Traversing the free-energy pathways of intricate biomolecular processes

*Enhanced simulation development and applications*

Pérez de Alba Ortíz, A.

**Publication date**
2021

**Citation for published version (APA):**
Introduction

How can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?

Erwin Schrödinger

A consensus on the definition of life is notoriously difficult to reach, but most efforts at doing so agree that living systems perform complex functions such as reproduction, metabolism, homeostasis and response to stimuli [1–4]. In living organisms, these functions are carried out by a fascinating machinery of biomolecules [5]. The inner workings of these systems, however, often fall in a vexed range of time and space scales [6]. Many molecular processes occur too quickly and at too small scales to be resolved by experimental techniques. Ironically, these same processes also happen at system sizes and rates that are too large and infrequent to be sampled by most standard molecular simulation methods. While various enhanced sampling techniques are able to expand the timescales accessible on simulations, the prowess of these schemes steeply decays for highly intricate biomolecular transitions. In this thesis, we aim to unveil some of the mysterious microscopic mechanisms of biomolecular processes by advancing the capacities of simulation methods. Particularly, we focus on incrementing the level of complexity that can be handled with current state-of-the-art enhanced sampling techniques. Therefore, as part of our quest to understand complex biomolecular function, we build a framework to sample intricate biomolecular transition pathways and apply it to conformational and chemical changes in various types of biomolecules. The resulting advancements provide us with unique microscopic insights about: the rugged free-energy landscapes, the mechanistic pathways, the kinetic rates and thermodynamic equilibria, and, ultimately, the intricate functioning of various examples of life’s little machines.
1.1. Biomolecules: structure and function

The astonishing variety of terrestrial living organisms is in fact notably uniform at the molecular scale [5]. While eukaryote and prokaryote cells might differ in their structure—i.e. by the presence or absence of a well-defined nucleus, respectively—they are built from similar biomolecules, with only slight variations. This uniformity points to the existence of a common ancestor, from which modern organisms have evolved. Across different branches of the tree of life, biomolecules have similar structures and perform similar functions. For example, nucleic acids generally act as genetic information carriers, containing and transferring the instructions of how to assemble proteins. In turn, proteins are the workhorses of living systems, and perform a plethora of functions; including providing structure, regulating responses and catalyzing the formation of other biomolecules, such as polysaccharides.

Biomolecules are usually divided into two categories. The first type includes small molecules, such as lipids, vitamins or metabolites. The second kind includes large biopolymers, such as nucleic acids and proteins, which are built by covalently bonded monomeric units. The properties of biopolymers are dictated by their particular monomer composition. Strong covalent bonds are key to understand the formation and long-term structure of the different biopolymers. Weaker, noncovalent forces—such as electrostatic, Van der Waals and hydrogen-bond interactions—allow for flexibility and play key roles in numerous biological functions. In fact, biomolecular structures and processes can be understood in terms of an interplay of covalent and noncovalent interactions between biopolymers, as well as with their environments [5]. Water, an essential ingredient for life on earth, mediates interactions by screening electric charges—not only of biomolecules, but also of ions and polar species—and by solvating and stabilizing molecules via hydrogen-bonding. Furthermore, water molecules also have a crucial role in biochemical reactions [7]. In this thesis, we investigate conformational and chemical changes in different biomolecules, which occur in various environments and fulfill specific biological functions. The systems that we study belong to the three main kinds of biopolymers—polypeptides, polynucleotides and polysaccharides—which are described in further detail below (see Fig. 1.1).

