A Computational Investigation of Allostery and Binding Interactions in Proteins

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Natural Sciences By Research

by

Akshay Prabhakant
20161074
akshay.prabhakant@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500032, India
October 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**A Computational Investigation of Allostery and Binding Interactions in Proteins**" by **Akshay Prabhakant**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Marimuthu Krishnan

To my parents, friends who have helped along the way, and a brighter future

Acknowledgments

Without further ado, I would like to extend my heartfelt gratitude to my advisor, Professor Marimuthu Krishnan, whose unwavering guidance, boundless expertise, and patient mentorship have been instrumental throughout my research journey.

To my dear mother, for her immeasurable support and caring spirit that provided comfort during the most challenging moments of this endeavour, and to my father, for his enduring encouragement, wisdom, and a never-give-up attitude that shaped my approach to life's challenges.

I am profoundly thankful to my dedicated co-researchers, Abhinandan and Nikhil, who stood alongside me, bravely overcoming the adversities we encountered during the pursuit of both research projects. Your resilience and collaboration have been invaluable.

To my dear friend Yagya, whose constant motivation and unwavering belief in my capabilities served as a driving force, helping me reach the completion of this dissertation.

I would like to extend my gratitude towards the High Performance Computing Centre here at IIIT Hyderabad. No simulations could've been carried out had it not been for their generous provision of the required computational facilities.

Lastly, I extend my appreciation to all those remarkable researchers whose passion for knowledge and relentless pursuit of excellence continue to inspire others, igniting a thirst for learning and improvement in all of us. Your contributions to the world of academia have left an indelible mark.

Abstract

Protein allostery is a fundamental biological mechanism that plays a pivotal role in regulating various cellular processes. It allows proteins to change their shape and function in response to specific signals or ligands. In the context of nucleotide excision repair (NER), proteins involved in the repair process, such as the Rad4/XPC protein, undergo conformational changes when they bind to damaged DNA, which begin the lesion excision process by locating the lesion site, which is followed by the recruitment of other repair factors and facilitates the removal of damaged DNA segments. Essentially, protein allostery ensures that the repair machinery is only activated when and where it is needed, contributing to the precision and efficiency of DNA repair processes in cells.

Molecular dynamics (MD) simulations enable the investigation of the conformational changes that facilitate complex, selective, and biologically significant processes involving proteins binding to a variety of substrates. Hence, the present work uses MD to study two processes that proteins participate in: (a) transcription modulation through mediator allostery and (b) locating UV-lesions present in DNA for their ultimate excision and repair.

F-helices in CAP are able to ultimately recognise and bind to DNA for transcription when they undergo reorientation after cAMP, and a mediator has situated itself into the binding pockets of CAP. The present study uses a simulation-based approach to investigate the mechanism of cAMP-induced changes in the conformation and energetics of F-helices observed during the allosteric regulation of CAP by cAMP. The free energy profiles obtained by two-dimensional umbrella sampling of CAP and cAMP-bound CAP provide a detailed picture of the elasticity modifications observed in the DNA-binding domain of CAP when cAMP is appropriately situated. Residue-wise interaction energy maps w.r.t. CAP residues under the different conformations of CAP, cAMP-bound CAP, and cAMP-bound DNA-complexed CAP are created, which ultimately offer clues on the microscopic origin of the inter-subunit cooperativity and dimer stability of CAP.

UV radiation-induced DNA damage has adverse effects on genome integrity and cellular function. The most prevalent DNA lesion is the cyclobutane pyrimidine dimer (CPD), implicated in a variety of genetic skin-related diseases and cancers in humans. Rad4/XPC is a damage-sensing protein that recognises and helps repair CPD lesions with high affinity. This binding efficiency of Rad4 depends on how efficiently the BHD2 and BHD3 domains have been associated with the CPD-containing lesion site of DNA. The present thesis investigates the mechanism, energetics, dynamics, and molecular basis for this Rad4-DNA association using CPD-containing perfectly matched DNA. This key molecular event that occurs in NER is studied using suitable reaction coordinates, and the resultant free energy surface when compared with the same of TTT/TTT mismatched DNA reveals that Rad4 has a higher tendency to stay in the associated conformation with CPD-containing DNA than TTT/TTT mismatched DNA, hence having a higher lesion-recognition efficiency on the former than the latter.

Contents

List of Figures

List of Tables

Table Page

Chapter 1

Introduction

Contents

At the end of the eighteenth century, the biochemical makeup of the animate world was found to be astoundingly distinct from that of the inanimate world. Antoine-Laurent Lavoisier was one of the first to identify the sophisticated chemical nature of the animate world and distinguish it from the classical makeup of the *mineral world*. He also found out that living matter had an abundance of elements such as oxygen, nitrogen, carbon, and phosphorus.

All known life forms to date are found to be carbon-based. The tetravalency of carbon has a major role to play in this. It can form single bonds with hydrogen atoms and both single and double bonds with oxygen and nitrogen atoms. Moreover, it can form very stable single bonds with up to four other carbon atoms. Almost all biomolecules stem from hydrocarbons, with hydrogen atoms being replaced by a variety of functional groups that impart specific chemical properties to the molecule, forming various families of organic compounds. A broad spectrum of proportions and configurations enables biomolecules to fulfil a variety of functions. The four abundant, functionally important and human-relevant classes of biomolecules are carbohydrates, lipids, nucleic acids, and proteins (Figure 1.1). Carbohydrates are involved in serving as a source of stored energy, provision of energy through respiration; Lipids constitute the cell membrane, thus responsible for selective acceptance and rejection on the entry of foreign bodies/biomolecules, act as energy-rich fuel stores, pigments, and intracellular signals; Nucleic Acids store an organism's genetic code; Proteins are involved in serving as transporters for moving nutrients

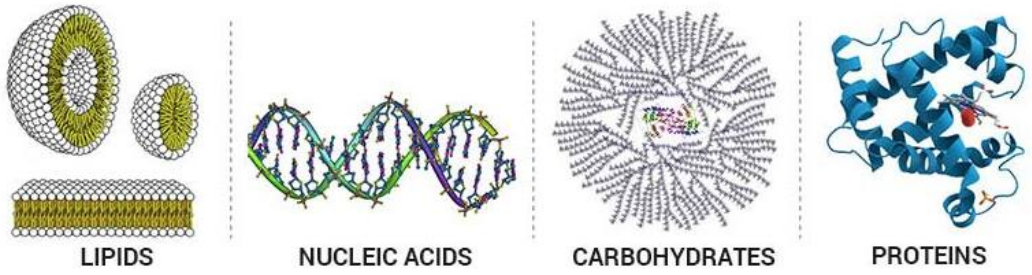


Figure 1.1: **The "Big-4" of biomolecules** Image borrowed from [1].

and other molecules in and out of cells, the formation of antibodies for immunity against foreign bodies, possess catalytic activity. The study of biomolecules, aimed at understanding the behaviour of the cell and the physiological processes that occur, ultimately governing it and thus providing deeper insights into the functioning of living organisms, is known as biochemistry.

In addition to exploring how these molecules function, interact, and undergo reactions, biochemistry delves into the pivotal role of genetic techniques in shedding light on various aspects of this field. This is especially significant since the biochemistry of nucleic acids forms the core of genetics. For instance, biochemical assays reveal how DNA and RNA participate in processes like replication, transcription, and translation; biochemical methodologies are used for various applications such as recombinant DNA technology and gene cloning. Studying living organisms and their immune systems functions has a near-complete coincidence with biochemistry. Techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and mass spectrometry used to determine the three-dimensional structures of proteins belong to biochemistry. Enzyme-catalysed reactions (belonging to the protein biomolecule class) are what break down and digest drugs in living organisms, thus forming a solid foundation in pharmacology and pharmacy. Examination of illnesses like inflammation, cellular damage, and cancer involves the increasingly prevalent use of biochemical methodologies [2].

1.1 Proteins

Proteins have a driver's seat in almost all cellular processes, thus proving their vast coverage of functionalities. By virtue of this coverage, a need to study multiple proteins is created in order to explore the molecular mechanism of a biological process. All proteins are basically a sequence of the same set of 20 most commonly occurring amino acids, linked covalently to each other. These building blocks help a variety of living organisms adapt and survive in their habitat, such as the production of feathers in birds, spider webs, rhinoceros horn, the lens protein of the eye, antibiotics, etc. [3].

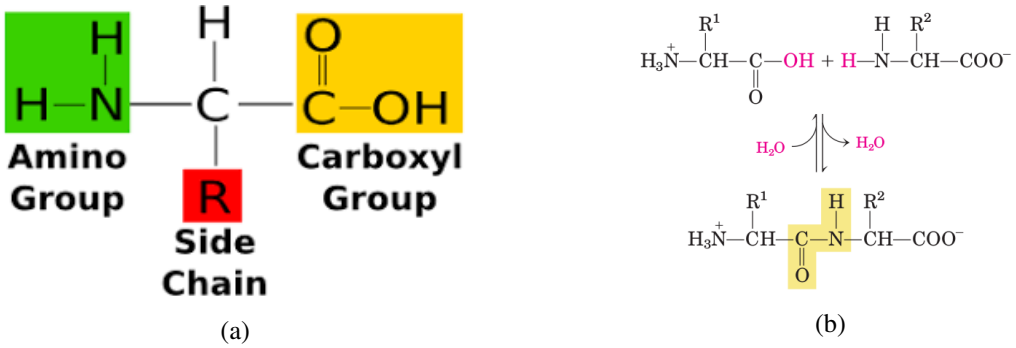


Figure 1.2: **Amino Acids introduction.** (a) A skeletal model of a generalized amino acid showing the amino(green) carboxyl (yellow) and R-groups(red) attached to a central(α)carbon, Image borrowed from [4], (b) Formation of a peptide bond by condensation: The α-amino group of one amino acid (with R^2 group) acts as a nucleophile to displace the hydroxyl group of another amino acid (with R^1 group), forming a peptide bond (shaded in yellow). Image borrowed from [3].

As early as the eighteenth century, Antoine Fourcroy distinguished the ability of proteins to coagulate or flocculate under treatments with heat or acid, marking the beginnings of protein history [5]. This led to the identification of ubiquitous proteins such as albumin, fibrin, gelatin, and gluten. In 1838, Jacob Berzelius proposed the name **protein** to designate *the primitive or principal substance of animal nutrition* [6]. This was followed by the discovery of the peptide bond by Emil Fischer and Franz Hofmeister [7, 8]. In 1934, the first sharp X-ray diffraction pattern for a crystalline protein, pepsin, was obtained by J. D. Bernal and Dorothy Crowfoot Hodgkin, confirming its compact globular shape and further discovering the importance of water for maintaining conformational stability. By 1936, all 20 ubiquitous amino acids had been identified. By 1958, a low-resolution crystal structure for myoglobin, the first folded protein 3D structure, was published [9].

Amino acids form the building blocks of proteins. They are organic compounds that contain amine ($-NH_2$) and carboxyl (-COOH) functional groups, along with a side chain (R group) specific to each amino acid. The most prominent amino acids are the α-amino acids, which have both amino and carboxyl groups attached to the α-carbon of the structure (Figure 1.2a). These side-chain groups vary in structure, size, and electric charge, which influence the solubility of the amino acid in water. All proteins, irrespective of the complexity of the biological life from which they come, are constructed from the same ubiquitous set of 20 amino acids. The individual amino acids are usually represented by either three-letter codes or a single-letter code (Figure 1.3). The side groups (R groups) of amino acids determine their properties. For instance, glycine has the smallest side chain and thus can be accommodated in places inaccessible to other amino acids; it often occurs where peptides bend sharply [2]. During enzymatic catalysis and electron transport in respiring mitochondria, amino acids with charged side chains function in charge relay systems via the formation of salt bridges [2].

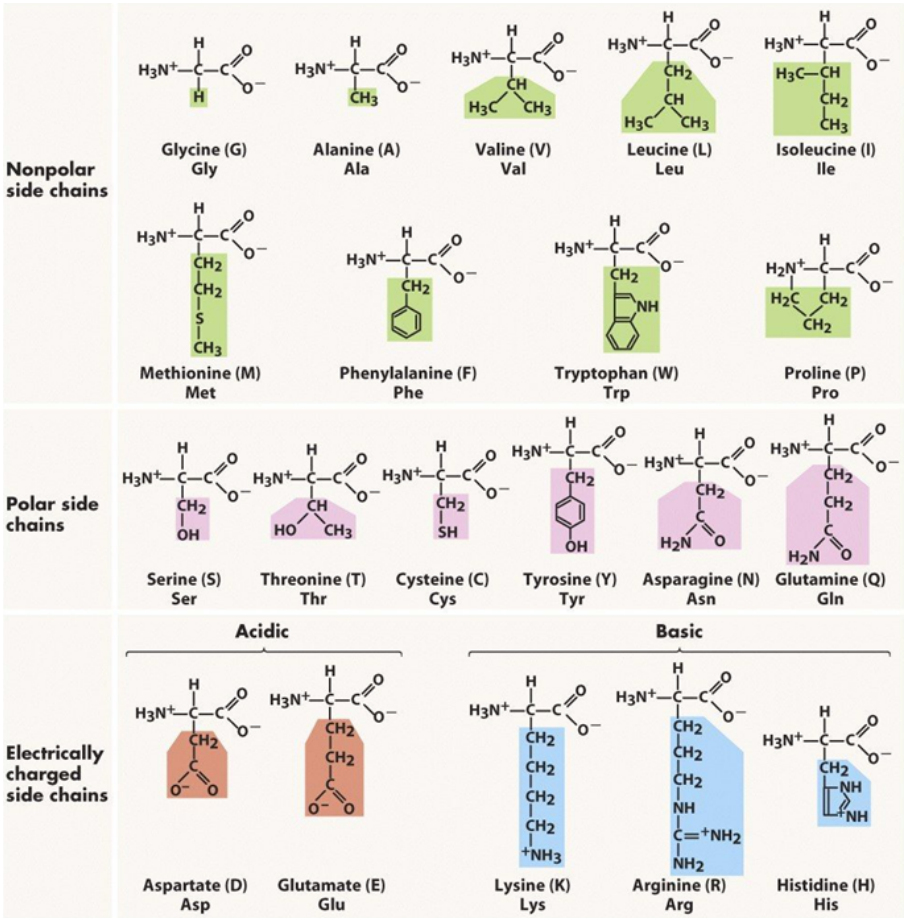


Figure 1.3: **The set of 20 ubiquitous amino acids.**The structural representation of each is displayed, the name, the single letter notation inside the parentheses and the triple-letter notation are depicted. An additional information, regarding the nature of side-chain(R group) of each of the amino acids is mentioned. Acidic denotes amino acids with a net-negatively charged side-chain, basic denotes amino acids with a net-positively charged side-chain.

A molecule known as a dipeptide results from the condensation reaction of two participating amino acids. This reaction involves the loss of a hydroxyl ion from the carboxyl group (COOH) of the first amino acid and a proton from the amino group (NH_2) of the second, releasing a water molecule (H_2O) and leading to a peptide bond (-CO-NH-) formation Figure 1.2b). Within this dipeptide, the amino-terminal(a.k.a. the N-terminal residue) is the amino acid with a free α-amino group; the residue at the other end having a free carboxyl group is the carboxyl-terminal (C-terminal) residue.

Polymeric forms of peptides include tripeptides (formed when three amino acids are connected by two peptide bonds), tetrapeptides(linking of four amino acids), pentapeptides (five amino acid linkage), and so forth. When a few amino acids are joined in this fashion, the structure is called an oligopeptide. When many amino acids are joined together, the product is called a polypeptide. Although the terms

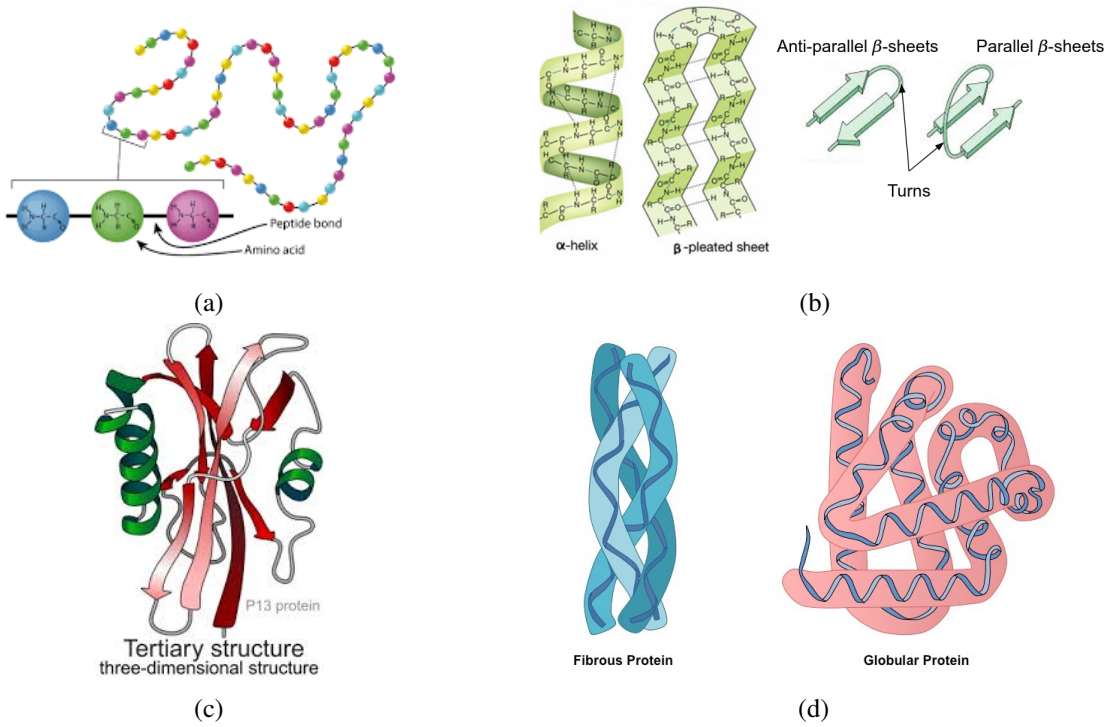


Figure 1.4: **Levels of structure in proteins.** (a) The primary structure consists of a sequence of amino acids linked together by peptide bonds and includes any disulfide bonds. (b) The resulting polypeptide can be arranged into units of secondary structure, such as an α-helix or a β- sheet. These 2 major secondary structures are joined by another secondary structural unit called turns. (c) The helix/sheet is a part of the tertiary structure of the folded polypeptide, which is itself one of the subunits that make up the quaternary structure of the protein. (d) When the protein comprises of multiple units that assemble together, its quaternary structure is formed. In some proteins, pairs of very long α-helices are interwound in a left-handed sense to form two-chain coiled coils. A typical example is α-keratin, found in hair, which is also an example of a fibrous protein. An example of globular proteins is haemoglobin. All images are borrowed from [3].

protein and polypeptide are at times used synonymously, polypeptides generally have molecular weights below 10,000 Da and proteins have higher molecular weights [3]. Amino acid sequences of proteins are read from left to right, i.e., from the N-terminal to the C-terminal.

The simplistic way of representing a protein as a sequence of amino acids does not relay the relevant functional information of a protein completely. The shape of a protein plays a very important role in determining its function. The four levels of protein structure: primary, secondary, tertiary, and quaternary, need to be understood in order to aptly know how a protein obtains its final conformation (Figure 1.4).

A description of all covalent bonds (mainly peptide bonds and disulfide bonds) linking amino acid residues in a polypeptide chain constitutes its primary structure (Figure 1.4a). The most important element of primary structure is the sequence of amino acid residues.

Particularly stable arrangements of amino acid residues that form recurring structural patterns are referred to as the secondary structure of a protein. These structural patterns emerge as the spatial relationship between the constituent amino acid residues of a protein. The three basic units of secondary structure are the α-helix, β-strand, and turns. The distinction between residue sequence and secondary structure came into play when the crystal structure of proteins started to be resolved. Alanine, Leucine and Glutamine are found more frequently in α-helices whilst Proline, Glycine and Aspargine were found less frequently than average. Using this analysis of the primary sequence, a helix propensity scale was derived, which is still used in predicting the occurrence of helices and sheets in folded soluble proteins [10]. Usually right-handed in nature, the α-helix is the most common structural motif found in proteins; in globular proteins, over 30% of all residues are found in helices. The regular α-helix has 3.6 residues per turn, with each residue offset from the preceding residue by 0.15 nm. This helical arrangement of amino acids exists due to the formation of hydrogen bonds between the backbone atoms. The hydrogen bonds occur between the backbone carbonyl oxygen (acceptor) of one residue and the amide hydrogen (donor) of a residue four positions ahead in the polypeptide chain. The arrangement of hydrogen bonds shows variation in length and angle with respect to helix axes for various kinds of proteins. The β-sheet was the second type of secondary protein structure identified by the model-building studies of Pauling and Corey [11]. The backbone of the polypeptide chain is extended into a zigzag rather than a helical structure and, in turn, is arranged side by side to form a structure resembling a series of pleats. Hydrogen bonds are formed between adjacent segments of the polypeptide chain. A single β-strand is not stable, largely because of the limited number of local stabilising interactions. However, when two or more β-strands form additional hydrogen bonding interactions, a stable sheet-like arrangement is created. These adjacent polypeptide chains in a β-sheet can be either parallel or antiparallel, i.e., have the same or opposite N-terminal to C-terminal orientations, respectively (Figure 1.4b). These β-sheets result in significant increases in overall stability and are stabilized by the formation of backbone hydrogen bonds between adjacent strands that may involve residues widely separated in the primary sequence. Turns refer to short segments of amino acids that join two units of secondary structure, such as two adjacent strands of an antiparallel β-sheet. In some proteins, the proportion of residues found in turns can exceed 30%, and in view of this high value, it is unlikely that turns represent random structures [12]. The polypeptide sequence is able to alter its direction, all thanks to the existence of turns. The reverse turns or bends arise from the geometric properties associated with these elements of protein structure [12].

Tertiary structure describes all aspects of the three-dimensional folding of a polypeptide (Figure 1.4c). While the secondary structure constitutes the spatial arrangement of adjacent segments of amino acid residues in a polypeptide, the tertiary structure depicts even longer-range facets of the amino acid sequence. Interacting segments of polypeptide chains are held in their characteristic tertiary positions by several kinds of weak interactions (and sometimes by covalent bonds such as disulfide crosslinks) between the segments [3]. The tertiary structure indicates the manner in which multiple

secondary structures assemble to form domains and the way in which these domains relate spatially to one another. A domain is a section of protein structure sufficient to perform a particular chemical or physical task, such as binding to a substrate or other ligand.

When a protein has two or more polypeptide subunits, their arrangement in space is referred to as quaternary structure [3]. A number of tertiary structures may fold into a quaternary structure (Figure 1.4d). In considering these higher levels of structure, it is useful to classify proteins into two major groups: globular proteins, which have their polypeptide chains folded into a spherical or globular shape, and fibrous proteins, which have their polypeptide chains arranged in long strands or sheets [12].

