation of the p53-Mdm2 complex from unbinding work measured by dynamic force spectroscopy”. In: *Phys. Chem. Chem. Phys.* 13 (2009), pp. 2738–2743.

[246] M. Palassini and F. Ritort. “Improving Free-Energy Estimates from Unidirectional Work Measurements: Theory and Experiment”. In: *Phys. Rev. Lett.* 107 (2011), p. 060601.

[247] B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Society for Industrial and Applied Mathematicsr, 1980.

[248] Anna Rita Bizzarri and Salvatore Cannistraro. “Single molecule force spectroscopy by AFM indicates helical structure of poly(ethylene-glycol) in water”. In: *New Journal of Physics* 1 (1999), pp. 6.1–6.11.

[249] P. C. Weber et al. “Crystallographic and Thermodynamic Comparison of Natural and Synthetic Ligands Bound to Streptavidin”. In: *J. Am. Chem. Soc.* 114 (1992), p. 3197.

[250] P. C. Weber et al. “Energetics of ligand binding to proteins”. In: *Adv. Protein Chem.* 29 (1975), p. 1.

[251] E. P. Wojcikiewicz et al. “Force Spectroscopy of LFA-1 and Its Ligands, ICAM-1 and ICAM-2”. In: *Biomacromolecules* 7 (2006), p. 3188.