1.1.1. Proteins

Proteins are large molecules that perform multiple functions in living organisms, such as responding to external stimuli, catalyzing reactions, replicating genetic information, providing structural support and transport, among others [5]. Each protein is formed by a specific sequence of amino acid residues connected by peptide bonds, i.e. a polypeptide [8]. In some proteins, the polypeptide chain can be cross-linked. The particular sequence of amino acids—which is usually referred to as the protein’s primary structure—is determined by genes [9]. Each amino acid consists of a central carbon atom, $C_\alpha$, an amino group ($-\text{NH}_3^+$, in aqueous solution at neutral pH), a carboxyl group ($-\text{COO}^-$), and a side chain. Remarkably, almost all proteins are built by combinations of 20 encoded amino acids, which have been subject to natural selection in order to maximize the coverage of various properties, while minimizing the energy cost [10]. Indeed, amino-acid side chains vary widely
1.1. Biomolecules: structure and function

Figure 1.1: Examples of the biomolecules studied in this thesis. (A) An oligopeptide, polyproline, with its helical structure shown in red (studied in Chapters 4 and 5). (B) A light-sensing protein, with α-helices and β-sheets shown in purple and yellow, respectively; and a chromophore shown in licorice representation (studied in Chapters 8). (C) A DNA double helix, with colored bases—A in red, T in green, G in blue and C in yellow—showing the specific pairing (studied in Chapters 6 and 7). (D) A polysaccharide, hyaluronan, with its disaccharide unit made of N-acetylglucosamine and glucuronic acid (studied in Chapter 3).

in shape, size, charge, hydrophobicity, hydrogen-bonding and reactivity; thus providing varied building blocks for proteins.

While a polypeptide’s main chain, i.e. the backbone, is somewhat flexible, it is also restricted to some favorable conformations. The peptide bond—formed between the carboxyl group of one residue and the amino group of the next one—is essentially planar. This means that the dihedral angle defined around a peptide bond, \( \omega \), can either be in cis—with both \( C_\alpha \) atoms on the same side of the bond—or in trans form—with both \( C_\alpha \) atoms on opposite sides of the bond (see Fig. 1.2). Most peptide bonds are trans. Two other types of dihedrals determine the local shape of the backbone. A \( \phi \) dihedral describes the rotation around a bond between an amino N atom and a \( C_\alpha \) atom, while a \( \psi \) dihedral describes the rotation around a bond between a \( C_\alpha \) atom and a carbonyl C atom. Due to steric hinderance, not all configurations on the \((\phi, \psi)\)-plane are feasible [11]. This restriction of conformational space is key to allow for protein folding—i.e. the process through which a linear polypeptide acquires a specific 3D structure [12]—and function. Most natural proteins have chains of \( \sim50 \) to \( \sim2000 \) amino acids. Sequences of fewer residues are referred to as peptides, which, like proteins, are also involved in multiple biological functions. In Chapters 4 and 5, we study conformational changes of peptides (see Fig. 1.1).

The primary structure of a protein dictates its three-dimensional form, also known as secondary structure. There are regular structures in which polypeptides fold. For example, α-helices [13] are tight coils—usually right-handed—in which the backbone forms an inner rod held by hydrogen-bonds, with the side chains extending outward. Another common structural motif, β-sheets [14], are formed by β-strands—i.e. sections where side chains of adjacent amino acids point in opposite directions— which are linked by hydrogen bonds (see Fig. 1.1). Adjacent β-strands may run in parallel or antiparallel directions. Additionally, β-turns mark reversals in the direction of a polypeptide chain.
Introduction

Figure 1.2: Protein backbone dihedral angles: $\omega$ describes the rotation around a C-to-N peptide bond between two amino acid residues; $\phi$ describes the rotation around a residue’s N-to-$C_\alpha$ bond; and $\psi$ describes the rotation around a residue’s $C_\alpha$-to-C bond.

The tertiary structure of a protein is determined by the placement of secondary structures—i.e. helices, sheets or turns—stabilized by hydrophobic, electrostatic and hydrogen-bonding interactions, as well as cross-linking via disulfide bonds. The quaternary structure refers to assemblies of several protein subunits—either of the same or of different kind—which together can perform complex functions, such as signal transduction. Moreover, proteins can bind other kinds of molecules to fulfill specific functions. For example, light-sensing proteins—such as the ones studied in Chapter 8—bind molecules that absorb photons within a certain range of wavelengths in order to trigger specific responses (see Fig. 1.1).