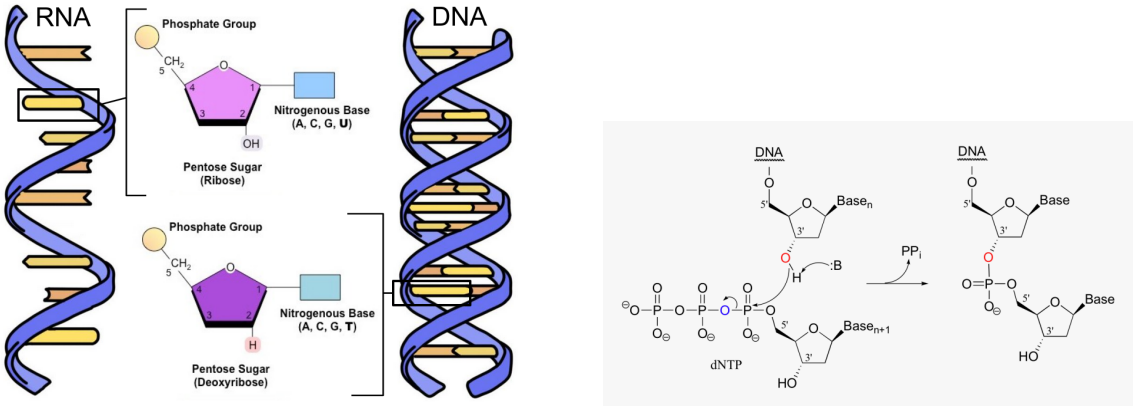
Proteins are complex macromolecules tasked with carrying out critical intra- and extracellular processes. The cytoskeleton of a cell comprises a dense protein network and is responsible for maintaining its shape and structural integrity. The scaffolding of elastic movement in muscles is possible due to Actin and myosin filaments. Haemoglobin transports oxygen, while the circulating antibodies detect foreign invaders. Enzymes catalyse reactions that generate energy, synthesise and degrade biomolecules, replicate and transcribe genes, process mRNAs, etc. Cells have the ability to detect environmental changes and act accordingly thanks to the secretion of hormones, which are then detected by receptors situated on these target cells. Proteins undergo both structural and functional alterations that mimic the developmental stages of the organisms they belong to [2]. Conformational changes may very well be allosterically regulated in nature, at times, be responsible for activating or deactivating the protein. For instance, binding of cylic adenosine monophosphate (cAMP) to Catabolite Activator Protein (CAP) brings about the rotation of the α-helices of the latter, which enables a promoter-DNA sequence to bind to it via docking on these parallel helices.

The classical principles of genetics were deduced by Gregor Mendel in 1865 on the basis of the results of breeding experiments with peas [13]. Early 20th-century genetic studies focused on the identification and chromosomal localization of genes that control readily observable characteristics, such as the eye colour of the fruit fly *Drosophila*.

The early 20th century saw the discovery of a strong connection between enzymes and genes. This was further reinforced when phenylketonuria, a hereditary disease, was discovered to stem from a genetic defect in the metabolism of phenylalanine, one of the 20 ubiquitous amino acids [14]. This defect was hypothesised to result from a deficiency in the enzyme needed to catalyse the relevant metabolic reaction, leading to the general suggestion that genes specify the synthesis of enzymes. From the 1941 experiments of George Beadle and Edward Tatum, the **one gene-one enzyme** hypothesis was concluded (each gene specified the structure of a single enzyme) [15]. During the early days of genetic research, there was a prevailing belief that genes found within chromosomes were composed of proteins alongside DNA. These proteins were believed to pass on genetic makeup from a parent to its offspring. The prevailing notion among most researchers stemmed from the fact that proteins were known as macromolecules characterised by extensive diversity and specific functionality. An inadequate knowledge of

nucleic acids in this part of the century led to the assumption that all nucleic acids shared similar properties. This invariance led the community to believe that nucleic acids could not possibly possess the wide coverage of properties that a macromolecule that is a candidate for genetic material should have. Another reason for this thought was the fact that the 20-amino-acid alphabet of proteins could potentially be configured into more unique information-carrying structures than the four-letter alphabet of DNA. It was thought that perhaps DNA acted as structural support for the chromosomes.

The actual clarification came during the First World War, when Frederick Griffith began studying *Streptococcus pneumoniae* in an attempt to develop a vaccine against it, since at the time a lot of servicemen died due to a lung infection caused by the same. Through his *transformation* experiment, he discovered that the R-strain of this bacteria (not pathogenic) when injected into healthy mice, transformed into an S-strain (pathogenic), thereby causing an infection in the host [16]. This was followed by the experiments of Oswald Avery, Colin MacLeod, and Maclyn McCarty, which later on, along with other studies of the activity of DNA in bacterial transformation, led to the acceptance of the idea that DNA is the genetic material.



(a) (b)

Figure 1.5: **Formation and differences of nucleic acids.** (a) Schematic representational comparison of the basic unit, nucleotide, in an RNA against a DNA. Two key differences are to be noted. Firstly, the entire sugar molecule is same, except for a hydroxyl group which is absent in the nucleotide of DNA, but present in that of RNA. Secondly, Uracil(U) cannot be a nitrogenous base in the nucleotide of a DNA molecule, and Thymine(T) cannot be a nitrogenous base in the nucleotide of an RNA molecule. Image source [17, 18] (b) In DNA and RNA, the phosphodiester bond is the linkage between the 3' carbon atom of one sugar molecule and the 5' carbon atom of another, deoxyribose in DNA and ribose in RNA. The above example, although shown only for DNA, is generalizable for RNA as well. Image source [19]

1.2 Nucleic Acids

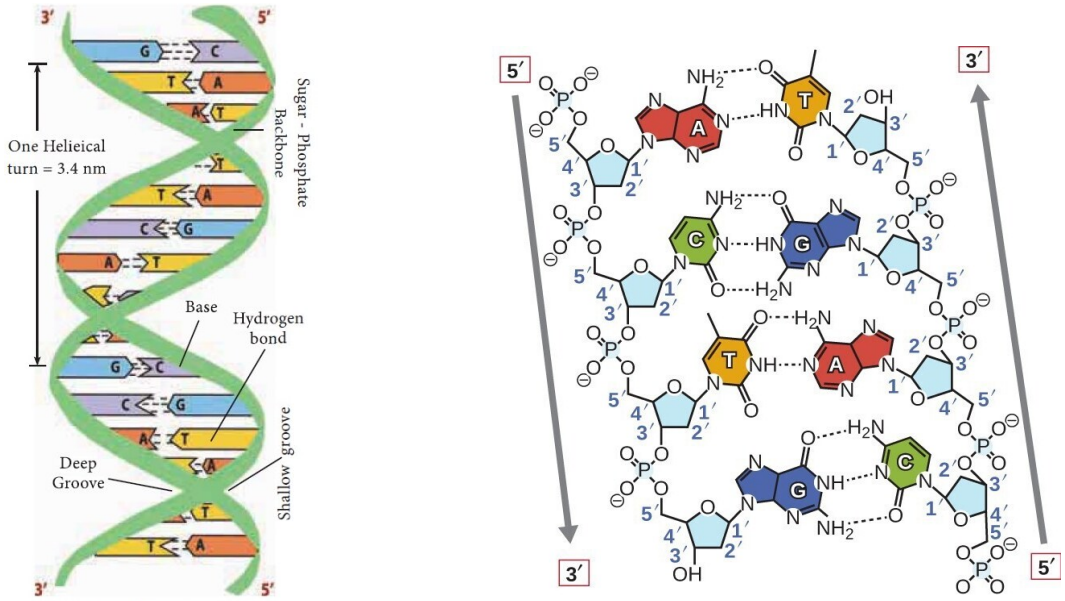
Found in abundance in all living species, nucleic acids are large biomolecules essential to all known forms of life. The two major types of nucleic acids are RNA (RiboNucleic Acid) and DNA (Deoxy-ribo Nucleic Acid). Nucleic acids are responsible for creating, encoding, and storing information in every living cell of each living organism. They are also responsible for passing on this stored information to the biological successor. By directing the process of protein synthesis, they determine the inherited characteristics of every living being. They are composed of a series of nearly identical building blocks called *nucleotides*.

A nucleotide is made of three components: a pentose (5-carbon) sugar, a phosphate group, and a nitrogenous base. When the sugar component is ribose, the resulting polymer is called RNA. If the sugar is deoxyribose (a derivative of ribose), the polymer formed is DNA (Figure 1.5a). The five most prominent nitrogenous bases found across various RNA/DNA are: adenine(A), thymine(T), guanine(G), cytosine(C) and uracil(U). Among these, A,T,G, and C are found in the nucleotides that build up to form DNA, whereas A,U,G, and C are the ones found in the nucleotides of RNA. A polymeric structure is formed when successive nucleotides are covalently linked through phosphodiester linkages (Figure 1.5b) [20]. All nucleic acids are represented as a sequence of single-letter representations of their constituent nitrogenous bases. One end of this chain is labelled as 5' and the other as 3' because of the order in which the formation of this nucleic acid chain has occurred, as shown in Figure 1.5b, wherein the last base that will be added will have a free 3' end, and the top-most base, through which the sequence formation had begun, would have its 5' as the free end.

DNA is considered to be the molecular reservoir of genetic information. The arrangement of every biomolecule and cellular element is determined by the information encoded in a cell's DNA sequence. Each nucleotide of DNA consists of either one of the following four distinct nucleobases: adenine (A), cytosine (C), guanine (G), and thymine (T). A combination of the linear sequence of these bases along with their three-dimensional orientation is widely understood to be responsible for preserving the instructions crucial to the functioning and reproduction of the cell. Such an important of a biomolecule is DNA that its discovery should not go unnoticed.

Soon after Griffiths's **transformation principle** experiment, the scientific community slowly began taking an interest in and exploring more about nucleic acids. The team of Oswald Avery, Colin MacLeod, and Maclyn McCarty around 1940, began experiments on a very similar trajectory. They began to isolate the substance responsible for this transformation. Their work finally concluded in 1944 with the discovery that DNA was the molecule responsible for this transformation and that DNA was the genetic material [21]. In 1952, experiments by Alfred Hershey and Martha Chase on the T2 virus further supported the findings from this work. They found that DNA carried the instructions to make new viruses, which were passed on to subsequent generations because, on radioactive labelling of DNA,

the subsequent generation of viruses were also radioactive in nature, as opposed to radioactive labelling of proteins, which produced non-radioactive viruses [22]. At this point, the functional importance of DNA was widely understood by the scientific community, and parallel structural studies were also being carried out.



(a) (b)

Figure 1.6: **Watson-Crick double-helix DNA model.** Nucleotides are arranged forming a B-DNA structure. Adenine(A) is paired with Thymine(T) and Gaunine(G) is paired with Cytosine(C). The Deep and Shallow grooves are usually referred to as the Major and Minor grooves. (a) DNA double helix, shown in the B-DNA form. Image source [23] (b) The 2-strands of DNA run anti-parallel to each other for proper alignment of their 3' and 5' ends. Image source [24]

Erwin Chargaff, in the year 1950, proposed a rule regarding the nucleobase composition of DNA, which stated that the number of adenine (A) and thymine (T) units matched each other, as did the number of guanine (G) with those of cytosine (C). This would later be known as the *tetranucleotide hypothesis*. He also determined that the DNA composition differs among species, resulting in varying quantities of each base. This observation strengthened DNA's credibility as the genetic material compared to proteins [25, 26]. Other researchers had made important but seemingly unconnected findings about the composition of DNA; Alexander Todd had observed a recurring pattern of phosphate and deoxyribose sugar groups on the backbone of the DNA molecule, and high-resolution X-ray images of DNA fibres obtained by Maurice Wilkins and Rosalind Franklin advocated for a helical, corkscrew-like shape of DNA [27–29]. It was finally in 1953 that, through unifying these disparate findings into a coherent theory of genetic transfer, two researchers by the names of James Watson and Francis Crick proposed the double-helix (twisted ladder) structure of the DNA [30].

The primary features of the Watson-Crick double-helix model are as follows(ref. Figure 1.6):

(i) The DNA molecule consists of two polynucleotide chains or strands that spirally twist around each other and coil around a common axis to form a right-handed double helix.

(ii) The two strands are antiparallel, i.e., they run in opposite directions so that the 3' end of one chain faces the 5' end of the other.

(iii) The sugar-phosphate backbones remain on the outside, while the core of the helix contains the purine and pyrimidine bases.

(iv) The two strands are held together by hydrogen bonds between the purine and pyrimidine bases of the opposite strands.

(v) Most DNA double-helices are right-handed. B-DNA is the most common right-handed form. DNA may exist in a different form, the A-DNA form, a structure shorter and wider than the B-DNA but with a narrower tendency to exist. Only one type of DNA, called Z-DNA, is left-handed.

(vi) Adenine (A) always pairs up with thymine (T) by two hydrogen bonds, and guanine (G) always pairs up with cytosine (C) by three hydrogen bonds. This complementarity is known as the base pairing rule. Thus, the two stands are complementary to one another.

(vii) The base sequence along a polynucleotide chain is variable, and a specific sequence of bases carries the genetic information.

(viii) The base compositions of DNA obey Chargaff's rules.

(ix) The diameter of DNA is 2nm. Adjacent bases are separated by 0.34 nm along the helical axis of the DNA. The length of a complete turn of helix is 3.4 nm, i.e., there are 10 base pairs per turn.

(x) The DNA helix has a shallow groove called the minor groove ($\approx$1.2nm) and a deep groove called the major groove ($\approx$2.2nm) across.

Chromatin is a compact structure composed of both DNA and proteins, primarily located within the cell nucleus and mitochondria. Across individuals of the same species, each cell contains an equivalent quantity of DNA (6 picograms in the case of humans). Notably, both female and male gametes possess half of this DNA quantity. Chromatin is formed when DNA binds with proteins known as histones in the case of eukaryotes. This binding is visualised as the DNA molecule wrapping around a central core of eight histone units at regular intervals, forming a double helix structure. In contrast, chromatin in bacteria, plasmids, mitochondria, and chloroplasts consists of circular DNA.

RNA differs from DNA in primarily three aspects: 1. It has a single chain instead of two intertwined strands; 2. It contains ribose sugar instead of deoxyribose; and 3. It contains uracil instead of thymine.

Different types of RNA exist in cells, including messenger RNA (mRNA), which transmits genetic information from the cell nucleus to the cytoplasm; transfer RNA (tRNA), which carries amino acids to the site of protein synthesis; and ribosomal RNA (rRNA), which has a large (60S) and a minor particle (40S) and is involved in amino acid assembling during protein synthesis. Other types of RNA include small nuclear, small cytosolic RNA, microRNA, small silencing RNA, and long noncoding RNA.

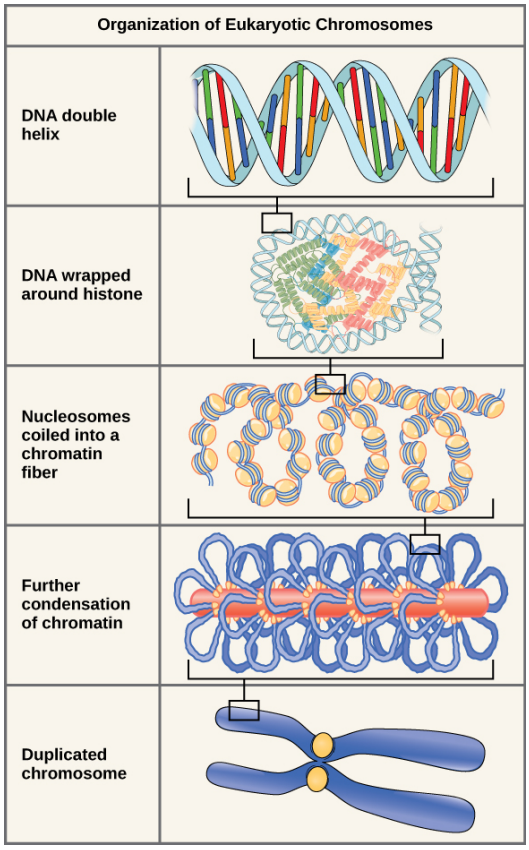


Figure 1.7: **Chromosomal arrangement of DNA inside a cell.** Chromosomal DNA is intricately organized within tiny nuclei through the assistance of histones. These are proteins with a positive charge that bind strongly to the negatively-charged DNA, creating structures known as nucleosomes. Each nucleosome consists of DNA coiled approximately 1.65 times around a core of eight histone proteins. As these nucleosomes come together, they fold into a 30-nanometer chromatin fiber, which, in turn, forms loops spanning around 300 nanometers on average. These fibers are further compacted and folded to generate a narrow 250-nanometer fiber, which is tightly coiled into the chromatid of a chromosome. Image source [31]

In eukaryotes, the DNA is found inside the nucleus of the cell. The length of this nuclear DNA is far greater than the size of the region that houses it. Thus, the DNA has to be condensed. In humans, the packing ratio of this DNA, i.e., the degree to which DNA is condensed, is estimated to be around

7000. To achieve such large values, the packing of DNA has to be distributed across several hierarchies of organisations (Figure 1.7) [32, 33].

The first stage of DNA condensation involves wrapping the double helix around an octa-core histone protein at regular intervals to produce a bead-like structure known as a nucleosome [34, 35]. The part of DNA connecting the nucleosomes is called linker DNA. This gives a packing ratio of about 6, wherein the beads are about 10 nm in diameter, in contrast with the 2-nm diameter of a DNA double helix. The second level of condensation is the coiling of beads in a helical structure 30 nm in diameter called the chromatin fiber. This structure increases the packing ratio to about 40. The final packaging occurs when the fibre is organised in loops, scaffolds, and domains that yield the appropriate final packing ratio. A variety of fibrous proteins are used to pack the chromatin, and they ensure that no two chromosomes occupy overlapping areas of the nucleoplasm.

Defects in the compression of the chromatin structure can lead to several diseases. Thus, chromatin has to be organised in an error-free manner for the easy occurrence of gene expression or for the ease of accessibility of hereditary information residing in the DNA.

For each organism, this genetic information has to be passed to the two daughter cells created when each cell of this organism undergoes cell division. This necessitates the replication of each of the many molecules that form the cell, thus all DNA molecules [37–39]. Each parental DNA strand serves as a template for the synthesis of a new complementary daughter strand, thus making the replication process *semi-conservative* [40].

The preparatory proteins for DNA replication undergo a series of complexations and decomplexations, finally resulting in the activation of the Mcm helicase bound to the parent double-helix DNA, which causes the latter to unwind. This results in the formation of a structure called a *replication fork* [38]. For the part of the template DNA just ahead of the replication fork and existing in an unwound state, its strands provide enormous torsional resistance. To release this, proteins called topoisomerases temporarily break this DNA region by adding negative supercoils to the DNA helix. These two unwound strands serve as the template for the formation of the two complementary strands, the *leading* and *lagging* strands. Activation of the helicase is followed by docking of α-primase and other DNA polymerases onto either of the template DNA strands [41]. DNA polymerases are a family of enzymes that carry out DNA replication but can only extend an existing DNA or RNA strand paired with a template strand [42, 43]. Thus, a short sequence, called a *primer*, must be created and paired with the template DNA strand (for each of the 2 template strands).

After the template strands are separated, the α-primase makes an RNA primer (a short stretch of nucleic acid complementary to some template) for each of the template strands and adds them next to their respective neighbouring (to be extended) strands [44]. DNA polymerase then extends this primer along its 3' end by joining new nucleotides matched with the sequence of the template strand via phosphodiester linkages. Hence, the DNA polymerase can build up DNA only in the 5'-3' direction. Therefore,

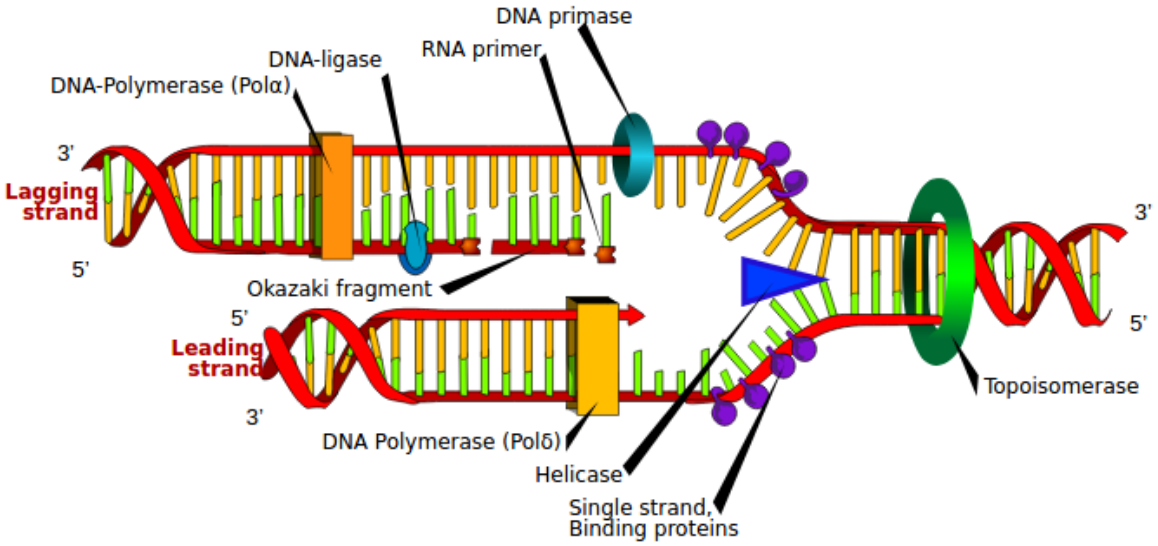


Figure 1.8: **DNA-Replication.** Helicase(blue-triangle) creates the replication fork, by unwinding the DNA-double helical strands at target site. Topoisomerase(green ring) is responsible for stabilising the part of these unwound strands, ahead of the replication fork. Single-strand binding proteins(purple ball and stick) prevent all sorts of internal binding within the nucleobases of the unwound strands by themselves binding at the outer region. DNA-primase(blue ring) generates an RNA-primer as an inception for replication of both of these unwound template strands. DNA-Polymerase(Pol-δ) locomotes in the 3'→5' direction, detecting the nucleobase at each nucleotide encountered and generating a complementary nucleotide, thus building up the leading strand in a 5'→3' direction. For the lagging strand however, DNA-Polymerase(Pol α) attaches itself at the neighboring site of this RNA-primer, travelling in the 3'→5' direction opposite to that of the lagging-template strand and detecting the nucleobase at each nucleotide encountered, generating a complementary nucleotide, thus building up the lagging strand in a 5'→3' direction. This pattern of RNA-primer generation followed by Pol-α locomoting in the opposite direction is repeated several times, throughout the length of the template-DNA strand, whereas for the leading strand this has to be done only once. These DNA-nucleotides fragments thus generated at the lagging strand are called as Okazaki fragments(maroon strand). RNA-primers are then removed, thus creating a nick, which is then filled up by DNA-ligase. Image source [36]

one of the two complementary strands that gets built along the same direction is termed the leading strand, while the other is referred to as the lagging strand (built along 3'-5').

The leading strand is built in a continuous fashion via the addition of nucleotides complementary to its template strand, one at a time. The lagging strand, on the other hand, is made in fragments because, as the fork moves forward, the DNA polymerase attached to the lagging strand must come off and reattach to the newly exposed DNA template. The small fragments are called Okazaki fragments. Hence, the lagging strand needs a new RNA primer for each of the short Okazaki fragments [45]. The RNA primers are removed and replaced by DNA through the activity of DNA polymerase I. DNA ligase is responsible for connecting the lagging strand with these smaller DNA strands generated by DNA polymerase I.

After DNA replication, it needs to be wound around histones and undergo two more successive compressions to form the replicated chromosomes. The replicated DNA double-helix must be coiled around histones at the same places as the original DNA. To ensure this, histone chaperones disassemble the chromatin before it is replicated and replace the histones in the correct place. Some steps in this reassembly are somewhat speculative.

DNA must be replicated with high fidelity. The hydrogen bonds linking two complementary bases make a significant contribution to the fidelity of DNA replication. However, DNA polymerases replicate DNA more faithfully than these interactions alone can account for. Proofreading mechanisms are required to ensure that the accuracy of replication is compatible with the low frequency of errors that is needed for cell reproduction.

As seen in replication, DNA exists in a complexed state with a variety of proteins. Proofreading mechanisms to ensure its accuracy also require complexation with proteins. DNA tends to get damaged, either due to internal or external factors, at certain regions throughout its length. To rectify this damage, a process called NER (nucleotide excision repair) occurs to repair the damaged portions of the DNA sequence. Appropriate proteins are involved in carrying out NER by complexing with DNA. Hence, the formation of protein-DNA complexes is a key player in many of the physiological processes involving DNA [46–48]. It is therefore extremely important to examine the nature of complexes that are formed between proteins and DNA, as they form the basis of our understanding of how these processes take place.

1.3 Protein-DNA complexes

The class of proteins involved in binding with a DNA molecule in order to carry out a physiological process is called *DNA-binding proteins*. These possess certain regions called *DNA-binding domains* that actually form interactions with either specific or general single- or double-stranded DNA. Interaction of these proteins with the major groove of the DNA is facilitated by: (a) a significantly lower electronegative potential and (b) an increased exposition of functional groups that identify a base pair

than the interactions with the minor groove of the DNA. Proteins that modulate transcription (known as transcription factors), polymerases that help in elongation of DNA, helicases and ligases that modify the DNA during the process of DNA replication, histones involved in packing the DNA to the chromatin level are some of the well-known examples of DNA-binding proteins.

Protein-DNA binding is not only limited to double-helix DNA but can also occur with a single strand of DNA. Replication protein-A is a classic example since it is adept at binding to single-stranded DNA [49]. It's specifically tasked with preventing the single-stranded DNA obtained during the creation of the replication fork during the process of DNA replication from forming any stem-loops or degradation by nucleases. The sequence of the DNA involved in protein-DNA binding can also strengthen or weaken the binding. For instance, transcription factors bind specifically to a set of DNA sequences, thus activating or inhibiting the transcription of genes that have these sequences near their promoters. On the other hand, some proteins bind in a sequence non-specific manner. For instance, structural proteins such as histones are prominent examples of proteins that bind non-specifically to DNA. Interactions between the histone-octacore and the region of DNA strand wound around it comprise mainly the ionic bonds between the basic residues in the histones and the acidic sugar-phosphate backbone of the DNA and are therefore largely independent of the base sequence [50].

One of the important transcription factors is Catabolite Activator Protein (also known as CAP). It goes by such a name since it involves the transcription of genes involved in many catabolic pathways. For instance, in bacterium such as Escherichia Coli, when the amount of glucose transported into the cell is low, it unusually modifies ATP into a cyclic molecular loop called cyclic AMP (adenosine monophosphate). This increase in cAMP levels is sensed by CAP, which binds to the former, resulting in an allosteric conformational modification wherein its *recognition* α-helices, a characteristic *helix-turn-helix* motif, undergo a major rotational and slight translational conformational change. This is followed by an increased affinity for binding to the DNA, leading to the formation of the CAP-cAMP-DNA complex.

This is required as CAP then coaxes RNA-polymerase into place, an enzyme required for the occurrence of transcription. By binding to the α-subunit of RNA Polymerase (RNAP), CAP triggers the onset of transcription. This binding enables the generation of the RNAP-promoter closed complex and a further conformational change of this complex to the open state. This protein-protein engagement induces a distinctive bending of the DNA in proximity to the transcription initiation site. This phenomenon plays a pivotal role in catalysing the transcription process of genes linked to lactose catabolism [51].

DNA-Protein complexes may also be formed as a means of rectifying any lesions that may have crept into the DNA sequence. Proteins such as XPC (xeroderma pigmentosum C) and RAD4 (radiation-4) via DNA-binding are responsible for locating and removing any damaged nucleobase or nucleotide entirely. XPC primarily deals with lesions generated from UV-exposure, such as pyrimidine-dimers. The auxiliary proteins UV-DDB bind to the lesions, thus helping XPC localise the lesion. It subsequently

recruits TFIIH, whose XPD helicase verifies the lesion [52–55]. Excision of the damaged nucleotide(s) is carried out by endonucleases, followed by repair synthesis and nick sealing by DNA ligases.

Generally speaking, each interface between a protein and the DNA sequence it is bound to comprises multiple forms of interaction. Some of these may be highly specific contacts, such as hydrogen bonds between the protein and the DNA nucleobases, existing due to a different set of non-specific interactions between elements on the protein and the DNA backbone that orient the DNA-recognition domains the right way. Protein-DNA interactions can lead to conformational as well as functional changes in both biomolecules involved.

1.4 Research Focus

1.4.1 Allosteric Response of DNA Recognition Helices of Catabolite Activator Protein to cAMP and DNA Binding

Transcription factors (TF) are molecular entities that instigate the process of transcription and thus have a vital role in gene expression. Catabolite Activator Protein (CAP) is one such TF, found in a variety of species ranging from the microscopic Escherichia Coli to humans, is known to regulate the metabolism of different organic macromolecules. E.Coli CAP has been extensively covered in numerous studies for more than half a century now. Its TF abilities are regulated by a small effector molecule known as cyclic adenosine mono-phosphate (cAMP) [56–61]. This molecule has been known to bind with CAP, thus bringing it from its *inactive* state, wherein its binding with DNA is practically non-existent, to an *active* state, wherein its binding with DNA is energetically favourable [62–68].

This allosteric shift in conformation brought about by cAMP-binding is manifested in some major changes, such as the coil-to-helix transistion of the stretch VAL126 to PHE136, and a translational and rotational change of 7 Å and 60° respectively , observed in recognition (F-helices) helices, the entity in CAP responsible for binding with the incoming DNA molecule [69]. CAP is a ≈50 kDa dimer consisting of two identical 209 residue subunits, each of which is composed of two distinct domains: (i) an N-terminal cAMP-binding domain (CBD, residues 1 - 136) and (ii) a C-terminal DNA-binding domain (DBD, residues 138 - 209), which contains a helix-turn-helix motif for binding to DNA [69–75]. A short hinge region (residues 137-138) links these two domains. The CBD contains a cAMP-binding pocket region that seats the incoming cAMP molecule, and since this protein is dimeric, a total of 2 cAMP molecules can be accommodated, 1 per monomeric subunit.

The presence of two cAMP-binding sites calls for exploring the type of cooperativity that exists between them. Kalodimos et al. in 2006 characterised the negatively cooperative binding of cAMP to CAP and explained the allosteric regulation being mediated exclusively by the changes in protein motions rather than changes in the intra-protein bonding interactions [76]. The ambiguity in the dynamics

of the mechanism by which these structural changes are brought about by cAMP-binding, especially in the key region of the F-helices, is what propelled the study presented in the Chapter-3 of this thesis.

Chapter 3 in this thesis aims at unravelling the ligand binding effects on CAP via the use of MD simulations and novel reaction-coordinates fuelled umbrella sampling technique. Particularly since the F-helices are observed to have the most pronounced change on cAMP binding, the study will be primarily focused on their behaviour. Additionally, interactions of all residues (including those of the F-helices) with differing environments were studied. Lastly, some supplementary systems were derived from the original three structures, primarily to study their behaviour under MD simulations.

1.4.2 Energetics-based analysis of CPD-containing DNA binding to Rad4 to commence the NER process

DNA serves as the genetic blueprint for a biological cell, providing necessary information during its functioning and at the time of cell replication, making it a fundamental molecule for life. Thus, its damage can have egregious effects. Cyclobutane pyrimidine dimer (CPD), a DNA lesion induced via UV radiation exposure, is the most prevalent DNA lesion linked to a wide range of genetic skin-related diseases and cancers in humans [77–80]. Rad4/XPC is a damage-sensing protein that recognises and repairs CPD lesions with high fidelity, with assistance from an array of proteins.

Rad4 consists of an N-terminal transglutaminase domain (TGD) and three β-hairpin domains (BHD1, BHD2, and BHD3) [81, 82]. The binding of TGD and BHD1 domains to the undamaged segment of DNA helps maintain its structural integrity. The interaction between BHD2 and DNA involves its β-hairpin, which binds to the DNA minor groove near the lesion, establishing hydrogen bonds with the DNA backbone. On the other hand, the β-hairpin of BHD3 interacts with the DNA major groove, filling the space left by the flipped-out CPD and its adjacent bases from the undamaged DNA strand. The BHD2-BHD3 binding interface securely retains these displaced partner bases. Existing literature on the study of RAd4-recognition reveals three key constituent processes: 1. Association of Rad4 with DNA, mainly binding of the TGD and BHD1 to the undamaged strand and BH2 and BHD3 to the lesion site, 2. insertion of the BHD2 and BHD3 β-hairpins into the major and minor grooves of the DNA, respectively, 3. forcing the CPD lesion and its partner bases (referred to as 3'-dA and 5'-dA) to flip out of the DNA duplex.

Previous work shows that in the absence of Rad4, the partner bases of CPD are unable to flip out of the DNA duplex. This makes it more apparent that the Rad4-association would've preceded the flipping process. This Rad4-association event seems more simplistic than it actually is. It acts as the anchor-establishing process for Rad4 on the DNA, especially at the lesion site.

Chapter 4 in this thesis aims at elaborating this Rad4-association process via the use of MD simulations and reaction-coordinate-fueled umbrella sampling. The kinetic gating mechanism of the Rad4/XPC damage recognition [83] proves that the residential time of Rad4 on the DNA is what

determines the overall efficiency of the entire NER process. Thus, it's of considerable priority to study the energetics of Rad4-association and compare the results with those of a homologous DNA sequence with lesions that have been extensively studied.

Chapter 2

Computational Methods

Contents

2.1 Introduction

Understanding biological functions requires an extensive study of the involved macromolecular structures. Interactions within and between these structures are what carry out such functions. Prior to the discovery of computational methods, the most popular and acceptable sources of gaining information about a given macromolecular structure were experimental methods such as NMR (nuclear magnetic resonance) and X-ray crystallography. These experimental methods shared the issue of variability, which arose from the scarcity of experimental data in some specific regions of the structure. An introductory take on the flexibility and energetics of a macromolecule would be obtained from these *experimental ensembles*. However, the thermodynamic properties of the system, like entropy or free energy, could be easily derived by analysing what are known as *conformational ensembles* [84]. These ensembles were obtained from boosted conformational sampling, which was possible all thanks to the advent of computational methods.

A piece of software that explores an approximated mathematical model by engaging iterative methods is known as a computational simulation. When such a mathematical model consists of specific equations that can be effectively translated into a simulation, computer simulations are called upon. Another use case for computer simulations is when a model is better characterised as a set of evolutionary rules than traditional mathematical equations. It is often used as an adjunct to, or substitute for, modelling systems for which simple closed-form analytic solutions are not possible. A prime example of computational simulation is Molecular Dynamics based computational simulation, an ensemble of structural restraints employed on the biomolecule(s) being studied and Euler, Hamiltonian, Lagrangian and Newtonian mechanics baked together in order to obtain a big picture of the system being dealt with and its role in an interested physiological process.

The earliest practical applications of computational methods were seen in the Manhattan Project in World War II, which modelled the process of nuclear detonation. Understanding the behaviour of neutrons was the problem at hand. Since then, computer simulations have been used to study biochemical regulatory networks, to formally model theories of human cognition and performance,to model biomolecules themselves for drug discovery, and to model viral infection in mammalian cells, and the list gets appended to date [85, 86]. To gather a holistic view of a given system, agent-based simulations that capture intra- and inter-component interactions slowly became influential. Computer simulations have become ubiquitous in the twenty-first century. Many different fields, such as the natural sciences, social sciences, life sciences, and humanities, have found ways to use computer tools, like simulations, in their research methods.

One of the most prevalent applications of computer simulations has been computational biomolecular modelling. The preliminary step for the Manhattan Project was to build a successful model of 12 hard spheres about to undergo a Monte-Carlo simulation. The modelling of water and other kinds of fluids to cultivate an understanding of fluid dynamics so as to exploitatively apply them in the field of automobile and aerospace mechanics also constitutes a prime example of computational modelling. Computational modelling is highly scalable; not just 12 atoms but systems as large as 150,000 atoms can be easily modelled today, all thanks to the advancements in modelling and computing techniques that took place throughout the late 20^{th} century and still continue in the 21^{st} century.

2.2 Computational Modelling and Visualisation

The properties of biological systems often emerge from complex interactions between their components. Predictive computational models of biological systems are useful to fully understand their behaviour and generate hypotheses about their functions [87]. Mathematical and computational modelling of biological processes and biomolecules continue to play important roles in deciphering mechanistic insight into metabolic and gene regulatory networks, cellular signalling, disease formation, and drug action. These models also allow for the prediction of the behaviour of biological systems under different environmental conditions.

Computational molecular modelling has emerged as a useful tool to understand the mechanisms of protein-inhibitor/activator recognition and binding and to predict and characterise the structure, dynamics, and energetics of biomolecules, pathogens, and biomolecular assemblies and complexes [88–92]. Thus, it has become an invaluable tool in pharmacology and medicine, wherein it is routinely used in the design and discovery of new pharmaceuticals with improved efficacy and safety [93–98]. The rapid determination of biomolecular structures from X-ray diffraction and solution NMR techniques, together with the recent advances in genomics, molecular biology, synchrotron sources, and computer software for data processing, continue to fuel the widespread use of modelling techniques in various domains of biology. In addition, the evolution of databases, data mining methods, and the whole infrastructure of bioinformatics also catalyses the growth and applications of modelling in biological problems.

Computational visualisation forms an essential part of molecular modelling; after all, a computationally generated molecular model is virtually valueless if it cannot be viewed at. Visualising a molecule on the computer unlocks avenues for commenting about its stability simply by observing its structure, shape, and type of bonds, and distances between atoms. It helps in forming a preliminary understanding of the system being dealt with. Software such as VMD, Pymol, Avogadro, and UCSF Chimera have widespread visualisation usage across the simulation community [99–102]. Due to the many types of visualisation modes available, a molecule such as a protein can either be visualised atom-by-atom or as a composition of its secondary structures. The real-life analogue to computational visualisation would

generally be observation under a fine electron microscope or an X-ray diffraction-based spectrometer, both of which would be inaccessible to a lot of researchers today and would be expensive to have multiple units of installed at any facility. On the other hand, visualisation software is just a piece of code that can be put up on the internet, thus making it available to the public at a much lower cost.

2.3 Statistical Mechanics

Thermodynamics was originally developed in the 19^{th} century and was driven by the dawn of the industrial revolution [103] and a desire to understand and optimise the extraction of useful work from engines. It was originally introduced as a phenomenological theory based entirely on the interrelationship of macroscopic variables such as temperature, pressure, volume, and energy. However, studying systems at a microscopic level, such as the behaviour between various atoms and molecules in a closed system, led to the birth of statistical mechanics. Stochastic methods are used within this branch of physics to examine the movements and energies of the constituent particles of a system in an attempt to bridge the gap between microscopic and macroscopic worlds. Statistical thermodynamics was introduced as the study of equilibrium states, with no net tendency for the system to evolve over time unless driven from the outside. Boltzmann and others proposed theoretical descriptions for this evolution and subsequent relaxation [104].

2.3.1 Phase space and states

Consider an N-particle system, with each particle having s degrees of freedom. The l coordinates, $l = sN$, are used to describe the spatial orientation of the system. A corresponding set of other l coordinates, the conjugate momenta of each particle with its spatial orientation, are used additionally. These $2l$ coordinates along with the equations of motion, can now determine the future and past course of the system. Consider a Euclidean space of $2l$ dimensions with perpendicular axes representing the orthogonality of spatial coordinates, q_1, q_2.....q_l, against the particle momentum, p_1, p_2.....p_l . Such a space is referred to as *phase space*, and any particular point in this space, a *phase point* . The spatial coordinates would be hereon referred to as $q(t)$, which is the same as q_1, q_2.....q_l , and the corresponding momenta as $p(t)$, same as p_1, p_2.....p_l . A phase point is more commonly known by the phrase *microscopic state*.

2.3.2 Ensembles

Given the time evolution of $q(t)$ and $p(t)$, the time average of an observable A (i.e. $\langle A \rangle_{\text{time}}$) can be calculated from the following equation.

$$\langle A \rangle_{\text{time}} = \lim_{\tau \to \infty} \left(\frac{1}{\tau} \int_{t=0}^{\tau} dt A(p(t), q(t)) \right)$$

On carrying out evaluation of this over an infinitely long duration of time, the integral approaches the ensemble value of A. This is defined as:

$$\langle A \rangle_{ens} = \int\int d^N p d^N q A(p(t), q(t)) \rho(q, p)$$

The angular brackets represent the mean, a.k.a. the expectation value of a random variable, and the subscript ens indicates the average value of the property A across all replicas of the ensemble generated over the course of the simulation, from hereon to be referred to as the ensemble average. The double integral represents $6N$ integral signs; one for each of the $6N$ positions and momenta of all the atoms in the system. $\rho(q, p)$ denotes the probability density of finding an arrangement of atoms with positions $q(t)$ and momenta $p(t)$. By integrating over all possible configurations of the system, the above integral is evaluated.

Statistical mechanics had taken care of studying systems with fewer particles. However, systems that were closer to reality, i.e., having a number of particles closer to a mole, had a variety of complexities: an enormous number of particles; interpersonal interactions being influenced by factors such as collisions, electromagnetic forces, and quantum mechanical effects; a lack of positional and momentum determinism; and a higher dimensionality due to a huge number of degrees of freedom associated. Introduced by J. Willard Gibbs in 1902, ensemble is an idealisation consisting of a large number of virtual copies (sometimes infinitely many) of a system, considered all at once, each of which represents a possible state that the real system might be in. This system of virtual copies helps in virtualizing the diverse ways in which particles could arrange themselves. This probability distribution of the microscopic state helps gauge macroscopic properties. In other words, a statistical ensemble is a probability distribution for the state of the system.

2.3.2.1 Canonical ensemble

The system to be simulated is immersed in an infinite-dimension bath such that particle exchange is not allowed and the temperature remains almost constant. This bath is commonly referred to as a *thermostat*. In addition, either the volume, energy, or pressure could be kept constant. Particle collisions are said to occur on the boundary. No pressure bath is used within this ensemble, thus causing fewer disruptions to the trajectory, provided that pressure is an irrelevant macroscopic property. As heat exchange can occur with the thermostat, the total energy of the system is no longer conserved. The key idea is to consider the combination of the system and the bath as an isolated system; thus, no heat exchange occurs with this combination, i.e., its total energy and number of particles remain constant.

Using the Boltzmann's distribution, the probability of a system existing in a microstate i defined by the phase point $(q_i(t), p_i(t))$ with total energy E_i at a temperature of T is given by:

$$\mathcal{P}(q_i(t), p_i(t)) = \frac{1}{Z_{NVT}} exp\left[-\frac{E_i}{k_B T}\right] \tag{2.1}$$

$$Z_{NVT} = \sum_i exp\left[\frac{-E_i}{k_B T}\right] = \frac{1}{3N!}\frac{1}{h^{3N}}\int\int d^N p d^N q exp\left[\frac{-E(q(t), p(t))}{k_B T}\right] \tag{2.2}$$

where Z_{NVT} represents the canonical partition function of the system, i.e. the sum of the Boltzman factor, $exp\left[\frac{-E_i}{k_B T}\right]$, across all possible microstates. Since all particles are uniformly treated, the term N! has to be placed , and to keep the partition function dimensionless, the factor $1/h^{3N}$ is included.

2.3.2.2 Isothermal-isobaric ensemble

In the isobaric-isothermal ensemble, the number of particles (N), the pressure (P) and the temperature (T) are kept constant, thus is known as NPT ensemble. This ensemble is important as most chemical reactions are usually carried out under constant pressure.

The probability distribution for a possible microstate i follows the equation 2.1 with the partition replaced with the following:

$$Z_{NPT} = \int Z_{NVT}\, exp\left[\frac{-PV}{k_B T V_0}\right]\, dV \tag{2.3}$$

Where V is the volume of the system, and the integral is carried out over all the space accessible to the system, and V_0 is defined as a volume constant for normalization.

2.4 Ergodicity

The idea of experiments is to study the time evolution of a certain property (for instance, binding energy, free energy, entropy, etc.) when the system is left for a certain period of time. The idea of simulations is to replicate microscopic-scale events that occur in the experiments computationally; hence, even in simulations, a time average of the desired physical quantity needs to be calculated [105].

The average time spent in a given region of the state space is proportional to the number of feasible states the region contains after a system has achieved equilibrium. All accessible microstates are equiprobable over a large period of time and thus have the characteristic of being energetically equivalent.

The dilemma appears to be that one can calculate time averages by molecular dynamics simulation, but the experimental observables are assumed to be ensemble averages. Resolving this leads us to one of the most fundamental axioms of statistical mechanics, the ergodic hypothesis, which states that **the time average equals the ensemble average**. Ergodicity signifies a system will eventually visit all parts

of the space while moving in a uniform and random sense. One goal, therefore, of a molecular dynamics simulation is to generate enough representative conformations such that this equality is satisfied. If this is the case, experimentally relevant information concerning structural, dynamic, and thermodynamic properties may then be calculated using a feasible amount of computer resources. Because the simulations are of fixed duration, one must be certain to sample a sufficient amount of phase space.

If the ergodic hypothesis is true, then time averages equal ensemble averages, and equipartition is a valid assumption. As for ergodicity, equiprobability across all possible states of a system that are energetically degenerate is assumed. The reason it is typically assumed that the probability density is uniform in an isolated system is because of Boltzmann's ergodic hypothesis [106]. This entails two separate assertions: (i) that for an isolated system, all points in phase space with a given energy lie on a single trajectory, and (ii) that the probability density in phase space is uniform along this trajectory.

The validity of this hypothesis is very difficult to prove [107]. Some regions of the feasible phase space are blocked by a concept known as KAM tori (Kolomogorov-Arnold-Moser), despite the randomness of the trajectories. The KAM tori have a smaller dimensionality than the entire accessible space of the system at hand and remain isolated from one another. Apart from the laws of conservation of energy, other conservation laws dependent on parameters such as the starting structure and other macroscopic properties can also be seen being applied within KAM tori.

The probability density of the ensemble is given by

$$\rho(p^N, r^N) = \frac{1}{Z}\exp\left[\frac{-H\left(p^N, r^N\right)}{k_B T}\right] \tag{2.4}$$

where H is the Hamiltonian, T is the temperature, k_B is Boltzmann's constant and Z is the partition function defined in the previous section.

This integral is generally extremely difficult to calculate because one must calculate all possible states of the system. The configurations of a system of particles are updated at every timestep within an MD simulation. Hence, a span of the entire set of possible configurations under the given thermodynamic conditions is required to evaluate the above integral during an MD simulation.

In a molecular dynamics simulation, the points in the ensemble are calculated sequentially in time, so to calculate an ensemble average, the molecular dynamics simulations must pass through all possible states corresponding to the particular thermodynamic constraints.

Another way, as done in an MD simulation, is to determine a time average of A(a generalized representation of a thermodynamic quantity), which is expressed as

$$\langle A\rangle_{time} = \lim_{\tau\to\infty}\int_{t=0}^{\tau} A\left(p^N(t), r^N(t)\right) dt \approx \frac{1}{M}\sum_{t=1}^{M} A(p^N, r^N) \tag{2.5}$$

where t is the simulation time, M is the number of time steps in the simulation and $A(p^N, r^N)$ is the instantaneous value of A.

2.5 Potential energy dependent partition function

The total energy (E) and the canonical partition function (Z_{NVT}) of the system are defined in eq. (2.6) and eq. (2.7), respectively. Here, $K(p)$ and $U(q)$ denote the total kinetic and potential energies, respectively. By solving the kinetic energy part of the integral, the canonical partition function is expressed as a function of only the spatial coordinates of the system. C_0 in eq. (2.12) represents the final normalized constant obtained after the reduction.

$$E(q,p) = K(p) + U(q) \tag{2.6}$$

$$Z_{NVT} = \frac{1}{3N!}\frac{1}{h^{3N}}\int\int d^N p\, d^N q \exp\left[\frac{-E(q(t),p(t))}{k_B T}\right] \tag{2.7}$$

$$= \frac{1}{3N!}\frac{1}{h^{3N}}\int\int d^N p\, d^N q \exp\left[\frac{-K(p(t)) - U(q(t))}{k_B T}\right] \tag{2.8}$$

$$= \frac{1}{3N!}\frac{1}{h^{3N}}\int d^N p \exp\left[\frac{-K(p(t))}{k_B T}\right]\int d^N q \exp\left[\frac{-U(q(t))}{k_B T}\right] \tag{2.9}$$

$$= \frac{1}{3N!}\frac{1}{h^{3N}}\prod_{i=1}^{N}\int dp_i \exp\left[\frac{-\|p_i\|^2}{2m_i k_B T}\right]\int d^N q \exp\left[\frac{-U(q(t))}{k_B T}\right] \tag{2.10}$$

$$= \frac{1}{3N!}\frac{1}{h^{3N}}\left(\left(\frac{m_i k_B T}{2\pi}\right)^{3/2}\right)^N \int d^N q \exp\left[\frac{-U(q(t))}{k_B T}\right] \tag{2.11}$$

$$Z_{NVT} = C_0 \int d^N q \exp\left[\frac{-U(q(t))}{k_B T}\right] \tag{2.12}$$

Using eq. (2.3), even the partition function for an isothermal-isobaric ensemble could be represented in a similar manner. Thus, the problem is now broken down to the selection of an appropriate potential energy function in order to gauge the partition function and, in turn, the probability distribution of all possible micro-states.