1.1.2. Nucleic Acids

Nucleic acids—mainly, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)—are carriers of genetic information, essential for all forms of life on earth [5]. Biochemically, nucleic acids are polymers whose monomer units, nucleotides, contain a sugar, a phosphate and a base. The sugars of the nucleotides are linked by $3'$→$5'$ phosphodiester bridges, forming the backbone of the nucleic acid. DNA and RNA differ in their sugar—deoxyribose and ribose, respectively—and in one of the bases they use. DNA has two purine bases, adenine (A) and guanine (G), and two pyrimidine bases, cytosine (C) and thymine (T). In RNA, T is replaced by uracil (U).

One of the most important aspects of DNA is that it enables not only the storage, but also the replication of genetic information. As discovered in 1953 by Franklin and Gosling [15], and Watson and Crick [16], DNA has a double-helix structure, with two antiparallel polynucleotides coiled around the same axis. The nucleobases are inside of the double helix, and the backbone is outside (see Fig. 1.1). More importantly, the nucleobases of the opposite strands follow a specific pairing via hydrogen bonds—A-T and G-C—which is crucial to enable the replication of a complete DNA double helix from either single strand. In Chapters 6 and 7, we
study DNA base-pairing conformations. In eukaryotic organisms, DNA is packed in linear structures called chromosomes with are mainly stored in the nucleus of the cell. In contrast, prokaryotes store DNA in circular chromosomes in the cytoplasm.

DNA replication is done by several proteins, including polymerase enzymes [17]. During this process, the double helix is unwound and set apart by helicase enzymes. The two separated original strands serve as templates. DNA polymerases synthesize new complementary strands, by catalyzing the formation of phosphodiester bonds with the corresponding nucleotides according to the base-pairing rules. In some viruses, this process is carried out by RNA.

The sequence of bases in DNA, i.e. the genetic code, can be translated to sequences of amino acids in proteins. Indeed, each amino acid is encoded by a group of three bases, known as a codon [9]. Briefly described, a gene can be copied from DNA to messenger RNA (mRNA) in a process called transcription. Then, in a process called translation, mRNA is read by the ribosome, which uses ribosomal RNA (rRNA) to synthesize a new protein utilizing amino acids brought by transfer RNA (tRNA).

1.1.3. Polysaccharides

Polysaccharides—also known as polycarbohydrates—are long chains of sugars joint by glycosidic bonds, which function as energy storage and structural support in various organisms [5]. Distinct polysaccharides are assembled by several enzymes, called glycosyltransferases. Each glycosyltransferase catalyzes the formation of a glycosidic bond between two specific monosaccharides. The wide variety of monosaccharides and possible linkages originates the structural complexity of polycarbohydrates.

In animals, the most common homopolysaccharide is glycogen, which is a large branched polymer of glucose used for energy storage. In plants, this glucose reservoir role is fulfilled by starch, which can be either unbranched (amylose) or branched (amylopectin). Another unbranched glucose polysaccharide found abundantly in plants, cellulose, serves a structural role. The types of glycosidic bonds are the key differences between these polysaccharides. The α linkages if glycogen and starch yield an open helix, which is suitable for an accessible sugar storage. In contrast, the β-linkages of cellulose give a straight chain, which can be used to construct strong fibers.

In animals, a key structural role in the extracellular matrix is fulfilled by glycosaminoglycans (GAGs), which are linear polyelectrolyte chains formed by repeating disaccharide units [5]. The GAG disaccharide unit has one amino sugar and one uronic sugar, and at least one of the two has a negatively charged functional group (see Fig. 1.1). GAGs perform multiple functions in animal organisms, such as regulating inflammatory [18, 19] and immune responses [20]. In Chapter 3, we study the glycosaminoglycan hyaluronan. Interactions between proteins and carbohydrates, which are a rich field of study, are not covered in this thesis.
1.2. Biomolecular dynamics simulations