2.6 Potentials Used in MD

In principle, the nature of intermolecular interactions in a system should be built on a quantum mechanical framework. However, the molecular dynamics simulation embodies a classical view of the same. Instead of describing the interparticle interactions in terms of overlapping electron clouds of atoms, MD simulations consider atoms as point masses coupled to each other via exotic springs. This methodology is adopted due to its reasonable simplicity and capability to be extended across large systems without substantially increasing the computational load, as opposed to the rigorous quantum mechanical description, which is still hard pressed in dealing with even the smallest systems. The Potential Energy Function, from hereon will be referred to as PEF, is a classical mechanics function

used to describe the potential energy of a system. Computational models for such *effective* potentials undergo refinement based on the differences produced in the experimental and model-based studies, and a substantial amount of evidence pointing against the model begs its ground-up redevelopment.

These PEFs are referred to as *force fields* in the context of MD simulations. The transformation of quantum mechanical to classical potential energy functions is feasible with two major approximations. The first is the Born-Oppenheimer approximation, which suggests considering the electrons to be treated separately from their nuclei since they can react instantaneously to the nuclei's motion due to their much faster dynamics. The second is that nuclei are to be treated as point particles that follow classical Newtonian dynamics. In conjunction, these permit the use of nuclear positions in conjunction with the force fields to calculate the actual potential energy of the system.

Often calibrated to experimental and quantum mechanical results, the generation of force fields aims at reproducibility and computational optimisation. Structural data obtained from X-ray crystallography and NMR, dynamic data obtained from spectroscopy, inelastic neutron scattering, and thermodynamic data play an important role in the same. An area of continual research, work is still being done on generating generalizable potential energy functions applicable to all kinds of biomolecules. The most commonly used force fields used for MD simulations are the AMBER, CHARMM, GROMOS and OPLS/AMBER [108–111].

The basic PEF that any of the above listed force fields use is the same. It is a function of the atomic positions of the constituent atoms of the system, r^N (assuming an N-particle system). These force fields only differ w.r.t. the value of the relevant constants pre-set for each interaction term present in the equations shown below.

$$\mathcal{U}(r^N) = \overbrace{\underbrace{\sum_i^{n_{bonds}} b_i\,(r_i - r_{i,eq})^2}_{\mathcal{U}_{\text{bond-stretch}}} + \underbrace{\sum_i^{n_{angles}} a_i\,(\theta_i - \theta_{i,eq})^2}_{\mathcal{U}_{\text{bond-bend}}} + \underbrace{\sum_{1,4\text{ pairs}} K_\phi\,(1 - \cos(n\phi))}_{\mathcal{U}_{\text{torsion}}} +}^{\mathcal{U}_{\text{bonded}}}$$
$$\overbrace{\underbrace{\sum_{i<j}^{n_{atoms}} 4\epsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]}_{\mathcal{U}_{\text{van der Waals}}} + \underbrace{\sum_{i<j}^{n_{atoms}} \frac{q_i q_j}{4\pi\epsilon r_{ij}}}_{\mathcal{U}_{\text{electrostatic}}}}^{\mathcal{U}_{\text{non-bonded}}} \tag{2.13}$$

Energy can culminate from internal, bonded interactions denoted by $\mathcal{U}_{\text{bonded}}$ and from external, non-bonded interactions denoted by $\mathcal{U}_{\text{non-bonded}}$. The bonded contribution is a cumulative of the interaction energy due to bonds, angles and bond rotations in a molecule. A pictorial representation of these components is provided in Figure 2.1 .

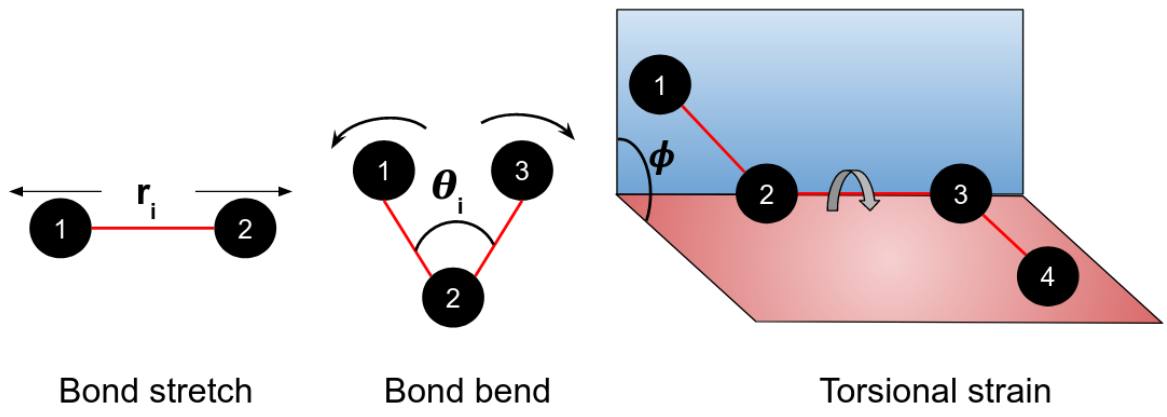


Figure 2.1: **Bonded interactions.**

1. **Bond stretch**: Covalently bonded atoms, also known as *1,2-pairs*, store harmonic potential energy as a virtue of their harmonic bond. An approximate function to this energy is what is seen in the above equation. This estimate depends on the displacement from the ideal bond length($\mathbf{r_{i,eq}}$), i.e., $\mathbf{r_i} - \mathbf{r_{i,eq}}$ and a force constant $\mathbf{b_i}$, which captures the strength of the bond. The chemical type of the bonded atoms is the determinant of these constants.

2. **Bond bend**: Two adjacent bonds having a common atom possess angular strain caused by the deviation of their bond angle from the equilibrium bond angle value($\theta_{\mathbf{i,eq}}$), i.e. $\theta_\mathbf{i} - \theta_{\mathbf{i,eq}}$. The strain constant $\mathbf{a_i}$ determines the extent of impact even the smallest of angular displacements will have on the aggregate stability of the molecule. Both of these constants depend on the bond type and the atom type that constitute these adjacent bonds.

3. **Torsional strain**: Steric influence between atoms separated by three covalent bonds, also known as *1,4-pairs* results in the possession of torsional strain. . Three consecutive covalent bonds form a dihedral angle, and a rotation about any of the bonds results in the building or loosening up of this potential. This potential is assumed to be periodic and is often expressed as a cosine function. Torsional constant $\mathbf{K}_\phi$ and ϕ itself depend upon the nature of the atoms and the type of covalent bonds in between the consecutive atoms.

The non-bonded interactions are an aggregate of electronic cloud repulsion set up due to interacting dipoles/induced-dipoles, and static coulombic interactions, observed in non-bonded atoms, which may be of the same of different molecules. The variation of either of these energies w.r.t. interatomic separation is provided in Figure 2.2.

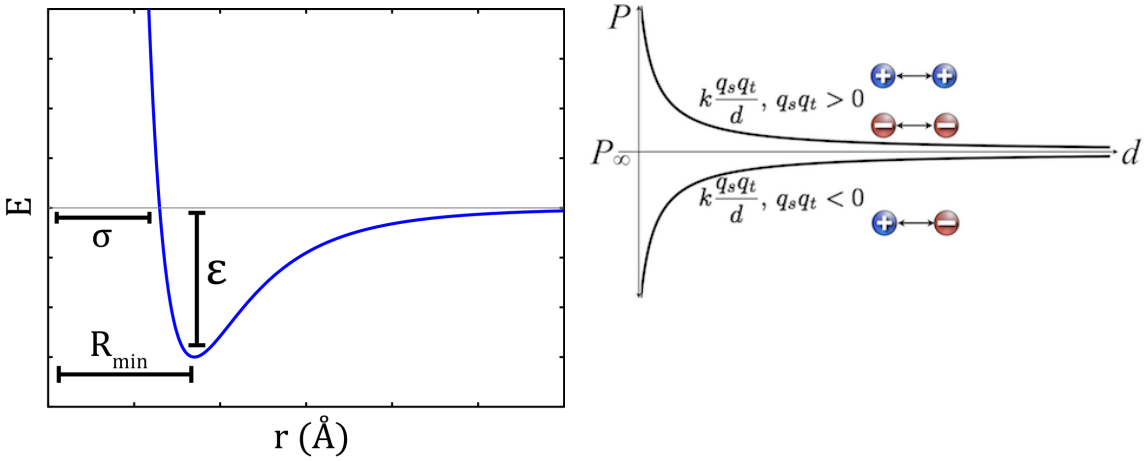


Figure 2.2: **Non-bonded interactions.** The image on the left-hand side represents the variation of the Lennard-Jones potential used to model the vanderWaals interactions, $\mathcal{U}_{\text{van der Waals}}$, with the interatomic separation between the interacting species, r. σ represents the separation value at which potential becomes 0, R_{min} represents the separation corresponding to the minima of this potential and ϵ represents the value of this minima. The image on the right-hand size represents the variation of coulombic interaction potential, $\mathcal{U}_{\text{electrostatic}}$, with interatomic separation, d. The curve drawn above and below the x-axis represents this variation for like and unlike respectively.

1. **Van der Waals interactions**: These arise from a balance between repulsive and attractive forces. The former are set up at short distances, where the electron-electron interaction is strong, and the latter, also referred to as the dispersion force, arises from fluctuations in the charge distribution of the electron clouds. These fluctuations result in an instantaneous dipole which in turn, induces a dipole in a second atom or molecule giving rise to an attractive interaction.

   The van der Waals interaction is most often modelled using the Lennard-Jones 6-12 potential, which expresses the interaction energy using the atom-type dependent constants σ and ϵ which are experimentally defined in nature, and the interatomic separation denoted by **r**. These interactions tend to zero at infinite atomic separation and become significant as the distance decreases. The attractive interaction is longer range than the repulsion, but as the distance become short, the repulsive interaction becomes dominant, thus resulting in an energy minima. The positioning of the atoms at optimal distances stabilises the system.

2. **Electrostatic interactions**: Interactions set up by coulombic forces between non-bonded atoms are included in this term. ϵ_r is the relative permittivity of the medium in which atoms of charges q_i and q_k separated by a distance r, electrostatically interact.

The empirical PEF is smudged by several limitations, resulting in inaccurately calculated potential energy. A major limitation is forbidding any sort of radical change in electronic structure, i.e., events like bond making or breaking cannot be modelled. Mixed *quantum mechanical-molecular mechanical* force fields are under development in order to tackle this limitation.

Another limitation stems from a fixed set of atom types employed when determining the parameters for the force field. Atom types are used to define atoms that are involved in a particular type of bonding. For instance, an aliphatic, sp^3-bonded carbon atom has different properties than a carbon atom found in the cycloalkane ring. Such atoms are grouped in the same categories in order to minimise the total number of atom types, in lieu of denoting them as different atoms with unique parameters. This can lead to *type-specific errors*. A single set of parameters could still work for environmentally passive atoms like hydrogen and carbon, but reactive atoms like oxygen and nitrogen involved in various bonding settings require more types and parameters to completely capture their contribution.

An important point to note is that the PEF by itself does not include entropic effects. Thus, a minimum value of $\mathcal{U}$ actually corresponds to the minimum value of free energy and not to an actual equilibrium, the most probable state. Since routine simulations are implemented with isothermal-isobaric conditions, the equilibrium state corresponds to the minimum of Gibb's Free Energy. These ignored entropic effects are usually included while evaluating PEFs in a molecular dynamics simulations. Several other approximations in order to efficiently compute potential energy have been discussed later on in a section titled *Simulation Tactics*.

2.7 Energy Minimization

The process of locating the energy-minimum configuration of a system on its PES is termed *energy minimization* (EM). The net force on individual atoms of the system in the energy-minimum configuration is approximately zero. For instance, the energy minimum configuration of a water molecule in the gas phase would correspond to a structure in which the O-H bond lengths and H-O-H bond angle would be equal to their equilibrium values. The energy minimised structures may correspond to local or global minima of the PES. Local minima are usually reached when EM stops after finding the first stable configuration. Since this configuration may not be the most-stable one, suitable EM algorithms that allow the system to cross-over energy barriers on the PES can be used to reach the most stable global minimum configuration of the system. In all EM methods, the atomic coordinates of a given many-body system are gradually changed to generate configurations with lower and lower energies until the minimum is reached [112].

EM methods can be broadly classified into two types: derivative-based and non-derivative-based methods. Derivatives of the potential energy function with respect to atomic coordinates possess key information on the shape and stationary points of the PES. Several points need to be considered before

choosing an EM approach: efficiency, accuracy, computational resources, and time required for execution. No single viable-for-all EM approach exists, and depending on the molecular model, it may need to be hand-picked and even fine-tuned. Efficient EM approaches, which usually have a quantum mechanical foundation, may atrociously fail when applied to molecular mechanics due to the number of atoms in the latter being several multiples higher than those in the former. Procedures such as the inversion of Hessian or other similar atomic matrices may suffer due to the same issues. Molecular mechanics usually requires an algorithm with a larger number of steps over which energy will be minimized. Therefore, we have various methods in various popular software packages [113].

No EM algorithm to date has been able to identify the global minima efficiently if forced to start from a single, random point. Hence, algorithms popularly in use start off with several points on the PES, and all of them are minimized. At times, the global energy minimum may not be the most populous minimum, i.e., the most probable structure. For instance, a global minimum could be characterised as a steep valley having a small population, perhaps owing to a smaller number of vibrational states having the energy to access it, when compared to a local minimum shaped as a wide basin having a higher population owing to a larger number of vibrational states being energetically capable of accessing it. Thus, a combination of the energy minimum value and its population determines the most probable structure. Hence, the common structure taken by a system could very well not be the global minimum, or the one with the highest population, or to any local minimum found in its PES [114].

2.7.1 Derivate based minimization methods

The derivative of the PEF w.r.t. each degree of freedom is calculated. Usually, a Cartesian or internal coordinate representation is used for degrees of freedom. Both analytical and numerical methods are used to calculate the derivatives, but the former being easier and quicker to generate in addition to being more accurate, is preferred over the latter. If it's impossible to generate derivatives using analytical methods, then it's advisable to follow the non-derivative minimization-based approach since it's more efficient than the numerical one [114].

2.7.1.1 Steepest Descent method

Coordinates of the system are modified in an iterative manner, such that those in the current iteration, x_k^i, are used to calculate the gradient and in turn x_{k+1}^i, coordinates of the subsequent iteration. Using the PEF-gradient:- $\frac{\partial \mathcal{U}}{\partial x_k^i}$ for the k^{th} iteration, the corresponding update equation is given by:

$$x_{k+1}^i = x_k^i - \eta^i . \frac{\partial \mathcal{U}}{\partial x_k^i} \tag{2.14}$$

η^i represents the decay rate for updating the i^{th} coordinate and is usually kept constant while carrying out all the iterations. A 1D representation of the steps of this algorithm is shown in Figure 2.3. Explo-

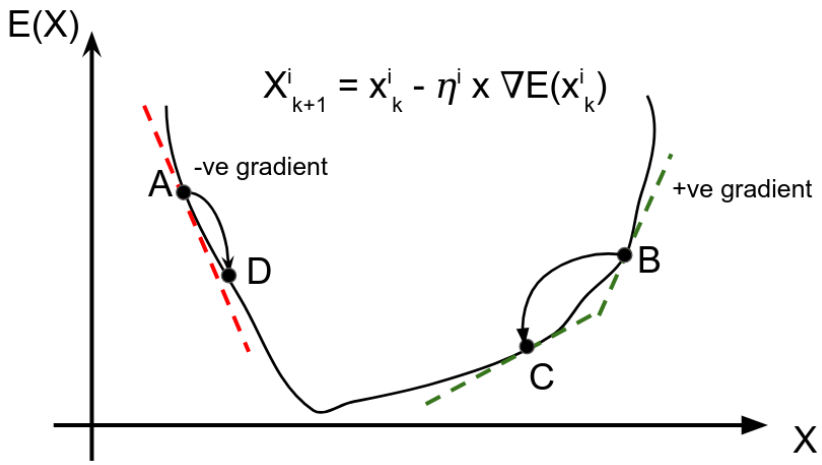


Figure 2.3: **Steepest Descent for 1-D PEF.** A simple display of the PEF, E(χ) (which is same as $\mathcal{U}(\chi)$) varying with the reaction coordinate, χ. On taking either of the starting points A or B, the algorithm is bound to arrive at the minimum since the sign of gradient decides the direction of movement. The value of the gradient and the learning rate η determines the step-size.

ration along the direction of net force, which is perpendicular to that of the previous iteration, results in an inefficient way of reaching the minima, especially on energy surfaces with narrow valleys, since the method leads to numerous oscillations while descending to the minimum. A slight modification in the form of updating the position any time the trial point along the gradient has a lower energy will result in an efficiency increase due to the decrease in the net function evaluations, leading to a drastic decrease in the computational time.

Relying on gradients is both an advantage and a limitation of this method. Although as the minimum is reached, the gradient approaches zero, slowing down convergence, it is extremely robust even when systems are far from harmonic. Hence, it is also termed as *A Robust but Slowly Converging Algorithm*. It is often used when initial configurations are considered to be far from the minimum, i.e., large gradients, for instance in relaxing poorly refined crystallographic or computationally modelled configurations.

2.7.1.2 Conjugate gradient method

Even in this method, gradients are used to obtain the optimum. The gradients are perpendicular to the search direction at every point, and steps are taken in a conjugate manner (and not perpendicular, as was the case in steepest descent), which is why it is more correctly known as the *conjugate direction* method (Figure 2.4). The energy minimum is located through the use of a set of conjugate directions existing for a quadratic function of M variables. Here, $M = 3^N$, since the total number of coordinates across which minimization needs to be carried out is 3^N. The conjugate gradient method moves in the direction

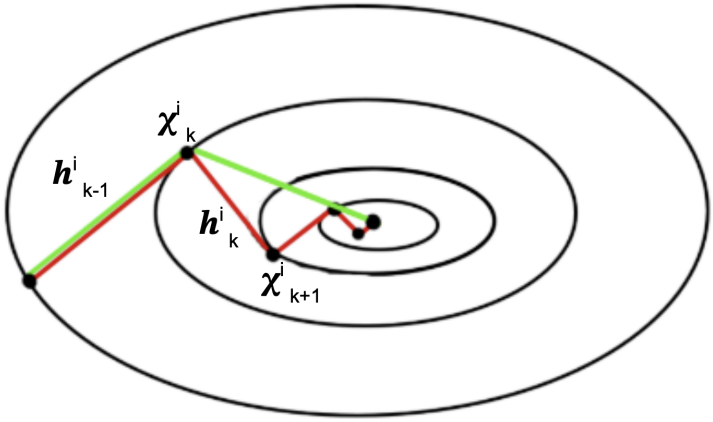


Figure 2.4: **Conjugate Gradient.** The contour plot of the PEF, with the steps of the steepest descent method in green and that of the conjugate gradient method in red. The convergence of the conjugate gradient method is $\mathcal{O}(n)$ steps, n being the dimensionality (here $n = 2$)

h^i_k from point x^i_k where h^i_k is computed from the gradient at the point and the previous direction vector h^i_{k-1}.

Update equation:

$$x^i_{k+1} = x^i_k - \eta(\nabla\mathcal{U}(x^i_k) + \beta h^i_k) \tag{2.15}$$

$$h^i_k = \nabla\mathcal{U}(x^i_k) + \beta h^i_{k-1} \tag{2.16}$$

Since the step history is used to accelerate convergence, the algorithm requires storage of all previous searching directions and coordinates and thus can be computationally expensive. Hence, a robust but slow algorithm like the Steepest Descent is used, which is not as computationally intensive, to converge into a small region containing the optimum, followed by running this method for a smaller number of iterations, which pin-points the actual optimum, thereby limiting the computational resource demand.

2.7.2 Non-derivatives based methods

2.7.2.1 Simplex Method

To minimize a function having M dimensions, a geometrical figure with $M + 1$ vertices connected in a pairwise manner is employed which is known as *a simplex*. Subsequently, a PEF of $3N$ Cartesian coordinates will have a corresponding simplex described by $3N + 1$ vertices; if internal coordinates are used the number of vertices will decrease to $3N - 5$.

The simplex algorithm finds the lowest energy point by exploring the potential energy surface. The simplex constructed using these $3N + 1$ vertices rests on the PES such that its vertices cut the PES at unique points. The simplex algorithm consists of three primary movements. A popular move is to reflect the vertex with the highest PEF value across the simplex to its opposite side. If this move results

in a reduced PEF value, the corresponding point replaces the original point in the simplex figure. The *reflection and expansion* move is exercised when reflection doesn't yield a point with a lower PEF value. If a minimum is reached, which is not necessarily the global one, then to progress, a *contraction along the highest dimension* is performed. The simplex will *contract along all of its dimensions* as a last resort for finding a lower energy point along the PES.

Employing the simplex method of minimization requires the creation of a preliminary set of vertices. The $M + 1$ dimensional figure created as a result will only pass through a single vertex within this set. By changing the coordinate values for all other points such that they ultimately lie on the PES, an initial simplex is created.

Simplex method is the most efficient, provided the initial conformation of the system is highly energetic. The computational time blows up as the system size increases; hence, this method can only be efficiently used for smaller systems. To add to this dismay, the procedure cannot be readily parallelized to use distributed computing resources to speed up the computation time. In simulation software, this algorithm is iterated for a few epochs to energy-minimise the initial structure, followed by the use of an efficient method that can be used for further calculations.

2.7.2.2 Sequential Univariate Search Method

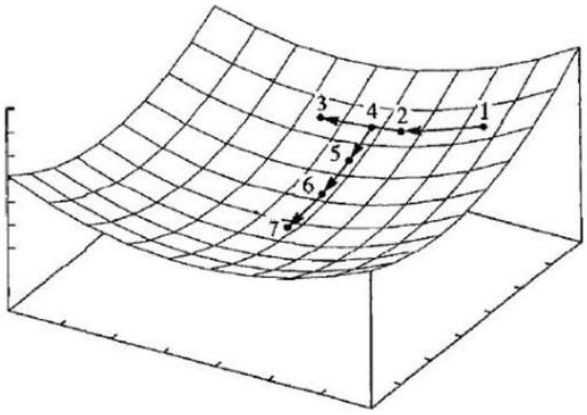


Figure 2.5: **The sequential univariate search approach.** Beginning from point 1, points 2 and 3 are created after exercising the first step of this approach. Point 4, a minimum along a parabola fit w.r.t. these points is located. Iterative repetition of these algorithmic steps yields rest of the points. Image source [115]

Since the simplex algorithm requires numerous energy calculations, the sequential univariate search method is preferred when the computational demand is high enough. For every unique degree of freedom, two new states are created by making changes in the existing ones (i.e., using a degree of freedom having value as x_i, 2 configurations with values = $x_i + \partial x_i$ and $x_i + 2\partial x_i$ are generated). These three points are then fit on a parabola. The minima of this parabola is located. The subsequent step involves

warping the coordinate to this located minima, i.e., the conformation corresponding to this minima is taken as the new starting configuration, and the parabola-construction procedure is repeated. A minimum is achieved when infinitesimally small fluctuations across all directions (degrees of freedom) are taken. Fewer PEF evaluations are carried out than the simplex method. A slow convergence is observed when: 1. two points taken share a strong bond; or 2. the PES is shaped like a gradual shallow valley.

2.8 Molecular Dynamics Simulation

Alder and Wainwright were the first scientists to introduce the MD method in the late 1950s [116, 117]. Their work involved simulating the behaviour of hard spheres. Their results led to an increased understanding of simple liquids. In 1964, Rahman carried out a simulation of liquid argon using a realistic potential for the first time [118]. Later on, in 1974, along with Stillinger, he carried out what was the first of its kind: molecular dynamics simulation of a realistic system: liquid water [119]. This was followed by the simulation of BPTI (bovine pancreatic trypsin inhibitor) in 1977, which was the first protein to be computationally simulated [120].

Fast forward to today, and studies based on MD simulations of solvated proteins, protein-DNA complexes, and lipid systems have addressed a variety of issues, including the thermodynamics of ligand binding and the protein folding of small proteins. Specialised simulation techniques tailored for handling specific problems have been discovered, such as mixed quantum mechanical-classical simulations to study enzymatic reactions. NMR experiments used to resolve the structures of various compounds suffer from high variability when experimental data regarding specific regions of the structure is inadequate. Instead of analysing a single set of coordinates obtained from the PDB file of a molecule to be simulated, MD simulations employ *conformation ensembles* as a means of increasing the sampled space. This is possible due to significant progress made in the computational efficiency of the underlying simulation algorithms. The data from these ensembles can be fed into the appropriate mathematical functions to evaluate the macroscopic properties of the foundational system. Complicated events such as certain configuration shifts and protein foldings can be simulated thanks to ensembles. Even experiment-based studies can be replicated since they measure nothing but the time-averaged properties of a conformational ensemble.

The MD technique shall be listed from hereon.
Consider a system of N particles, with:

- atomic positions of the entire system represented as $\mathbf{r} = \{r_1, r_2, ...r_N\}$
- atomic momenta of the entire system represented as $\mathbf{p} = \{p_1, p_2, ...p_N\}$
- the PEF of the system represented as $\mathcal{U}(\mathbf{r})$
- the total kinetic energy of the system denoted by $\mathbf{K}(\mathbf{p})$

- m_i and $F_i(t)$ represents mass and net force on particle i respectively

The equations of motion for the i^{th} particle can then be presented as:

$$\dot{r_i} = \frac{p_i}{m_i} \tag{2.17}$$

$$F_i(t) = \dot{p_i}(t) \tag{2.18}$$

The net force applied on each particle is exactly equal to the negative of the gradient of the PEF w.r.t. its coordinate. Mathematically speaking, $F_i(t) = -\nabla_i \mathcal{U}(\mathbf{r})$. Thus, on initialisation of velocities via Boltzmann distribution at some temperature T of the system, and the calculation of the PEF using the force fields files for the system, the force on each atom can be calculated, thus enabling the estimation of momentum and finally that of the position, for each particle, using the equations (2.17) and (2.18). These steps when re-iterated for a certain number of timesteps, constitute of the basic MD algorithm. A numerical integration algorithm is required for updating the momentum and position values, since these kinetic and potential energy functions are highly complicated, thus rendering a generalised analytical solution unavailable.

2.8.1 Verlet Integration

The first numerical integration algorithm to be proposed goes by the name *Verlet* integration. Expressing the position as a taylor series w.r.t. time, we get:

$$r_i(t + \Delta t) = r_i(t) + \Delta t \frac{\partial r}{\partial t} + \frac{1}{2!}(\Delta t)^2 \frac{\partial^2 r}{\partial t^2} + \frac{1}{3!}(\Delta t)^3 \frac{\partial^3 r}{\partial t^3} + O((\Delta t)^4) \tag{2.19}$$

$$\Rightarrow r_i(t) + \Delta t \dot{r} + \frac{1}{2!}(\Delta t)^2 \frac{F_i}{m_i} + \frac{1}{3!}(\Delta t)^3 \frac{\partial^3 r}{\partial t^3} + O((\Delta t)^4) \tag{2.20}$$

In theory, this value of Δt tends to 0, thus higher order terms are ignored since they would have little to no impact due to higher powers of this Δt term. This Δt is referred to as the integration timestep.

Similarly, the position of the previous timestep, using that of current timestep is:

$$r_i(t - \Delta t) = r_i(t) - \Delta t \dot{r} + \frac{1}{2!}(\Delta t)^2 \frac{F_i}{m_i} - \frac{1}{3!}(\Delta t)^3 \frac{\partial^3 r}{\partial t^3} + O((\Delta t)^4) \tag{2.21}$$

Adding equations 2.20 and 2.21, followed by rearranging so that $r_i(t + \Delta t)$ becomes the LHS, we get:

$$r_i(t + \Delta t) = r_i(t - \Delta t) - 2r_i(t) + (\Delta t)^2 \frac{F_i}{m_i} + O((\Delta t)^4)$$

Hence, to estimate the particle position at the next timestep, the algorithm requires storage of the current and previous timesteps. Using PEF, the force on each particle will be calculated, and the fourth-order terms can be conveniently ignored. Notice that velocities are not required for updating the positions. However, on thinking carefully, it seems that during the starting up of the algorithm, $r_i(0)$ denotes the initial coordinates of the system fed into the algorithm, but for calculating $r_i(\Delta t)$ we have no $r_i(-\Delta t)$ since this timestep does not exist by definition. Thus, velocity calculation is required only once, in order to start up the algorithm.

2.8.2 Velocity Verlet integration

A related and more commonly used algorithm is the velocity Verlet algorithm, first introduced in 1982 [121]. As the velocities and positions of particles are updated at the same time, this algorithm bypasses the problem faced in the first time step of the basic Verlet algorithm.

The update equations of this integration method are extracted via expressing them in terms of intermediate timestep quantities, i.e. values of quantities at $t + \frac{\Delta t}{2}$. The intermediate momentum can be expressed as:

$$p_i\left(t + \frac{\Delta t}{2}\right) = p_i(t) + \frac{F_i(t)}{m_i}\frac{\Delta t}{2} \tag{2.22}$$

And the final momentum, in terms of this intermediate momentum can be expressed as:

$$p_i(t + \Delta t) = p_i\left(t + \frac{\Delta t}{2}\right) + \frac{1}{2}\frac{F_i(t + \Delta t)}{m_i}\Delta t \tag{2.23}$$

Adding 2.22 and 2.23, we get:

$$p_i(t + \Delta t) = p_i(t) + \frac{[F_i(t) + F_i(t + \Delta t)]\,\Delta t}{2m_i}$$

The steps of the algorithm are as follows:

1. Initialise the coordinates $r_i(0)$ from the starting the structure and the momenta $p_i(0)$ from a boltzmann distribution at the simulation temperature.

2. Update the position from the current timestep t to next timestep $t + \Delta t$ as:

$$r_i(t + \Delta t) = r_i(t) + \frac{p_i(t)}{m_i}\Delta t + \frac{1}{2!}\frac{F_i}{m_i}(\Delta t)^2$$

   For this step, the knowledge of the momentum(velocity), PEF and thus net force for the current timestep of each particle is required.

3. Since the positions at $t + \Delta t$ is known, compute the PEF and thus the net force on each particle, i.e. $F_i(t + \Delta t)$.

4. Update the momentum

$$p_i(t + \Delta t) = p_i(t) + \frac{[F_i(t) + F_i(t + \Delta t)]\,\Delta t}{2m_i}$$

Returning to the discussion of MD simulations, simulating systems with 50K-100K atoms has now become a common practice, and provided that the computational power is low, simulating 500K atoms can also be performed effortlessly. The development of high-performance computing (HPC) facilities has a subpar but important role to play in the evolution of computational MD simulations, along with the simplicity of the basic MD algorithm.

CPU parallelization and accelerators have now made their way into most of the day-to-day computers. Through the message-passing interface (MPI), certain computer programmes can now utilise multiple cores of CPUs concurrently, which is being taken advantage of by some of the widely used simulation packages such as AMBER, GROMACS, NAMD, and CHARMM [108, 122–124]. Each CPU is responsible for simulating a small but continuous portion of the actual ensemble, thus being able to leverage the proximity of the corresponding interactions. It is important for CPUs that simulate adjoining sections of a system to transfer information between each other.

The use of GPUs as fully programmable, high-performance processing units has been a solid breakthrough in simulation packages. All of the aforementioned widely used simulation packages already have GPU-optimised, parallelised versions. Since GPUs excel at parallelism and speed of calculations, a wider adoption of GPU-optimised simulation software has taken place. Energy efficiency is an added benefit realised when opting for GPU-optimised packages.

2.9 Potential of Mean Force

MD simulations are also routinely used to examine the variation of thermodynamic quantities (free energy or the potential of mean force (PMF), in particular) with respect to a few key collective variables of the system of interest. The reaction coordinates could be as simple as the distance between two atoms or as complex as the number of contacts between two functionally important domains of the system. Mathematically, the PMF of a system of N particles is the potential that gives the average force exerted by all the $n + 1...N$ particles on a particle j, over all the configurations, where for each configuration a set of particles $1...n$ is kept fixed [125].

$$-\nabla_j w^{(n)} = \frac{\int e^{-\beta \mathcal{U}}(-\nabla_j \mathcal{U})\,dq_{n+1}...dq_N}{\int e^{-\beta \mathcal{U}}\,dq_{n+1}...dq_N}, j \epsilon [1, 2, 3.....n] \tag{2.24}$$

$-\nabla_j w^{(n)}$ is the mean force acting on particle j, and $w^{(n)}$ is the potential of this force, i.e. PEF of this force. To simplify, if $n = 2$, the resultant PEF or potential of force $w^{(2)}(r_{12})$ would represent the average work to be done in order to pull 2 particles initially at an infinite separation to a distance of r units.

The potential of force along a reaction coordinate (degree of freedom) is usually expressed as a natural logarithmic function of its probability distribution. Since MD simulations are usually carried out in an ensemble, the ensemble average of this probability distribution is usually used to calculate the PMF. Formally, it is given by

$$\mathcal{P}(\chi) = \langle \delta[\chi - \chi(q)] \rangle$$

For a canonical ensemble(NVT constant):

$$\mathcal{P}(\chi) = \frac{\int dq\, \delta[\chi - \chi(q)] \exp.(-\mathcal{U}(q)/k_B T)}{\int dq \exp.(-\mathcal{U}(q)/k_B T)} = \frac{Z(\chi)}{Z}$$

Z indicates the partition function and $Z(\chi)$ is also a partition function but for a system which is constrained to lie on a surface defined by the equation $\chi(q) = \chi$. The Landau free energy is then given by:

$$\mathcal{F}(\chi) = F - k_B T \ln((P)(\chi))$$

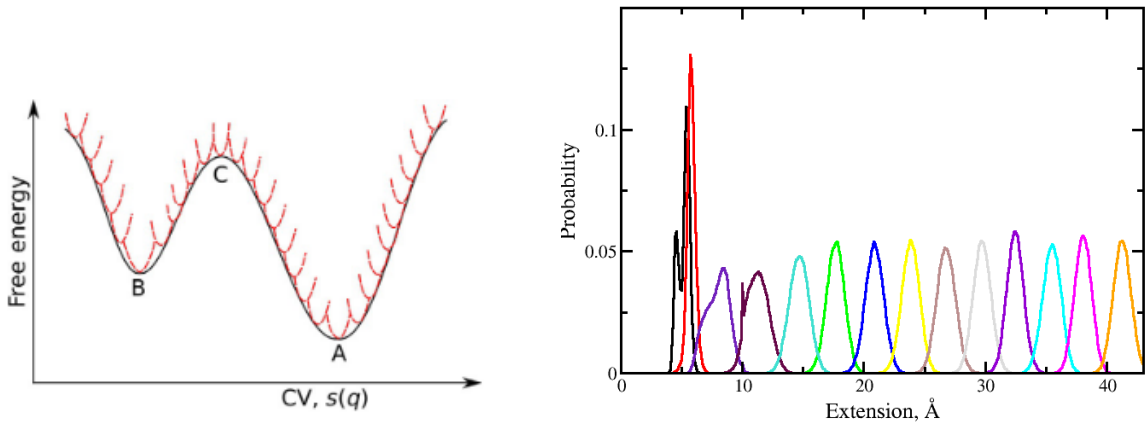
where F is a arbitrary fixed constant.

For processes with an activation barrier higher than $k_B T$ the distribution function $\rho(\chi)$ cannot be computed by a straight molecular dynamics simulation. Such computations would not converge due to low sampling in higher-energy configurations. Special sampling techniques (non-Boltzmann sampling) have been developed to obtain a PMF along a coordinate χ. One of such enhanced sampling techniques is Umbrella Sampling. Often PMF simulations are used in conjunction with umbrella sampling due to the aforementioned problem of convergence and low sampling for high-energy configurations [113].

2.10 Enhanced Sampling: Umbrella Sampling

Having its foundation laid on previous works [126, 127], umbrella sampling was founded by Torrie and Valleau [128, 129]. Umbrella sampling was employed to achieve and study molecular interactions in fluids well below their critical point. The method gained popularity in the 1980s with advancements in computational resources and simulation algorithms. Researchers started applying umbrella sampling to study a wide range of biological, chemical, and physical processes, including protein folding, ligand binding, and phase transitions.

The free energy profile of a system may be characterised by numerous local minima, local maxima, and saddle points. It may so happen that some of these minima are separated by saddle points having a barrier height much larger than the thermal energy ($k_B T$) for that particular temperature, and thus a



(a) (b)

Figure 2.6: (a) Schematic illustration of the umbrella sampling method. x-axis represents the reaction coordinate value and y-axis represents the corresponding free energy value. Red lines represent the PMF functions F_i, for each window, in other words harmonic bias potentials added to the system Hamiltonian at different windows along the reaction coordinate($s(q)$) space. Black represents the actual, unbiassed PMF function F. Figure borrowed from [130]. (b) Biassed probability distribution of the reaction coordinate(in this figure, Extension) for each window. x-axis and y-axis represent the reaction coordinate and corresponding probability value respectively.

lot of time may be spent transitioning from one minima to the other, resulting in their poor sampling. This lack of sampling signifies the generation of an inaccurate PMF. Even for small systems, simulation timescales in the order of a millisecond MD are required at the very least to obtain useful sampling. This is obviously not tractable, and so a smarter approach must be used.

A way to go about handling this would be to perform MD simulations using a bias potential exactly equal to $-F(s)$, where $F(\chi)$ represents the PMF along a reaction coordinate χ (if its already known, or accurately estimated that the free energy surface will have a functional representation of $F(\chi)$). This will make the biassed free-energy landscape flat and barrier less. This potential acts as an *umbrella* that helps in safe crossing of the transition state in spite of its high free energy.

Since it would be almost impossible to generalize functional estimation of the free energy profile along a reaction coordinate, if the location of the barrier, w.r.t. this reaction coordinate is accurately estimated, the sampling of this barrier could easily be favoured by addition of an artifical harmonic restraint on the CV, for e.g. in the form of the following potential:

$$\mathcal{U}(\chi) = \frac{k}{2}(\chi - \chi_0)^2$$

After applying such a bias, the sampled distribution will change from $\mathcal{P}(q)$ to $\mathcal{P}'(q)$, and these are related as:

$$\mathcal{P}'(q) \propto \mathcal{P}(q).exp\left[\frac{-k\,(\chi(q)-\chi_0)^2}{2k_BT}\right] \tag{2.25}$$

For large values of k, only points close to s_0 will be explored. By combining simulations performed with different values of χ_0, one could obtain a continuous set of simulations going from one minimum to the other, crossing the transition state(s). The ensemble in such a biassed sampling scenario would be known as **biassed** or **extended** ensembles, and will no longer remain a pure NVT/NPT/NVE ensemble.

If a very small force constant (k) of bias is used for the enhanced sampling, the system may still be unable to cross the aforementioned barriers. On the contrary, if a large force constant is used, the barrier will be crossed, but the sampling in turn suffers from extremely narrow and non-overlapping probability distributions. To tackle this, the resolution of the chosen windows needs to be finer, i.e., more intermediary windows need to be sampled. This may, however, lead to a manifold increase in the computational demand, and thus the selection of a balanced force constant paired with an apt resolution is required for optimising both sampling and computational demand.

There are numerous methods to obtain the actual PMF from umbrella sampling simulations. One of the most popular methods developed by a pioneer in the field, Shankar Kumar, is the Weighted Histogram Analysis Method [131]. The corresponding software tool was developed by another pioneer, Alan Grossfield [132]. Since it's a free-to-use software, it forms the perfect basis for young researchers to learn and understand umbrella sampling and obtain free energy profiles for a variety of reaction coordinates, thus garnering such a wide audience.

2.10.1 Weighted Histogram Analysis Method

This method, popularly known as its abbreviated form, WHAM, involves solving of the following 4 equations [133].

$$e^{-\beta F_i} = \int \mathcal{P}^u(\chi) e^{\beta \mathcal{U}(\chi)} d\chi \tag{2.26}$$

$$\mathcal{P}^u(\chi) = \sum_i^{N_w} p_i(\chi) \mathcal{P}_i^u(\chi) \tag{2.27}$$

$$p_i(\chi) = \frac{a_i(\chi)}{\sum_i^{N_w} a_i} \tag{2.28}$$

$$a_i(\chi) = N_i e^{-\beta \mathcal{U}_i(\chi) + \beta F_i} \tag{2.29}$$

2.26 was seen earlier on during the discussion of umbrella sampling, its the one used to calculate the free energy value for the i^{th} window, F_i, using the global unbiassed probability distribution of the

reaction coordinate, $\mathcal{P}^u(\chi)$, along with the biasing potential used, $\mathcal{U}(\chi)$. This global distribution is computed from a weighted average of the distributions of the individual windows, as given by 2.27, wherein p_i is the weight for the probability distribution $\mathcal{P}_i(\chi)$ for the i^{th} window and N_w represents the total number of windows.

The input equations 2.26 and 2.27 are solved under the following 2 conditions, in order to minimize the statistical error(standard deviation of global unbiassed probability distirbutions, σ) in $\mathcal{P}_i^u(\chi)$ [131]:

$$\frac{\partial \sigma^2 \left(\mathcal{P}^u(\chi)\right)}{\partial p_i(\chi)} = 0 \tag{2.30}$$

$$\sum_i^{N_w} p_i(\chi) = 1 \tag{2.31}$$

This eventually leads to the obtaining the equations 2.28 and 2.29. The N_i in 2.29 represents the number of steps sampled for window i. Its easily observable that for calculating 2.29, the values from 2.26 need to be borrowed, hence these equations are to be computed iteratively, in a self-consistent manner, until convergence is achieved.

2.11 Simulatory Optimization Tactics

2.11.1 Implementing Ensembles - Canonical ensemble via thermostat

Although MD simulations at high temperatures are able to circumvent the inadequate sampling of the high-energy barrier scenario, the computational time itself at such considerably high temperatures is quite large, especially when utilising explicit solvent. However, engaging Langevin dynamics at realistic temperatures could be used as a successful alternative [134]. Since the solvent collisions make the dynamics stochastic, Newton's equations could be well replaced with Langevin's as the following:

$$m\frac{d^2x}{dt^2} = F(t) - \zeta\frac{dx}{dt} + R(t) \tag{2.32}$$

$$\langle R(t)R(t')\rangle = 2m\gamma kT\delta(t-t') \tag{2.33}$$

where, x is the position, m is the reduced mass, ζ is the friction constant, F is the systematic force (represents the real potential) , and $R(t)$ is a random force, assumed to be uncorrelated with the positions and velocities of the particle and is rather a Gaussian function with a mean of zero and variance given by 2.33. The collision frequency γ in 2.33 is defined as $\frac{\zeta}{m}$. Newtonian equations are recovered when $\gamma = 0$. The random force represents the thermal kicks from the small solvent-particles. The friction constant and the random force combine to give the correct canonical ensemble.

When choosing a timestep δt both the highest frequency of the systematic force and the collision frequency γ must be taken into account [135, 136]. The coordinates and velocities are updated according to the following steps:

1. fetch intermediate velocities as : $v(t+\frac{\Delta t}{2}) = v(t-\frac{\Delta t}{2}) + (x_n - x_{n-1})\,\frac{1-\gamma\frac{\Delta t}{2}}{1+\gamma\frac{\Delta t}{2}} + \frac{(F_n+R_n).(\Delta t)^2}{m.\left(1+\gamma\frac{\Delta t}{2}\right)}$
2. Calculate updated coordinates using intermediate velocities, x(t + Δt) = x(t) + v(t + $\frac{\Delta t}{2}$)
3. Get the velocities for the new time-step: $v(t+\Delta t) = \left(1+\gamma\frac{\Delta t}{2}\right)^{1/2}\left(v(t+\frac{\Delta t}{2}) + v(t-\frac{\Delta t}{2})\right)/2\Delta t$

2.11.2 Implementing Ensembles - Isothermal-isobaric ensemble via barostat

The commonly used isochoric and adiabatic ensembles that maintain a constant volume of the given system cannot be employed if the system being studied is characterised by a continuous loss of energy. This method was initially used to simulate single-component fluids and study phase transitions. Andersen was one of the first to propose a pressure-constraining method [137] for simulations, and its implementation by Haile and Grabben [138] yielded static properties that corresponded closely to (N, V, E) simulations, but the dynamic properties remained untested. A large array of pressure-constraining methods were suggested subsequently, which dealt primarily with modifying the Hamiltonian [139–142]. The Berendsen barostat method was finally brought forth in 1984.

This method weakly couples a system to be simulated within this ensemble with an external bath based on the theory of *least localised perturbation*, thus fulfilling any required systemic coupling. The impact of this coupling can be adjusted and examined by controlling its strength. This ensures the presence of any perturbation that is naturally found within a system in an out-of-equilibrium state.

This ensemble could be implemented using the method of weak coupling to a thermal bath proposed by Berendsen [143]. It proposes to add an extra term to the equations of motion which effects the pressure change:

$$\left.\frac{dP}{dt}\right|_{bath} = \frac{P_0 - P}{\tau_P}$$

where P_0 is the reference pressure, i.e. the pressure of the external bath, P is the instantaneous pressure and τ_P is a time constant. The pressure is given by:

$$P = \frac{2}{3V}\left(E_k - \Xi\right) \tag{2.34}$$

$$\Xi = -\frac{1}{2}\sum_{i<j} \mathbf{r}_{ij}.\mathbf{F}_{ij},\ \ \mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j \tag{2.35}$$

Here, P is the pressure, V is the volume, Ξ represents the virial of the pairwise additive potentials, and $\mathbf{F}_{ij}$ represents the force acting on particle i due to particle j. Within this scheme the coordinates and

the box sides are rescaled at every step, in order to maintain the constant pressure. Assuming the system is isotropic and within a cubic box the scaling factor μ is given by:

$$\mu = 1 - \frac{\kappa_T \Delta t}{3\tau_p}(P_0 - P)$$

κ_T is the isothermal compressibility. In theory, an inaccuracy in the value of isothermal compressibility only influences the accuracy of the non-critical time constant τ_P, its not consequential to the precision of the simulation. In reality, the value of κ_T should be reasonable, and may depend upon the simulation package used, for instance GROMACS and DL POLY use $\kappa_T = 4.6 \times 10^{-5} \text{bar}^{-1}$) at P_0 = 1 atm and T_0 = 300K.

The following steps would account for a brief employ of this barostat in an MD simulation:

1. Evaluate all forces on all atoms, $a_i(t) = \frac{F_i(t)}{m}$
2. Evaluate the virial and the kinetic energy, thus measuring the pressure, using 2.35 and 2.34.
3. Compute the intermediary-timestep velocities, i.e. $v\left(\frac{t + \Delta t}{2}\right)$
4. Evaluate the new atomic-coordinates, using the intermediate velocities computed in the preceding step and pressure-scale them by taking a product with μ .
5. Calculate the velocities for this new timestep.

2.11.3 Periodic boundary conditions

Commonly referred to as PBC, these involve periodic duplication of the system in all directions to represent an infinite system. A cubic lattice is typically used, with the central box housing our main system and other replica-cubes containing atoms that are the images of the central-box atoms. The coordinates of atoms inside these replica boxes can be obtained simply by adding or subtracting integral multiples of the central box dimensions. A particle leaving the box during simulation is replaced by an image particle that enters from the opposite side of the box, as shown in Figure 2.7. Hence, the net number of atoms in the central box remains constant.

On employing PBC, it's impossible to have fluctuations of wavelengths greater than the edge length of the cell. For simulations of systems with a *near liquid-gas critical point* this could cause some problems. The cell size should be larger than the range of interactions, or else some pseudo-long-range (mainly electrostatic) interactions may get imposed unnecessarily.