Modern biomolecular simulations are boosted by several developments, such as enhanced sampling, multiscale modeling, data-based analysis, experimental-data integration, machine learning, and so on [21]. Undoubtedly, the field of biomolecular simulations has come a long way from its beginnings. In 1959, the late B. Alder proposed the first general method for molecular dynamics (MD) simulations [22], and tested it on a Lennard-Jones fluid. A few years later, in the mid and late 1970s, the first simulations of biological systems were performed by, among others, M. Levitt, A. Warshel and M. Karplus [23–25]. Since then, the technique has boosted the study of almost every imaginable biological system. Biomolecular functions can be understood in terms of structures\(^1\), dynamics\(^2\) and free energies\(^3\), all of which can be meticulously extracted from MD simulations (see also Chapter 2).

1.2.1. State of the art

Nowadays, computing facilities and algorithms are able to run enormous systems, for longer times than ever before. This immense computational power is exemplified by milestones, such as the ms-long simulations of fast-folding proteins performed on the Anton supercomputer [26], or the recent billion-atom simulation of an entire gene (GATA4) with a stunning performance of 1 ns/day [27]. However, even these heroic simulations are dwarfed when compared to the actual biological timescales. If we were to simulate the whole process of the GATA4 gene transcription and translation, which takes \(\sim 1\) min [28], the MD run would take 164 million years with our current best performance. And this is not an isolated case. The majority of the relevant biological processes occur on timescales that cannot be reliably sampled with MD simulations. This is referred to as the rare event problem; meaning that the probability of observing such an occurrence during a simulation is vanishingly small. Note that these considerations are done for an affordable force-field level of MD simulations, where interatomic interactions are calculated based on empirical potentials. MD runs based on first principles, e.g. density functional theory, are even more limited by computational cost. Coarse-grained models can access longer timescales and larger systems, but sacrifice atomistic resolution (see Fig. 1.3).

Waiting for increases in computational speed to enable larger and longer simulations has never been an option for the impatient simulation community. In order to overcome the rare event problem, an arsenal of enhanced sampling techniques has been generated [6]. In particular, free-energy methods have become a method of choice to extract insight from molecular simulations. We present these techniques in more detail in Chapter 2. The essence of this class of methods relies on biasing key atomic degrees of freedom—e.g. bond distances or dihedral angles—to

\(^1\)The distinct stable atomic configurations of a biomolecule, which are determined by interatomic interactions, enable specific functions, e.g. a protein might have an open and a close state with distinct binding affinities.

\(^2\)The transitions of a biomolecule between different stable configurations occur via atomistic mechanisms, e.g. there might be one or several pathways through which a protein can open or close.

\(^3\)The feasibility of each configuration and each transition pathway is determined by their energetic and entropic contributions, e.g. a protein might be more stable in the closed conformation and prefer a certain closing mechanism.
favor the occurrence of the desired transition. We refer to these key descriptive degrees of freedom as collective variables (CVs). The genius of free-energy schemes is to bias the CVs with an external potential that can be quantified and corrected for afterward. Then, one can calculate the underlying free-energy surface of the system—projected onto the CVs—which delivers interpretable stable states, transition channels and barriers. These free energies also allow to relate the simulations to experimental measurements. However, as the number of CVs grows, the computational time for converging a free-energy estimation rises exponentially. Thus, our ability to simulate intricate processes, which involve many CVs, is limited.

1.2.2. Future directions
Path-based approaches offer a solution to handle many CVs [30–32]. The idea, in simple terms, is to exert the bias—and calculate the free energy—not on the CVs themselves but on the progress parameter along an adaptive path connecting two known stable states in CV-space. A minimum free-energy path (MFEP) can be found by updating the adaptive path on the fly. Since the progress along the path is a one-dimensional parameter, the convergence of the free energy is freed from the exponential scaling. Our in-house developed path-based approach [33, 34] is further explained in Chapter 2.