Apart from cubic systems, the rhombic dodecahedron [144] and truncated octahedron [145] can also employ PBC. The added benefit is the reduction in the number of solvent atoms required, resulting in a decreased computational load. For a given number of atoms, the distance between adjacent cells is

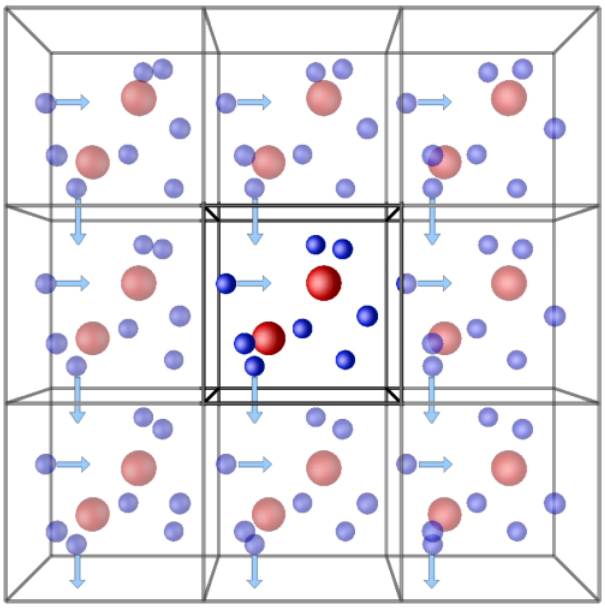
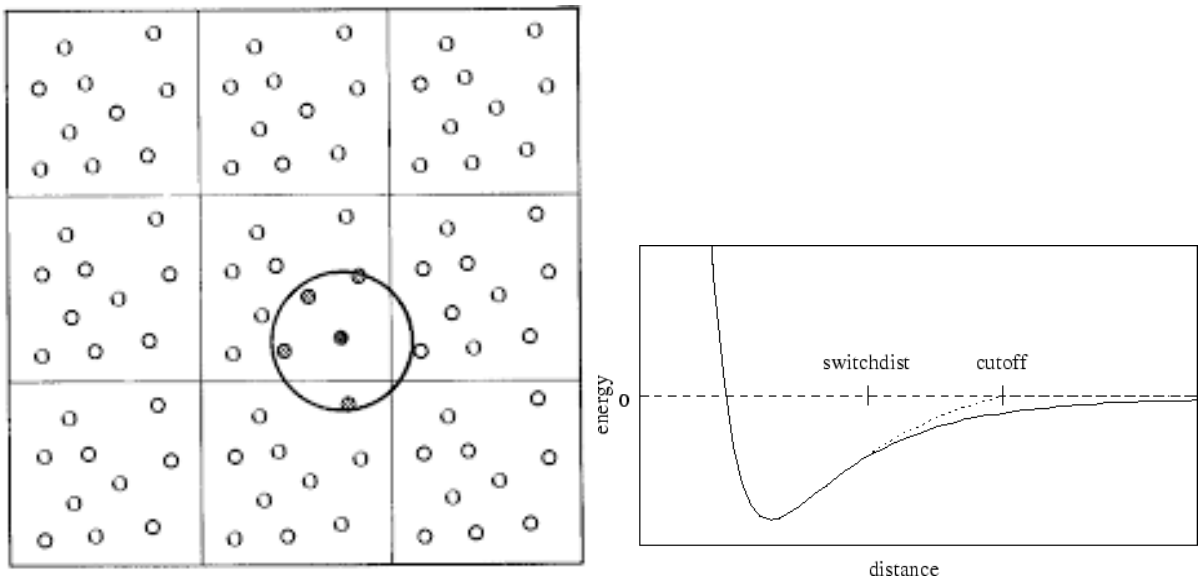


Figure 2.7: Periodic Boundary conditions in 3-D. The red and blue spheres represent atoms of two different elements. Arrows drawn on some of them represent their current direction of movement. Notice that in each image of the central box, the direction of movement of a particular atom and its corresponding images is the same. Image adapted from Central Michigan University.

larger for such cell types when compared to the cubic cell, so they require fewer solvent particles when compared to the cubic cell. Additionally, it's often sensible to pick a cell type with a shape that reflects the underlying geometry of the system. In the absence of periodicity, stochastic boundary conditions can be used for any system geometry [146].

Stochastic boundary conditions, a lesser-known boundary conditions tactic, are highly useful for local exploration of a system, such as a binding site. Enabling them allows the exclusion of a major chunk of the system from being simulated, thus saving considerable computational resources. A spherical, stochastic shell is said to enclose the local region and is characterised by stochastic dynamics, which can be evaluated using Langevin dynamics equations. A bath enclosing this stochastic shell contains and is responsible for maintaining the structural integrity of the rest of the system. Although proteins such as BPTI (bovine pancreatic trypsin inhibitor) have been studied under this approach [147], artificial density fluctuations that arise as a result of such constricting boundaries, even in the simplest of models, can alter the solvent-structure [148]. However, some models with improved features have been explored [149, 150].

Periodic boundary conditions are imposed with what is known as the *nearest image convention*.



(a) (b)

Figure 2.8: **Nearest image convention and potential truncation.** (a) The nearest atoms/copies wil be used for potential calculations, those which lie inside the circle. (b) The actual vanderwaal's interaction potential function used switches to a linear decay function when the interatomic separation reaches the *switchdist* value, and becomes 0 on reaching the cutoff.

2.11.4 Nearest image convention

Systems to be simulated usually have around 10^5 - 10^6 atoms. The computational load to compute pairwise non-bonded interactions scales as N^2, N being the total number of particles in the system, somewhere around 10^{10} - 10^{12}. For a large fraction of atoms, however, since the interatomic separation is considerably high, the theoretical value of non-bonded interactions would remain somewhere around zero, as suggested from equation 2.13. To compute such truncating non-bonded interactions, the PBC have to be augmented with what is known as the *nearest image convention*.

In this convention, each individual particle in the simulation interacts with the closest image of the remaining particles in the system. A radial cutoff is applied w.r.t. each atom, and all the atoms/nearest-images having separation greater than this cutoff are said to have zero interactions, and only the rest are assumed to contribute to the total interaction energy of the system. This cutoff should be large enough for reasons of accuracy but small enough to not include an interaction between a particle and its own image. This cutoff is also influenced by other systemic factors, such as the central-box shape and size of atoms. A cutoff of 2.5σ , where σ is the force constant seen in the Leonard-Jones potential equation, is known to produce a relatively small error. On the contrary, for long-range interactions such as coulombic interactions, using such a cutoff may produce considerable errors, depending on the system being simulated.

2.11.5 Neighbouring list

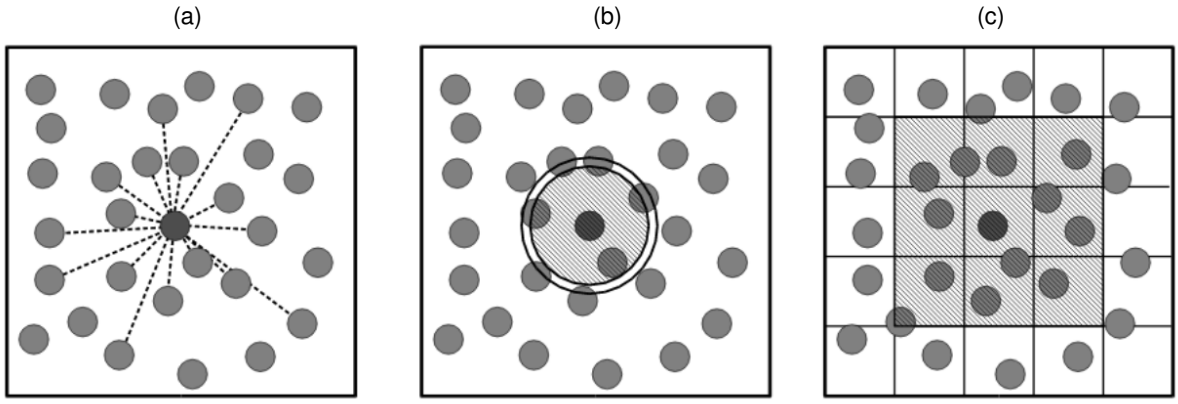


Figure 2.9: **Approaches to compute pairwise interactions.** (a) Consider all pairs, (b) using neighbour lists, where the concentric circles show the interaction range and the extra area covered by the neighbor list for one of the atoms. (c) Cell list method, the edge length exceeds the interaction range. All images have been borrowed from [151].

Although using the cutoff exempts the energy calculation for atoms of a certain interatomic separation, nonetheless, the separation still gets calculated, thus still increasing the computational time by an order of N^2. If there was a way to know a list of atoms to consider while calculating non-bonded interactions, a considerable amount of time would be saved by omitting the separation calculations for the rest. The first-ever list was proposed by Verlet, which went by the name *Verlet list* [152]. The interatomic separation for all pairs of atoms is initially calculated, and only those pairs are stored whose separation is under another cutoff called the *neighbour cutoff*.

Initialising this neighbour list although takes $O\left(N^2\right)$ computational load, is a one-time investment since the neighbour list is updated at regular intervals throughout the simulation, and the updates occur in a linear time. To reflect the change in the interatomic separation from a value greater than the neighbour cutoff to a value lesser than the non-bonded cutoff within the time interval of the update for the neighbour list, the neighbour cutoff has to be larger than the non-bonded cutoff. The use of the verlet list results in a net drop in the computational load from the initial $O\left(N^2\right)$ to $O\left(N^{\frac{5}{3}}\right)$.

The two main hyperparameters that tread on the efficiency-accuracy tradeoff are the update frequency and the neighbour cutoff value. The value of the update frequency is important since too high a frequency will increase the computational load, and too low a frequency may result in an incorrect calculation of PEF and thus forces due to the atoms moving within the non-bonded cutoff region, resulting in an inaccurate simulation. A larger neighbour cutoff would cause a surge in storage complexity, and a smaller value would result in the aforementioned inefficiencies.

Another, more efficient technique of optimising pairwise interaction calculations would be the use of a *cell list*. It is a data structure used to find all atom-pairs separated by a distance that is well within a

given cut-off distance. On dividing the simulation box into cubic cells, each with an edge length greater than or equal to the non-bonded cutoff, the particles are sorted into these cells, and the interactions are computed between particles in the same or neighbouring cells. This entire procedure is known as the cell index method. It brings down the computational time from the typical $O\left(N^2\right)$ to $O(N)$, thus an improvement over the neighbour list method.

2.11.6 Ewald sums

Interactions that tend to be considerable even at large values of interatomic separation are highly problematic to simulate. Such interactions need to be modelled with the utmost care since they affect fundamental properties such as the dielectric constant of the system. Thus, a variety of methods were developed to handle these long-range forces.

To study the energetics of ionic crystals, the Ewald sum came to light [153]. The short-range interactions are calculated in real space, whereas the long-range interactions are calculated using Fourier transforms. For the long-range interactions, instead of calculating point-charge interactions, a neutralising Gaussian charge distribution is functionally added to the actual system in the real space. Another distribution, to neutralise this, is added in the reciprocal lattice space. Thus, the amount of charge remains the same, and the initially conditionally convergent series of long-term interactions expressed as point-charges is now decomposed into two completely convergent series.

Since it rapidly converges in comparison to direct summation, it is the de facto standard method for long-range interactions when employing PBC. Time complexity of the method is still $O\left(N^2\right)$ due to the bottleneck calculations of the reciprocal space, but by using fast-Fourier transforms to handle the reciprocal space contribution, the time complexity can be reduced down to $O\left(N \ln N\right)$ where $\ln$ is the natural logarithm, i.e. logarithm to the base e. Nonetheless, it is still laced with certain limitations. Each inter-charge interaction decrements when the interatomic separation is half the central-box edge-length. Since the summation method considers the central and image boxes, rapid conformational changes in the former are reflected in the latter instead of fading away. All of these can be bagged into the Ewald method, inflating the errors introduced as a result of using PBC.

2.11.7 Bond-parameteric constraints

From a fixed computational power standpoint, factors such as the quantity of interactions to resolve per time-step, time spent extracting each interaction, the period of the time-step, and the total degrees of freedom to be handled affect the actual time a simulation takes. By decreasing the total degrees of freedom to be handled in the form of freezing high-frequency normal modes, the simulation efficiency could be incremented. This is usually done by constraining parameters of covalent bonds that involve hydrogen atoms using algorithms such as SHAKE [154, 155], RATTLE [156] and LINCS [157]. Hence

the current highest-frequency vibration has a frequency value much lower than these hydrogen-bond involving modes. Since the timestep has to be marginally smaller than the time-period of this highest-frequency mode, it can now be made larger, thus allowing for longer simulation time-scales. In practice, the time step can typically be increased by a factor of 3 compared to simulations with the original Verlet algorithm.

The SHAKE algorithm is an improvement over the Verlet algorithm in that the velocities of atoms involved in covalent bonding with hydrogen (and the hydrogen atoms themselves) are updated so as to constrain the corresponding bond lengths and bond angles at their respective equilibrium values. In LINCS, instead of altering the velocities, the atomic positions themselves are reset. This avoids a statistical error innate to the SHAKE method and enables an increment in the time-step value by a factor of four. Generalised SHAKE [158] adds support for general nonholonomic constraints, and no numerical drift is observed even when the number of constraints is large.

2.12 Need for Computational Studies

Before the advent of computers, one had to write force balance equations describing the equilibrium of forces and volume constancy equations and solve them by hand. This procedure wasn't scalable since the number of equations and variables involved in a realistic, single- or multi-molecule system were too many to keep track of, thus rendering this method impractical. To make matters worse, as the molecule(s) moved, their geometries changed, and the equations had to be re-derived and re-solved for each small increment of motion.

In the early 1970s, computers were first introduced in universities for research purposes. Sooner, it was discovered that code written and executed on them could be used potentially to solve the dense equations of molecular dynamics (and other sciences for that matter). This evolved into studying interactions between multiple biomolecules in one another's vicinity, which soon morphed into studying cell movements themselves. Computational algorithms experienced improvements in performance and reliability with the passage of time, bridging the gap between experimental and model-based studies. The extent of improvements in recent times has led to computational models being considered a reliable option when experimental studies yield less favourable results.

Evaluation of alternate scenarios via variation of the most important parameters opens up newer directions in which a biological system could be looked at, drawing more insights regarding the process(es) that it is involved in. Problems for which an analytical solution either does not exist or is very complex to arrive at can be solved, or at the very least, a leap towards the solution can be made by approaching them from a simulatory perspective. Simulations are also useful to assess the behaviour of a system in a nearly-real environment when a lack of experimental data regarding the same makes it difficult to study the system.

The existence of complicated and buggy computational software creates obvious doubts regarding the results derived from it being trustworthy. This also demands checks on the software being used at various levels and simpler ways to convey the physical meaning of the computational methods employed and the models used to peer reviewers. A fundamental problem with computational methods is that all such software is generally filled with complexities, even in the smallest of details, and a detailed understanding of the same is usually reserved for its authors.

A first order evidence is an evidence which by itself is enough to enunciate a hypothesis [159]. Usually experimental results of a study on a biomolecular entity are classic examples of first order evidence, the hypothesis being that the molecular entity participates in some process that is the objective of the study. Higher-order evidence, on the other hand, is evidence that leads to first-order evidence being real and important for some hypotheses [160, 161]. Results from simulations and observations can be constituted as higher order evidence. These generally need to match experimental results, or else either the simulation or the experiment, or both of them, could be wrong.

A computer simulation system is reliable in a domain of application if, and to the extent that, the majority of results that it would produce in that domain are true (or are accurate enough, given an agent's tolerance for error), when interpreted as claims about the world. If it yields very reliable results that are accurate to within a specified margin of error, then its result(s), attributing to some particular target feature of the biomolecular system being studied, can be evidence for a proposed hypothesis for the same. A sequence of steps could qualify as evidence for a hypothesis when represented as mathematical equations solved by hand that are reliable enough to accurately study a particular mechanism or reaction. Computer simulations are nothing but solving these equations via computers, using some approximations so as to ease out the calculations.

Chapter 3

Allosteric Response of DNA Recognition Helices of Catabolite Activator Protein to cAMP and DNA Binding

Contents

3.1 Introduction

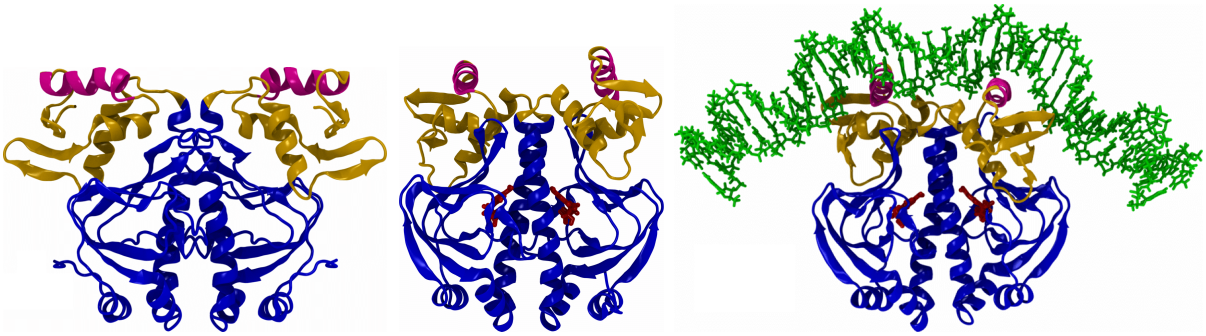
The modulation of proteomic activities across the cell can be influenced by the execution of gene transcription processes [162–166]. Molecules known as *transcription factors* (a.k.a. TF) are primarily responsible for impelling the gene-regulatory network [167, 168]. Comprehending the intricate processes within this network involves evaluating elements that impact the individual behaviors of transcription factors and DNA, as well as their binding dynamics [169–174].

Among these important TFs is a protein called **Catabolite Activator Protein**, hereafter referred to as CAP. Found in bacteria, it administers the metabolism of different organic macromolecules by responding to fluctuations in the cellular concentration of another molecular species called *cyclic adenosine monophosphate* (cAMP) [56–61].

CAP, a dimer of approximately 50 kDa, is constituted by two identical subunits of around 209 residues each. Each subunit consists of two unique domains: (i) an N-terminal domain (residues 1 - 136) responsible for cAMP binding, and (ii) a DNA-binding domain (residues 138 - 209) situated at the C-terminal, which incorporates a helix-turn-helix structural motif for DNA interaction [69–75]. The two domains are linked through a brief hinge region (residues 137-138). The cAMP-binding domain encompasses a pocket area for binding cAMP, allowing for the accommodation of two cAMP molecules due to the protein's dimeric nature, with one molecule per monomeric subunit. CAP manifests in two primary configurations: one with weak DNA binding and another where the affinity is markedly stronger. The introduction of cAMP results in an allosteric alteration, inducing a transition from the former, termed the *inactive* state, to the latter, termed the *active* state [62–68]. cAMP-bound CAP attaches to DNA at sites proximate to the target promoter region, thereby regulating the interactions between RNA polymerase and the target promoter, initiating the transcription process [56, 175, 176]. Since the influence of CAP's regulatory mechanism cascades into subsequent downstream biological processes, achieving a fundamental comprehension of this *allosteric shift* holds utmost significance [177–179].

MD simulations [180–185] and site-targeted NMR experiments [69, 76] have lately helped in unsheathing the intricacies of the allosterically mediated pathways involving regulatory sites and residues that are essential for protein allostery. Kalodimos et. al employed NMR and isothermal titration calorimetry experiments [76] to exhibit negative cooperativity (cAMP binding at one site decreases the binding affinity of the other cAMP molecule for the other binding site) with regards to the binding of the two cAMP molecules and the changes in protein motion. These changes mainly comprise of a coil to helix transition of the VAL126 to PHE136 segment in the CBD and conformational changes in the recognition helices (commonly known as F-helices) of the DNA, wherein they experience a net translation and rotation of ≈ 7 Å and $\approx 60^{\circ}$ respectively, relative to their orientation in the ligand-free CAP [69].

Establishing extensive connections with DNA base pairings and the sugar-phosphate backbone is promoted by these alterations in configuration, aiding the favorable docking of an incoming target DNA sequence onto these F-helices [69, 186]. The present study aims at understanding the cAMP- and DNA-induced energetics behind these changes by exercising molecular dynamics and two-dimensional umbrella sampling simulations on cAMP-free, cAMP-bound, and DNA-bound CAPs. A set of unique collective variables is proposed to capture the key interactions between CAP, DNA, and the cAMP molecules, and the corresponding free energy profiles are analysed to grasp their energetic perspectives.



(a) (b) (c)

Figure 3.1: Crystal structures of (a) unliganded CAP (Apo-CAP), (b) CAP bound with cAMP (CAP-cAMP) and (c) cAMP liganded CAP complexed with DNA (CAP-cAMP-DNA). The CBD (cAMP-binding domain) on the catabolite activator protein is coloured in blue, whereas the DBD (DNA-binding domain) is coloured in orange. The α-helices of interest in this study are coloured in magenta and belong to the DBD. The ligand (cAMP) is in red, and the complexed DNA in green.

3.2 Simulation Details

3.2.1 Models

The following three states of CAP were chosen for the present study as follows: (a) the ligand-free CAP (henceforth referred to as the apo-CAP), (b) the cAMP-bound state, wherein two cAMP molecules are bound to CAP (CAP-cAMP complex), and (c) the DNA-bound state, wherein cAMP-bound CAP is further bound to the DNA (CAP-cAMP-DNA complex). Solution NMR and X-ray crystal diffraction-determined PDB structures were chosen as an initial conformation for the 3 states, 2WC2 for apo-CAP [69], 1G6N for CAP-cAMP [71] and 1ZRC for CAP-cAMP-DNA [74]. The missing terminus residues of CAP for 1G6N (1-6, 207-217, 417-418) were modelled using Modeller [187]. A pictorial representation of the important domains for these 3 structures is rendered in Figure 3.1.

In addition to the aforementioned models, the following three models that represent the meta-stable intermediate conformational states were also investigated: (a) the cAMP-bound form of CAP with both the cAMP molecules removed from their respective cAMP-binding domains (henceforth referred to as the CAP-cAMP* model), (b) the DNA-bound form of CAP-cAMP-DNA with DNA removed from the DNA-binding domain (the CAP-cAMP-DNA* model), and (c) a composite system generated by docking the DNA molecule from the CAP-cAMP-DNA system onto the CAP-cAMP complex from the original CAP-cAMP system (the CAP-cAMP$\cdots$DNA model).

The pre-processing steps iterated for each model comprised: (a) addition of missing hydrogen atoms; (b) solvation in a cubic TIP3P water box of apt dimensions; and (c) the addition of Cl^- counterions to neutralize the system.

3.2.2 Molecular Dynamics Simulation

200 ns MD simulations of all six systems were performed using the AMBER 2016.10 simulation package [108] with the ff14SB force fields [188] for the protein, ParmBSC1 force fields [189] for DNA, and the TIP3P model [190] for water molecules. Force fields defined in [191] were used for cAMP molecules. The SHAKE algorithm was used to constrain the length of bonds involving hydrogen atoms [155]. Periodic boxes of the following dimensions were used: 115 Å $\times$ 97 Å $\times$ 78 Å for apo-CAP, 90 Å $\times$ 107 Å $\times$ 101 Å for CAP-cAMP, 108 Å $\times$ 94 Å $\times$ 143 Å for CAP-cAMP-DNA, and 90 Å $\times$ 90 Å $\times$ 90 Å for the remaining 3 systems, as part of employing periodic boundary conditions. The Particle Mesh Ewald method was used for evaluating long-range electrostatic interactions, with a 10 Å cutoff and tolerance of 0.00001 . A direct cutoff of 10 Å was used for the van der Waals interactions. A Berendsen barostat [143] was employed to maintain isobaric conditions 1 bar with a time constant of 1 ps, while a Langevin thermostat [134] kept the temperature at 300 K using a time constant of 1 ps. The velocity Verlet [121] algorithm was used to integrate the equations of motion with a time step of 2 fs.