Path-based methods have a promising future thanks to their ability to navigate high-dimensional CV-spaces. The advancement of computational power will continue to enable simulations of larger systems, which allow for more complex kinds of transitions. Machine-learning potentials will also enable first-principles accuracy in simulations of larger systems [35, 36], where one might also find elaborate mech-
anisms. Moreover, path-based approaches are well-suited to handle the sets of CVs discovered by novel machine learning and data-based approaches [37–40], as done in [41]. In this thesis, we develop and apply new strategies for path-based free-energy calculations, which we hope will expand their capability to tackle even more complex biomolecular processes hereafter.

1.3. This thesis

If we are to consider that the long-term goal of biomolecular simulation is to compute the atomic motions of a living organism, in all its complexity, for a relevant amount of time, then the purpose of this thesis is to make a small but significant step in that direction. The main objective of this thesis is to develop a robust enhanced simulation framework and apply it to study a variety of complex biomolecular processes. This is accomplished by means of path-based biasing methods, which are uniquely well-suited to navigate the high-dimensional CV-spaces of intricate transitions. Moreover, the path-formalism enables a seamless usage of standard free-energy methods, by simply applying them on the path progress parameter, rather than on the CVs. We extend path-based methods with capabilities used in other enhanced sampling schemes, such as parallel replicas. Furthermore, we also develop novel approaches, such as our scheme to explore multiple pathways simultaneously, as well as the switchings between them. Additionally, we employ path-based free-energy calculations in combination with multiscale modeling. Further developments, not included in this thesis but referred to in the appendices, deal with using machine learning and data-driven approaches to find sets of CVs, which can then be handled in path-based calculations. The flexibility, compatibility and efficiency of our path-based methodology allows us to overcome various challenges arising in the different biomolecular systems that we study: oligopeptides, polysaccharides, nucleic acids and light-sensing proteins. Across all systems, we identify and solve different sampling issues in order to extract valuable insight in the form of free energies, structures and mechanisms. This information can then be compared with, explain, or predict experimental measurements.

Part I of this thesis starts with the current introduction chapter. In Chapter 2, we provide an overview of the methods used in this thesis. Starting from the fundamentals of statistical mechanics, we explain how time-averaged atomic properties measured in an MD simulation can be matched to macroscopic experimental observables. We then dive into how an MD simulation is propagated in time and how the interatomic interactions can be calculated either at the force-field or the density-functional-theory level. Next, to address the rare event problem, we introduce enhanced sampling techniques. Specifically, we cover four well-established biasing methods: umbrella sampling, constrained MD, steered MD and metadynamics. Finally, we present the workhorse of this thesis, path-based free-energy methods; illustrating their theoretical foundation, numerical implementation, parameter options and software details.

Part II of this thesis is concerned with connecting experiments and simulations of polysaccharides. In Chapter 3, we combine 2D infrared spectroscopy, single-chain force spectroscopy and molecular simulations to understand the ion-responsive rigid-
ity of hyaluronan, a natural polyelectrolyte abundantly present in the extracellular matrix. Two-dimensional infrared spectroscopy and MD simulations agree on the formation of a key structure between Ca$^{2+}$ cations and specific hyaluronan side-chains, which weakens the polysaccharide’s intramolecular hydrogen bonds. Such binding does not occur with Na$^+$ cations. Free-energy calculations and force spectroscopy agree that the presence of Ca$^{2+}$ ions induces a $\sim 50\%$ decrease on the persistence length of hyaluronan. Thus, we discover the atomistic mechanism through which hyaluronan, known to regulate cell growth and migration, can modulate its flexibility.

Part III of this thesis is concerned with developing novel path-based schemes, particularly involving parallelism, and testing them in oligopeptides. In Chapter 4, we assess the performance of multiple-walker path-metadynamics by studying the right- to left-handedness transition in polyproline helices of varying length. Polyproline provides an ideal test system, in which adding one more residue to the chain also implies adding one more CV—i.e. one more $\omega$ dihedral—to describe the helix handedness. We study polyproline in tetrameric, pentameric and hexameric form, i.e. using three, four and five CVs. The convergence time for the pathways and free-energy profiles calculated with path-metadynamics scales sub-exponentially, even sub-linearly, with the dimensionality of the CV-space; thus surpassing the limitations of standard methods. Moreover, the use of multiple walkers accelerates the calculations beyond trivial parallelism, because the more uniform sampling along the path also helps in its optimization. Additionally, we learn details about the preferred mechanisms for the helix handedness change, either end-to-end or center-to-ends; thus explaining previous experimental and computational results.