Potential energy minimization for the entire system was split into 2 phases: the first involved restraining the heavy atoms of the solute (protein, cAMP, DNA) with a high force constant in an effort to relax all hydrogen atoms, and the second phase comprised a restraint-free minimization. Each phase was carried out by an initial 1000 cycles of the steepest descent algorithm and a subsequent 1500 cycles of conjugate gradient minimization. The structure obtained at the end of phase-2 was introduced in an NVT ensemble for annealing to 300 K for 20 ps, restraining the heavy atoms of the solute(with a harmonic spring constant of 10 kcal mol $^{-1}$ Å^{-2}). The Maxwell-Boltzmann velocity distribution corresponding to 300 K was used to assign random atomic velocities to all particles of the system. The annealed structure was brought into an NPT ensemble with the temperature and pressure held at 300 K and 1 bar respectively. This simulation was carried out for 2 ns with the positional restraints still intact. Each system underwent a 3 ns NPT simulation at 300 K and 1 bar pressure in the conclusive phase, after which all positional restraints were removed. This was followed by a 200 ns NPT production run with the same temperature and pressure values.

3.2.3 Umbrella Sampling

3.2.3.1 Collective Variables to Analyse Relative Motion of F-helices

As mentioned previously, exploring the underlying dynamics and energetics of the orientational changes of F-helices [69] on cAMP binding was the prime motto of this study. Therefore, the systems CAP-cAMP*, CAP-cAMP-DNA*, and CAP-cAMP$\cdots$DNA are expected to transition to apo-CAP, CAP-cAMP and CAP-cAMP-DNA, on simulating them as mentioned in the **Simulations** subsection. However, the time-averaged structures of CAP in these 3 systems are found to be comparatively more structurally similar to the CAP in CAP-cAMP, CAP-cAMP-DNA, and CAP-cAMP, respectively. The

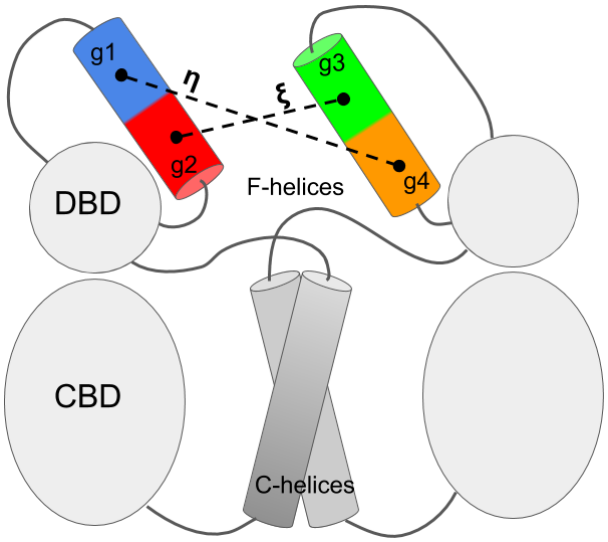


Figure 3.2: Collective variables η and ξ used to describe the relative motion of F_α-helices of CAP. The centers of mass (COM) of the four chosen groups on the two F_α-helices - g1 (blue), g2 (red), g3 (green) and g4 (orange) are shown as black dots. η and ξ are represented by dotted lines connecting the COM of (g1, g4) and (g2, g3) respectively.

rugged nature of the underlying energy surface can vouch for the fact that the time scales of the expected rare, transitionary events appear to be longer than those of the MD simulations used by the authors. Thus, it is practically infeasible to capture these binding-induced large-scale structural changes in CAP using conventional MD simulations. Thus comes the need for employing enhanced sampling techniques such as Umbrella Sampling.

In an effort to aptly capture structural changes (rotational and translational) in F-helices resultant of binding, two distance-based collective variables (CVs) η and ξ were defined, as shown in Figure 3.2. To define these, F-helices were first divided into 4 groups: groups 1 and 2 consisted of the C-α atoms of residues 186-191(upper half) and 179-185 (lower half), respectively, of the first F-helix, while the C-α atoms of residues 179'-185' (upper half) and 186'-191' (lower half) of the second F-helix formed groups 3 and 4, respectively (in the above and henceforth, the superscript ' denotes the residues of subunit-2 of CAP). The distance between the centres of masses of groups 1 and 4 was labelled to be η, and the distance between the centres of masses of groups 2 and 3 was referred to as ξ.

3.2.3.2 Simulation Parameters

Using the definitions of η and ξ as reaction coordinates, two-dimensional (henceforth referred to as 2D) umbrella sampling (US) simulations were performed on the systems Apo-CAP, CAP-cAMP and CAP-cAMP-DNA. This captured the variations in the potentials of mean force (PMFs) for the relative transformations of the two F-α helices. The resultant reaction coordinate sampling was subjected to

weighted histogram analysis to formulate the PMF profile. For each of these 3 models, the starting structure of their US runs was taken from the final timestep of their respective unbiased production runs. The collective variable η was varied from 28 Å to 48 Å, and ξ was varied from 18 Å to 35 Å, in steps of 0.5 Å for each of them. Thus a total of $101 \times 35 = 3535$ unique configurations or *windows* were simulated, per system. For each system, the starting structure underwent potential energy minimization to bring η and ξ to the centre of the chosen window using a harmonic biasing potential with a higher spring constant (100 kcal mol^{-1} Å^{-2}) on both the collective variables. Following this, in each window, the minimized structure was equilibrated for 100 ps followed by 2 ns of production run at 300 K and 1 atm pressure in an NPT ensemble. These umbrella sampling simulations were carried out using the same conditions as those of unbiased MD runs, but with η and ξ being restrained by a harmonic biasing potential of force constant 4 kcal mol^{-1} Å^{-2}.

To preserve the orientational sanctity of the F-α helices from the effects of the applied bias, the distances between all pairs of heavy backbone atoms in each of these helices were harmonically constrained using a spring constant of 2.5 kcal mol^{-1} Å^{-2}. This ensures the structural integrity of these helices and eliminates their internal motions during the PMF evaluation. Taking into account the 200 ns unbiased production runs per system (6 systems) and all the umbrella sampling runs (3 systems), the gross computational time of these simulations was $\approx 9.2\mu$s.

3.3 Results and Discussion

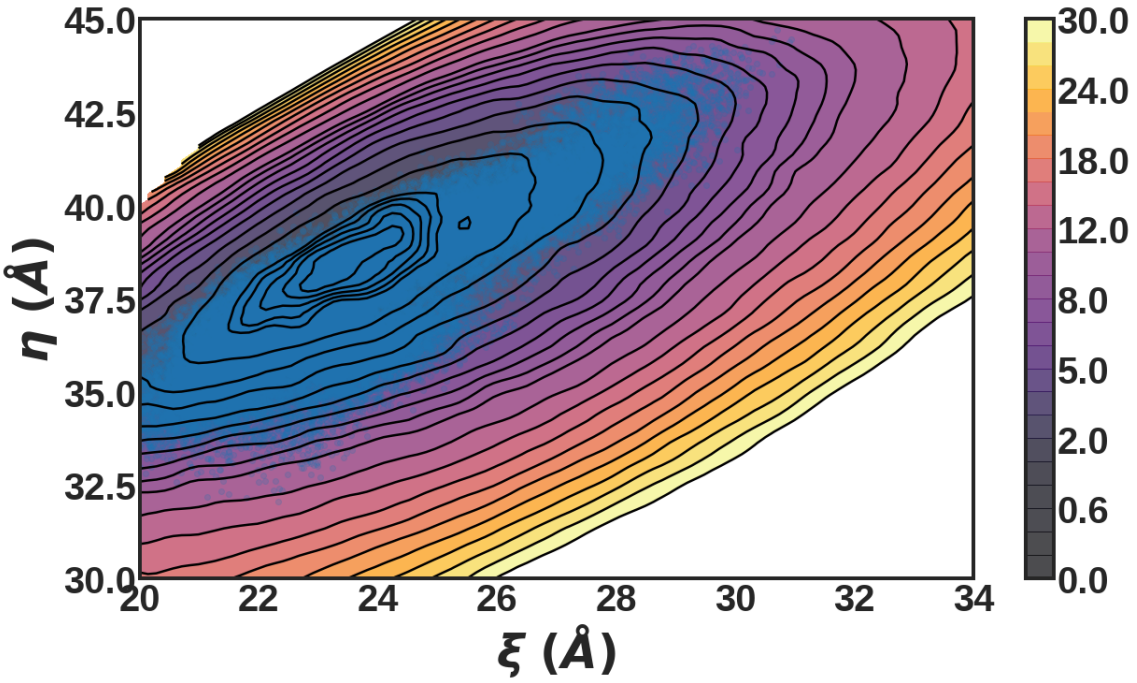
3.3.1 Free Energy Profiles

Figure 3.3 shows the 2D free energy profiles (henceforth referred to as FEP) from umbrella sampling calculated as per the details mentioned in the previous section, (η, ξ) indicates a unique point on the 2D contour plot that represents the FEP. The subscripts U and C are used for the unbound and cAMP-bound CAP respectively, thus $F_U(\eta, \xi)$ (ref. Figure 3.3a) and $F_C(\eta, \xi)$ (ref. Figure 3.3b) denote the free energy contours for Apo-CAP and CAP-cAMP respectively. Analysis of the effect of DNA-binding has already been discussed in [192], thus effects of the cAMP-binding event are analysed here. Hence, only the PMFs of Apo-CAP and CAP-cAMP are shown.

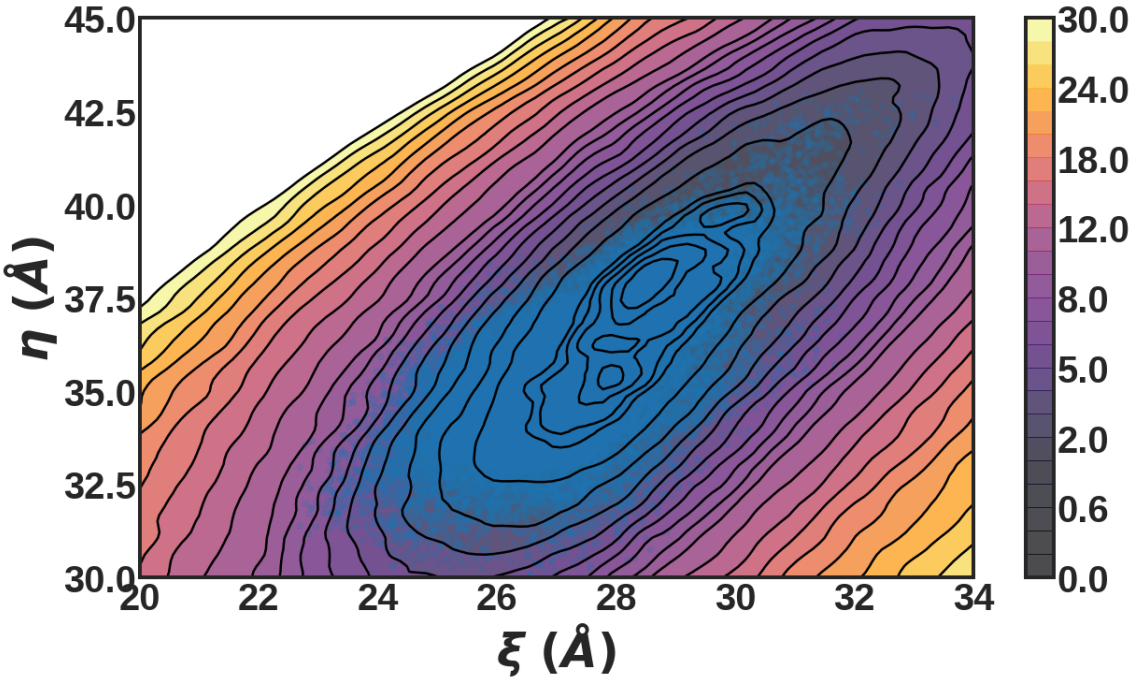
The global minimum of $F_U(\eta, \xi)$ is located at (38.8 Å , 23.9 Å). The elliptical contour lines seen in Figure 3.3a display a greater spacing along ξ when compared to that of η. The width of an energy basin of a system along η(denoted by $\delta\eta$) and ξ(denoted by $\delta\xi$) was devised as a numerical measurement of its freedom along these reaction coordinates, and are calculated as follows:

$$\delta\eta = \eta_{max} - \eta_{min} \ , \ \delta\xi = \xi_{max} - \xi_{min}$$

Subscripts "max" and "min" denote the maximum and minimum values, respectively, of that particular reaction coordinate along the contour line of free energy equal to 6 kcal mol^{-1} , for the FEP of



(a)



(b)

Figure 3.3: Calculated two-dimensional PMF profiles as a function of η and ξ for (a) apo-CAP, denoted as $F_U(\eta, \xi)$ and (b) CAP-cAMP complex, denoted as $F_C(\eta, \xi)$. The contour plots are truncated to leave out high-energy regions ($>$ 30 kcal mol^{-1}) on these PMF profiles. The reported energies are in units of kcal mol^{-1}. The distribution of η and ξ obtained from the unbiased MD simulations is superimposed on the PMF plots with translucent light-blue circles.

a particular system. The region demarcated by this free energy value is considered to aptly engulf the spread of η and ξ, obtained from unbiased MD simulations (blue data in Figure 3.3). For $F_U(\eta, \xi)$, $\delta\eta \approx 10$ Å and $\delta\xi \approx 13$ Å.

The $F_C(\eta, \xi)$ is characterised by a comparatively larger basin that $F_U(\eta, \xi)$, and its minima are located at (37.7Å, 28.8 Å) and (35.3 Å , 28.0 Å). This dual minimum indicates a back-and-forth existence of the CAP-cAMP complex in these conformations. The value of (η, ξ) for the crystal structure of the complex corresponds to (37.56 Å , 29.52 Å) which is situated well under the recorded minima of $F_C(\eta, \xi)$. When compared to $F_U(\eta, \xi)$, the elliptical contour lines seen in Figure 3.3b, i.e. $F_C(\eta, \xi)$ display a greater spacing along η than that along ξ, with the $\delta\eta$ and $\delta\xi$ values for CAP-cAMP being 15 Å and 11 Å respectively. A prominent shift in the energy minimum caused by cAMP-binding is observed and denoted along the reaction coordinates η and ξ as $\Delta_{U\rightarrow C}\eta$ and $\Delta_{U\rightarrow C}\xi$, respectively. The values of $\Delta_{U\rightarrow C}\eta$and $\Delta_{U\rightarrow C}\xi$,are -1.1 Å and 4.9 Å respectively. Therefore, a switch in the rigidity along the reaction coordinates observed from the PMF comparisons could be an effect of the resultant conformational changes of the cAMP-binding event, thus commissioning the binding of CAP with the promoter region on the DNA.

Regarding the unbiased MD-derived scattered spread of (η, ξ) of a system (blue data in Figure 3.3), it is observed that this spread is highly dense and uniform inside the region demarcated by the 6 kcal mol^{-1} contour line for both Apo-CAP and CAP-cAMP. The consistency of the free energy profiles with the unbiased simulations for either of the systems is evident from the spread being directed more along the major axis of the ellipsoidal free-energy contours.

3.3.2 Key Interactions Between Protein (CAP), Ligand (cAMP) and DNA

Changes in the reaction coordinates η and ξ, during the cAMP-binding event reflect interactional changes within the system from this event. Analysing potential energy changes associated with the event serves as a way of unveiling hidden features in the 2D PMF profiles corresponding to the event. Thus, electrostatic and van der Waals interaction energies of each amino acid of CAP with water, cAMP, DNA, and other amino acids of the protein itself were evaluated using the *linear interaction energy (lie)* tool available in the CPPTRAJ module [193] of AMBER16. U_U and U_C denote the potential energy of the protein CAP in Apo-CAP and CAP-cAMP respectively, and $\Delta U_{U\rightarrow C} = U_C - U_U$ denotes the change in interaction energy of the protein on the occurrence of the cAMP-binding event. Superscripts "intra" and "water" denote the interaction energy calculated for an amino acid of the protein with itself and with water respectively. Figure 3.4 depicts the cAMP-binding-induced changes in these interaction energies. Interactional effects of DNA-binding were already studied in [192], thus cAMP-binding effects will be the subject of discussion.

Although CAP is composed of two subunits that have identical structure and amino-acid sequence, the cAMP-induced change in their interaction energies with the environment begs to differ, as is seen

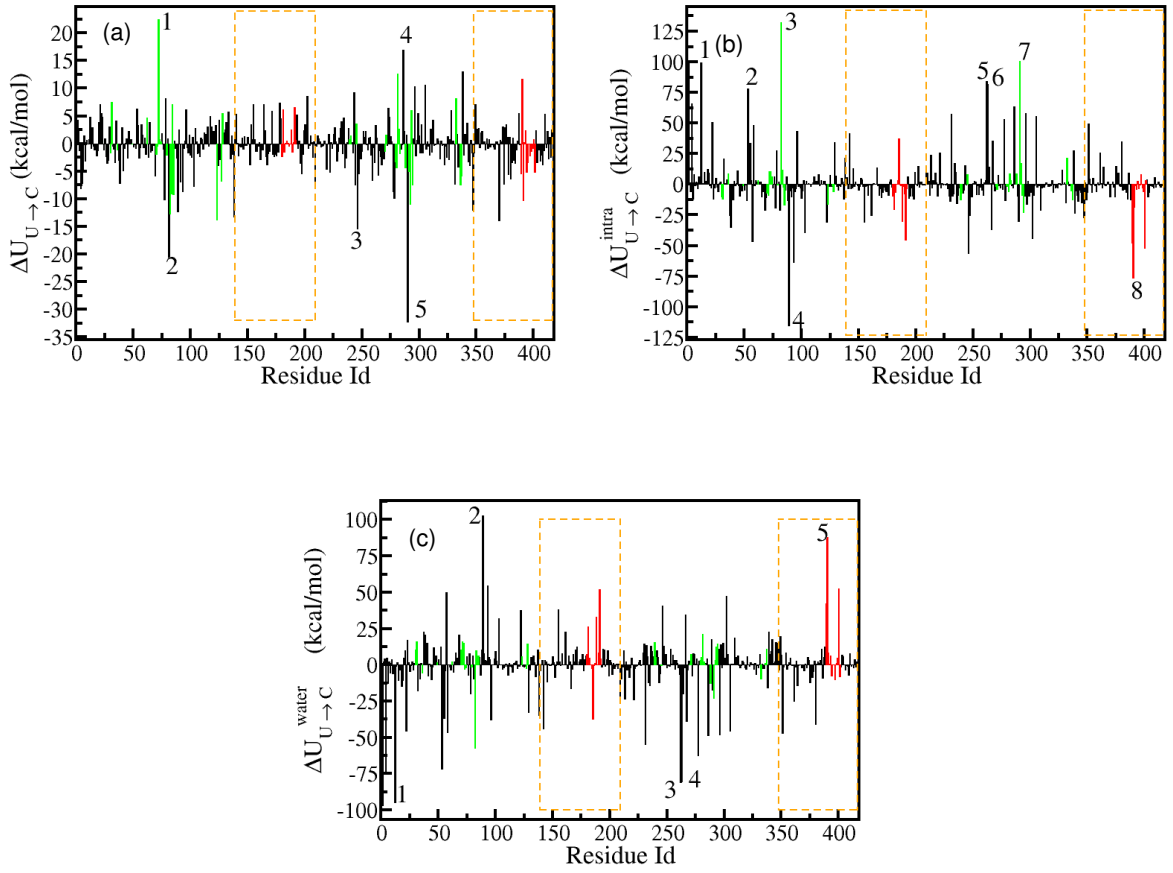


Figure 3.4: **Interaction energy changes in the cAMP-binding event.** The (a) cAMP-induced changes, $\Delta U_{U\rightarrow C}$ in the mean electrostatic energy of individual residues of CAP are shown. The corresponding changes in the mean intra-protein energy $\Delta U^{intra}_{U\rightarrow C}$ is presented in (b). The residues in the DNA-binding domains (bounded by the dashed orange rectangle), cAMP-binding pockets (green), and F-helices (red) are highlighted. The (c) cAMP-induced, $\Delta U^{water}_{U\rightarrow C}$ changes in the mean electrostatic energy of interaction between individual residues of CAP and water are shown. For each graph, residues are ranked according to their absolute interaction energy values, with a higher rank(1) and a lower rank(5) denoting a higher value and a lower value respectively.

in Figure 3.4. Amino acids primarily involved in salt-bridge and hydrogen-bonding stabilisation such as SER, THR, ASP, GLU, ASP and LYS, experience a large $\Delta U_{U\rightarrow C}$ thus showcasing the formation (negative change in $\Delta U_{U\rightarrow C}$) and breakage (positive change in $\Delta U_{U\rightarrow C}$) of salt-bridges and hydrogen-bonds. The most remarkable changes were observed in GLU-72, GLU-77, and GLU-81, which have a negatively charged side-chain, and ARG-82, which has a positively charged side-chain. These 4 residues are part of a subdomain in CAP that goes by the name *phosphate-binding cassette* (henceforth referred to as PBC) [194].

A signature motif of all cAMP-binding proteins, PBC stretches from GLY-71 to ALA-84. A bunch of amino acids in this subdomain form salt-bridges that provide scaffolding for the coiled region of VAL126 - PHE136, namely these contacts exist between the following amino-acid residues: (1) ARG-82 with GLU-129', (2) GLU-77 with ARG-122', and (3) GLU-78 and ARG-122' [195–197]. These are broken on the arrival of cAMP, but new bridges in the form of : (1) ARG-82 with cAMP's phosphate and (2) ARG-122' with GLU-129' stabilise the newly formed helical region of VAL-126 to PHE-136 [69]. Hence, the changes observed in cAMP-binding are in accordance with these experimental results. Although F-helices are nowhere in the vicinity of cAMP, their interaction energy with the environment experienced substantial changes during the cAMP binding event, as shown in Figure 3.4. The previously mentioned coil to helix transition impacts inter-amino acid contacts of CAP, thus impacting $\Delta U^{intra}_{U\rightarrow C}$ for the F-helices. As for the change in interactions with water, since F-helices form the exterior of CAP, changes in the solvent distribution around them occur due to their aforementioned conformational shift.

3.3.3 Allosteric Pathways

Prospective allosteric pathways in CAP were explored on account of the identification of amino acids in the CAP that experienced distinguishably large changes in their intra-protein interactions during the events of cAMP and DNA binding Figure 3.5. These residues are highlighted in Figures 3.5a, 3.5b, 3.5d . Figure 3.5c portrays a pathway of allosteric signals emanating at the cAMP binding pockets (CBP-1 and CBP-2), transmitted through the C-helices and finally received at F-helices, i.e. the DNA-binding domain, based on $\Delta U^{intra}_{U\rightarrow C}$. These inferred allosteric pathways rivetingly display the intertwined nature of the two subunits of CAP, with information flow occurring between the CBD of subunit-1 and the DBD of subunit-2 and vice versa. This seems to provide a basis for the intersubunit cooperativity at a compositional level.