In Chapter 5, we develop and assess novel path-based methods. First, we introduce multi-path-metadynamics, a technique in which multiple repulsive paths are used to simultaneously explore different competing transition mechanisms. The key to the technique is the use of special walkers that induce repulsion between the multiple paths, or pinpoint known distinct intermediates. We test the new technique in the prototypical system alanine dipeptide, with two paths, as well as on tetrameric polyproline, with six paths. The polyproline result compares well to the previous chapter. We also introduce path-based schemes to handle cyclic and switching paths. Moreover, we introduce another new concept: the PathMap. Once the multiple paths between two states have been captured, a PathMap provides an easily-interpretable view of the free-energy along each pathway, and of the switching from one path to another.

In Part IV, we use path-based methods to study a base-pairing transition in DNA. Hoogsteen (HG) base-pairing is an alternative motif to Watson-Crick-Franklin (WCF) in which the purine base rotates 180° around the glycosidic bond. This rotation can occur either inside or outside—i.e. with the purine flipping out—of the double helix. Recently, it has been discovered that WCF and HG base pairs exist in a dynamical equilibrium and that many biological functions require HG base pairs. In Chapter 6, we perform a mechanistic study of the Watson-Crick-Franklin to Hoogsteen transition of a specific A-T pair in a well-studied DNA sequence ($A_6$-DNA). We apply several enhanced sampling methods in combination with the path-
CV; including steered MD, metadynamics, umbrella sampling and constrained MD. This exercise shows the flexibility of the path-CV, as well as the advantages and disadvantages of each technique. The path also allows us to consider more CVs than just the purine base rotation, e.g. the purine base flipping, key hydrogen-bond distances, and distances between the neighbors and the backbone atoms near the transitioning pair. The relevance of the CVs is discussed in the appendix of the chapter. We identify key mechanistic details that distinguish the WCF-to-HG transitions occurring inside and outside of the double helix. Moreover, we find a marked preference for the latter pathway.

In Chapter 7, we exploit the multiple-path framework developed in Chapter 5. The efficiency of the method enables, for the first time, a systematic investigation of transient HG base-pairing in multiple DNA sequences. We study the WCF-to-HG transition in seven different DNA chains, based on variations of the original A6-DNA studied in Chapter 6. Importantly, we find dominant outside pathways for all sequences. We detect a weak influence of the immediate neighboring bases on the transition. Instead, we observe a strong influence of the relative length of the DNA towards each direction of the transitioning base pair. This finding points to a key role of DNA flexibility in modulating the dynamical equilibrium between WCF and HG base pairs, which might be particularly important in protein-DNA complexes.

Part V of this thesis is concerned with combining path-based free-energy methods with multiscale modeling; particularly with quantum mechanics/molecular mechanics. In Chapter 8, we investigate the dark-state recovery mechanism of light-sensing proteins. In particular, we study blue light-using flavin (BLUF) proteins, which are known for their hydrogen-bond network around the flavin chromophore and for their widely-varying signaling times. For example, the recovery rate of the BlrB BLUF protein is 750 times faster than that of the AppA BLUF protein. Our path-based schemes enable us, for the first time, to sample the dynamics of the Gln rotation and tautomerization that is associated with the BLUF dark-state recovery. Moreover, we obtain free-energy profiles for the BlrB and AppA proteins that validate a previously proposed mechanism. Our profiles also point to the key role of a Trp and a Met near the flavin chromophore as modulators of the recovery rate.

Finally, in Part VI, we provide a summary of the thesis, a list of associated publications, and acknowledgements to the people and organizations who supported this research.

References