3.3.4 Secondary Systems

As pointed out earlier on, the fact that the CAP in the systems CAP-cAMP*, CAP-cAMP-DNA*, and CAP-cAMP$\cdots$DNA instead of converging towards its structure as observed in Apo-CAP, CAP-cAMP and CAP-cAMP-DNA respectively, remains much closer to its initial orientations observed in

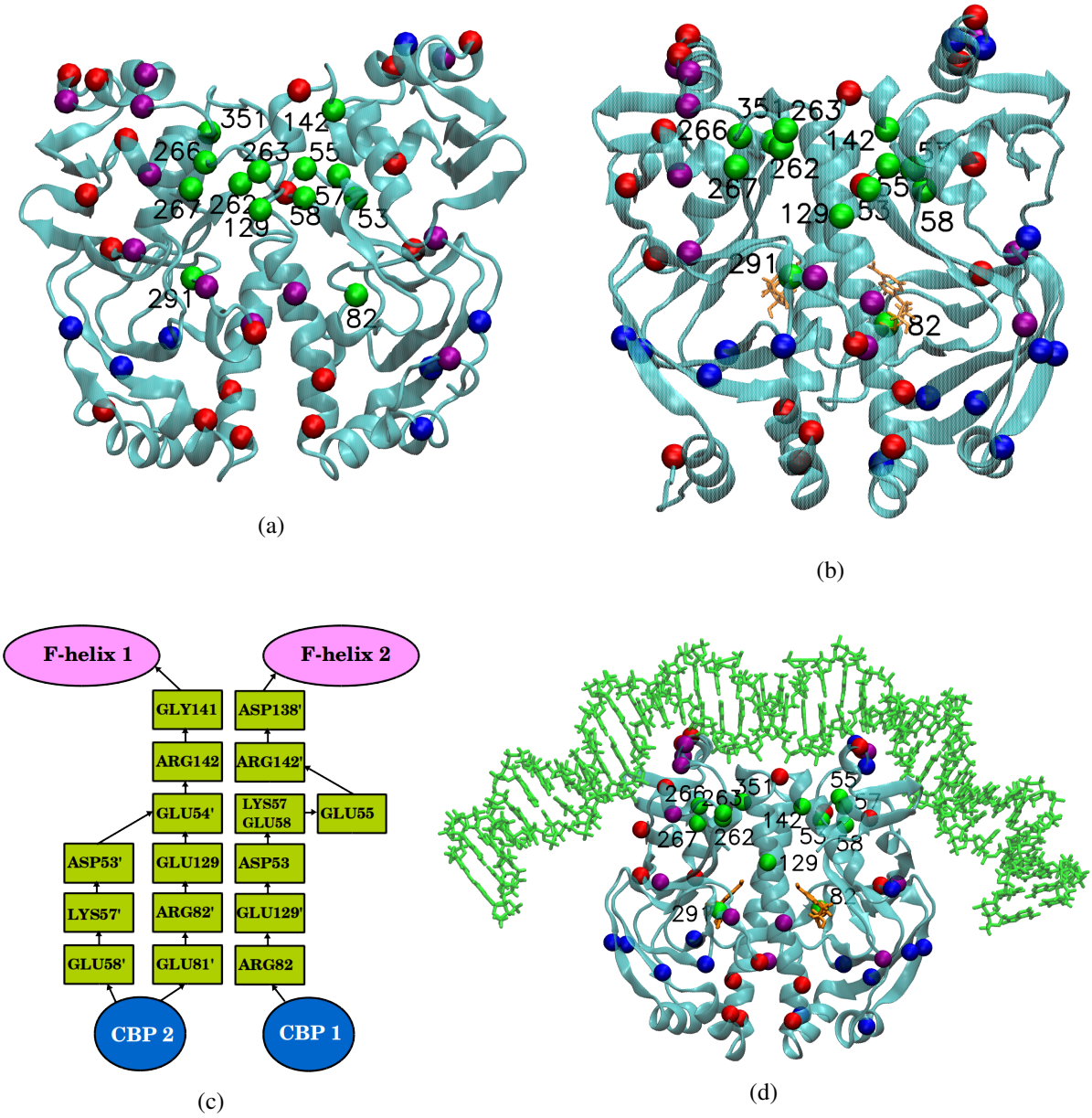
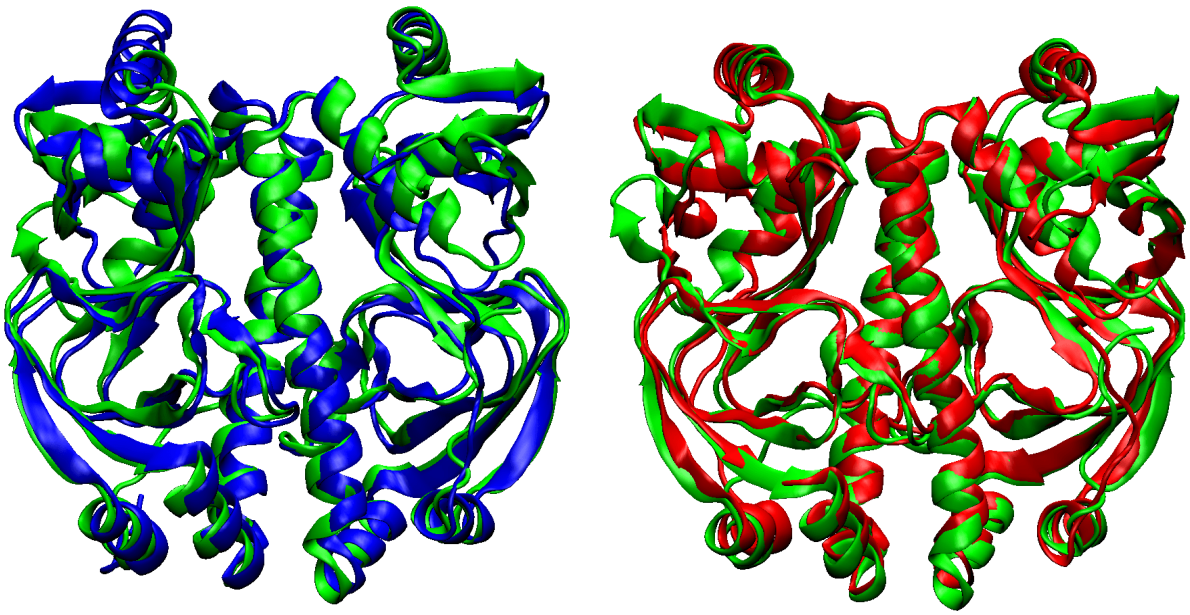


Figure 3.5: **Allosteric pathways.** The C-α atoms of the CAP residues that showed a marked change in inter-residue interactions in response to cAMP binding (blue) or to DNA binding (red), or to both DNA and cAMP binding (purple) are shown in (a) apo-CAP (b) cAMP-CAP and (d) cAMP-CAP-DNA(cAMP (orange), DNA (green)). (c) The predicted cAMP-activated allosteric pathways between the cAMP-binding pockets (CBP 1 and CBP 2) and the F-helices (F-helix 1 and F-helix 2) of subunits 1 and 2 of CAP. In (a), (b) and (d) , the C-α atoms of those residues mentioned in (c) are shown in green spheres and are labelled as well.



(a) (b)

Figure 3.6: MD-derived time-averaged structures of CAP in (a) CAP-cAMP (blue) and CAP-cAMP$\cdots$DNA (green) and (b) CAP-cAMP-DNA (red) and CAP-cAMP$\cdots$DNA(green)

CAP-cAMP, CAP-cAMP-DNA and CAP-cAMP respectively, persuaded the proposal of novel reaction coordinates in the form of (η, ξ) and carrying out an umbrella sampling simulation. The system of CAP-cAMP$\cdots$DNA however, additionally provides key insights into the symmetricity of F-helices throughout the two binding events of cAMP and DNA.

The CAP is a symmetrical molecule and the cAMP-binding seems to keep this symmetry. However, DNA-binding substantially reorients both helices and thus seems to break the symmetry. Since the focus of this study has been F-helices, the *symmetry* being talked about also concerns them. The time evolution of the collective variables, interaction energies and the structure of CAP-cAMP$\cdots$DNA are examined and compared with those of CAP-cAMP-DNA in Figure 3.6 and Figure 3.7 . F1-helix and F2-helix are defined as the F-helix on the first and second subunits, respectively.

In both of these models, the F2-helix seems to interact strongly with DNA than the F1-helix, thus indicating the symmetry breaking induced by DNA-binding (Figure 3.7). However, the structural comparison (Figure 3.6 and Figure 3.7) reveals that the CAP in CAP-cAMP$\cdots$DNA exhibits a greater structural similarity with the CAP in CAP-cAMP than with the CAP in cAMP-CAP-DNA (Figure 3.6) . The calculated RMSD of CAP in CAP-cAMP$\cdots$DNA with respect to CAP-cAMP-DNA is 2.87 Å .

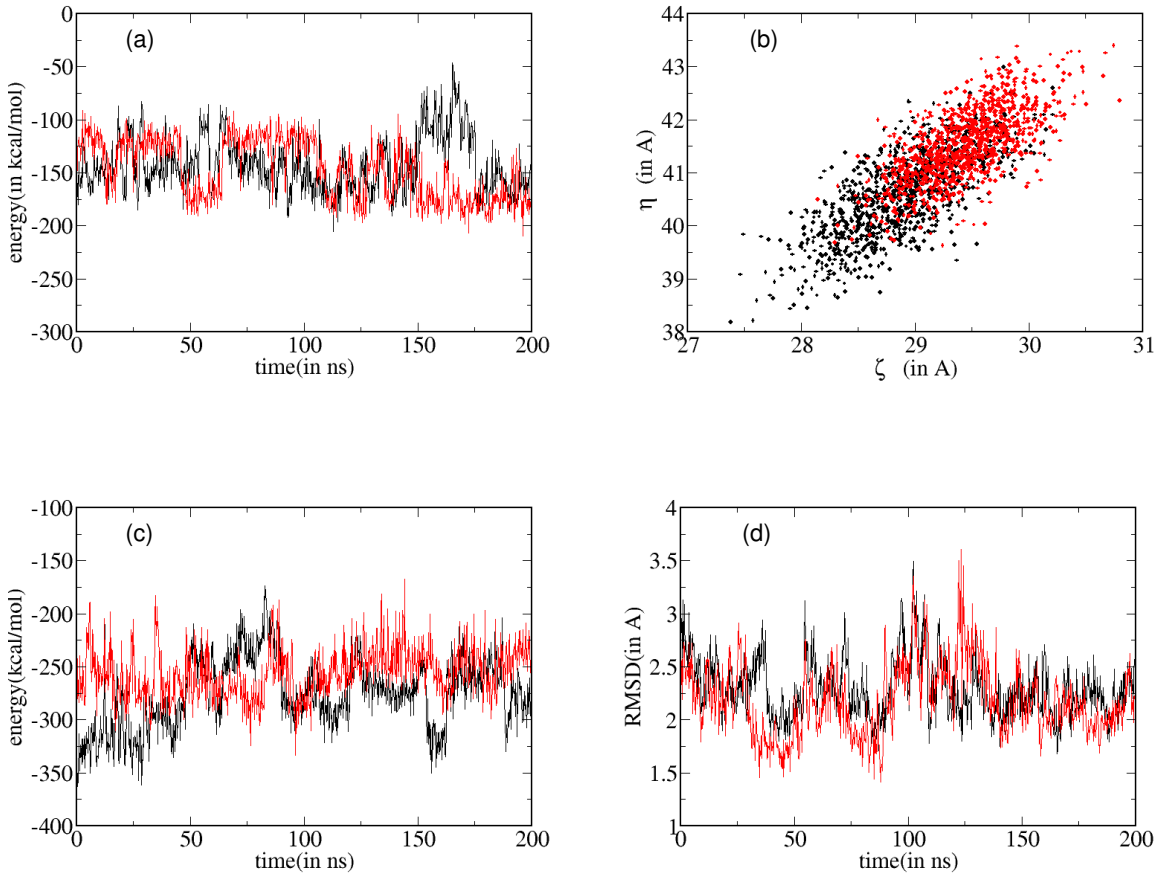


Figure 3.7: MD-derived time series of the energy of interaction between the (a) F1-helix and DNA, and (c) F2-helix and DNA; (b) the scattered plot for collective variables η and ξ obtained from the unbiased MD trajectories; In (a-c), CAP-cAMP$\cdots$DNA (red) and CAP-cAMP-DNA (black); (d) RMSD of F1-helix (black) and F2-helix (red) in CAP-cAMP$\cdots$DNA with respect to CAP-cAMP-DNA.

3.4 Conclusion

The mechanistic details and the molecular map of allosteric pathways of cAMP-mediated DNA recognition by CAP are investigated in the present study. Mainly, the cAMP-induced conformational changes in the F-helices are examined using MD and umbrella sampling simulations. By defining a set of unique reaction coordinates, the energetics and dynamics of the cAMP-binding event were captured by calculating and comparing the FEPs of the unbounded and ligand-bound proteins. The binding-induced shift in the energy minimum and changes in the shape and density of energy contours on the resultant free-energy profiles reveal important flexibility constraints imposed on DBD upon cAMP binding.

To aid the transmission of allosteric regulatory signals within CAP, residue-wise interaction maps were utilized to discern plausible pathways connecting CBD and DBD. These predicted allosteric pathways connect the CBD of one subunit to the DBD of the other subunit and proceed via *hotspot* residues of the unstructured part of the C-helices and of the CBD-DBD hinge of CAP. While the primary focus of this study was on characterising the structural alterations within CAP's F-helices, it's important to note that cAMP-induced changes can give rise to more intricate and interconnected movements. These encompass various additional secondary structural components (such as the β-strand-4/β-strand-5 flap segments and the C-helices) and molecular phenomena that extend beyond the orientational modifications observed in the F-helices. Future research should focus on understanding the coupling between some of these key molecular events using novel collective variables and different enhanced sampling methods.

Chapter 4

Energetics-based analysis of CPD-containing DNA binding to Rad4 to commence the NER process

Contents

4.1 Introduction

DNA lesions, including base alteration, base deletion, sugar alteration, and strand break, occur commonly either naturally or via environmental factors [198]. The UV light-induced cyclobutane pyrimidine dimer (CPD) is the most prevalent DNA lesion and is implicated in a variety of genetic skin-related diseases and cancers in humans [77–80].

Hydrogen bonding patterns are known to be modified when DNA contains CPD, which in turn affects the integrity of the DNA base pairing in and around the damage site. CPD-containing DNA is found to have its overall helical axis bent at ≈ 30 °toward its major groove and unwound by ≈ 9°as per molecular modelling, electron microscopy, and electrophoretic behaviour-based studies of dimer-containing oligonucleotides [199–201] . Inability to form hydrogen bonds between the 5'-dT of the CPD dimer with its partner bases leaves the CPD dimer with only one Watson-Crick base pairing, between 3'-dT and its partner Adenine [202]. Such DNA also experiences physical blockage in both replication and transcription that serve as a cell-cycle checkpoint to potentially avoid replication errors that could

alternatively lead to mutagenesis, chromosomal breakage, and DNA recombination [203–205]. Most importantly, UV photoproducts in the DNA molecule are the main cause of cell death by apoptosis post UV irradiation [80].

A specific repair protein detects the structural distortions in CPD-containing DNA, initiating the nucleotide excision repair (NER) mechanism. This mechanism involves the subsequent recruitment of other proteins to mend the DNA damage [206–214]. This repair protein probes for the lesion while sliding along the length of the DNA, halting and inducing significant conformational changes at the lesion site for damage verification and further NER actions [215–220]. Xeroderma pigmentosum C (XPC) , a key CPD repair protein in mammalian cells identifies the damage site and calls TFIIH (transcription factor) to unwind the duples and open a bubble in the DNA around the lesion site [211–214, 221–233]. Endonucleases XPG and XPF cut the lesion-containing oligonucleotide, and DNA polymerase fills up this gap, followed by the sealing action of DNA ligase to maintain the DNA structure [234–239].

Rad4 consists of an N-terminal transglutaminase domain (TGD), and three β-hairpin domains (BHD1, BHD2, and BHD3) [81, 82]. The binding of TGD and BHD1 domains to the undamaged segment of DNA helps maintain its structural integrity. The interaction between BHD2 and DNA involves its β-hairpin, which binds to the DNA minor groove near the lesion, establishing hydrogen bonds with the DNA backbone. On the other hand, the β-hairpin of BHD3 interacts with the DNA major groove, filling the space left by the flipped-out CPD and its adjacent bases from the undamaged DNA strand. The BHD2-BHD3 binding interface securely retains these displaced partner bases.

Given that the partner bases could not flip out of the CPD-containing DNA duplex in the absence of Rad4[199], it appears that the association of Rad4 with DNA must have preceded the flipping of the partner bases. In addition to this, the kinetic gating model proposed by Chen et. al (2015) [83] shows that Rad4/XPC needs to stay longer in order to form a stable complex with the DNA site of a smaller helical distortion. This makes association, i.e., the residential time of Rad4 on DNA, the most critical point that determines the overall lesion recognition efficiency. The present work attempts to explore the mechanisms and energetics of this association/dissociation of Rad4 with/from CPD-containing DNA using molecular dynamics and enhanced sampling simulations.

4.2 Simulation Details

4.2.1 Models

4.2.1.1 Rad4-DNA Complex

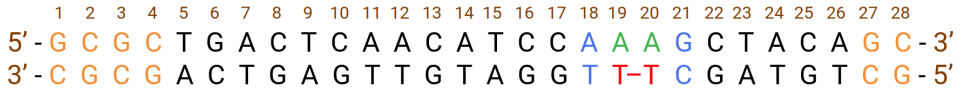


Figure 4.1: DNA sequence and nucleotide numbering scheme used in the study. The CPD (red), the partner bases (green), and the neighbouring base pairs (blue) are shown.

The Rad4-DNA complex's crystal structure (PDB ID: 2QSG) served as a model for the final associated open complex. In this bound state, Rad4's BHD2 and BHD3 are in close proximity to the damage site. The β-hairpins of BHD2 and BHD3 insert themselves into the minor and major grooves of DNA near the lesion, causing the CPD and its consecutive adenine partner bases on the undamaged strand to be entirely expelled from the DNA duplex. The DNA sequence from the crystal structure was extended to a 28-base pair sequence (Figure 4.1), incorporating a CPD-lesion at the 19^{th} and 20^{th} base pairs. The undamaged strand's CPD's partner bases, $A19_{\mathfrak{u}}$ and $A20_{\mathfrak{u}}$ (denoted by subscript $\mathfrak{u}$), are referred to as 5'-dA and 3'-dA, respectively. This nucleotide sequence corresponds to a CPD-containing DNA perfectly matched, differing from our previous studies focused on mismatched DNA [240]. To model the open complex since the CPD lesion's coordinates were unresolved in the crystal structure, we introduced the CPD lesion using the protocol used to construct the pre-association encounter complex. Moreover, in the open complex's crystal structure, we replaced each of the two mismatched thymine partner bases opposite the lesion with an adenine base utilising UCSF Chimera's Swapna module [241, 242]. While CPD-containing 'perfectly-matched' DNA might not be the most efficient substrate for direct recognition by Rad4/XPC compared to CPD within a three-base mismatch [243, 244], this model was chosen to explore the Rad4-DNA binding process in the absence of mismatches, focusing solely on the lesion's contributions. The model of the resultant associated open complex of Rad4 and DNA is depicted in Figure 4.2.

4.2.1.2 Intermediates of Rad4-DNA Complex

The above Rad4-DNA Complex is used to simulate the NER processes of de-insertion of BHD3-β hairpin from the lesion site, followed by flipping in of the partner bases 3'-dA and 5'-dA along with the flipping in of the CPD lesion. This model, termed as **model F** in the study, is used to simulate the dissociation of BHD2/BHD3 from the damaged DNA.

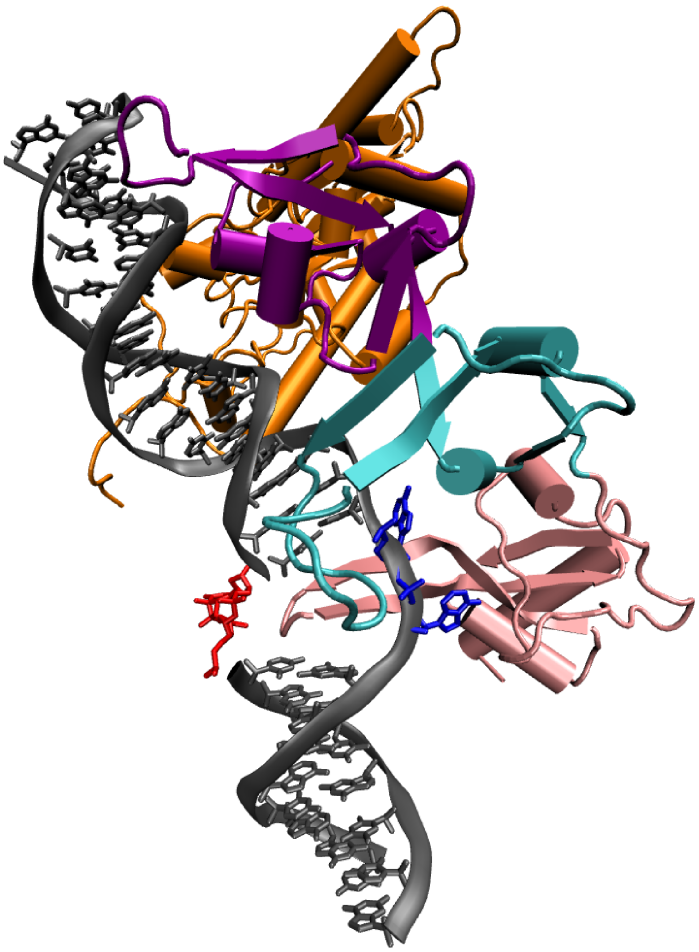


Figure 4.2: **CPD containing DNA-Rad4 complex.** The TGD, BHD1, BHD2, BHD3 domains of RAD4 are shown in golden-yellow, purple, cyan and pink respectively. As for the DNA, the CPD lesion is shown in red and its partner Adenine bases in blue. The image was generated using VMD [99].

To investigate the intermediate states, unbiased NPT molecular dynamics (MD) simulations were employed. These simulations began with diverse collective variable (CV) values specific to each intermediate process. Analysing the CV time series unveiled numerous trajectories that converged around a particular CV value. Clustering the structures from these trajectories, based on the root mean square deviation (RMSD) of key nucleotides surrounding the damaged DNA lesion site ($C17_u$ - $C22_u$ and $G17_d$ - $G22_d$), allowed the determination of the metastable state for each process by selecting the centre of the top-ranked cluster. Using this method, several significant metastable states of the Rad4-DNA complex were identified and are illustrated in Figure 4.3.

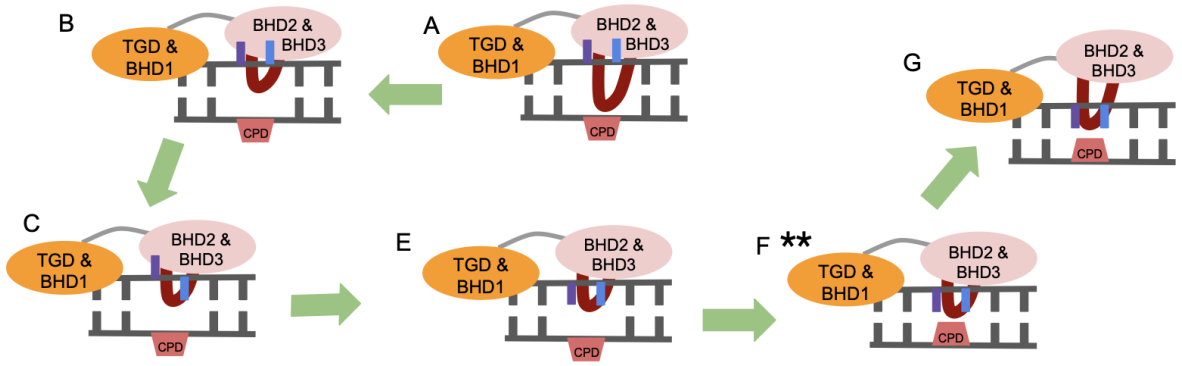


Figure 4.3: Models and sequences of events (denoted by numbered arrows) considered. (A) Rad4-DNA bound complex, (B) bound complex with BHD3 β-hairpin deinserted from the damage site, (C) same as (B) except for 3'-dA (blue) flipped into the DNA duplex, (E) same as (B) but both partner bases are flipped into the DNA duplex, (F**) same as (B) except that both partner bases and the CPD lesion are flipped into the DNA duplex, (G) same as (E) but the BHD2 and BHD3 domains are dissociated from the DNA. Structures marked with ** are the same meta-stable state formed after flipping in CPD.

4.2.2 Molecular Dynamics Simulation

AMBER 2018.1 simulation package [108, 245] was used to conduct all-atom MD simulations of all systems modelled. ff14SB [188] and ParmBSC1 [189] force fields were used for the protein and DNA respectively and the TIP3P model [190] for solvating the systems with a 20 Å of water molecule padding in each direction. 24 Na^+ ions were added to this water-boxed system for neutralization. Periodic boundary conditions were applied across all three dimensions, along with the usage of SHAKE algorithm to constrain hydrogen-related bonds. Particle Mesh Ewald(PME) [246] method having space cutoff, Ewald coefficient, and tolerance set to 10 Å, 0.27511 and 10^{-5} respectively, incorporating a 4^{th} order B-spline interpolation was used to gauge long-range electrostatic interactions. Long-range vanderWaals interactions were simulated with a distance cutoff of 10 Å.

During the initial stage of energy minimization, robust harmonic constraints were employed on the crystallographically determined atoms of the DNA-Rad4 complex to preserve its overall structural integrity, while weak constraints were placed on the yet-to-be-determined coordinates of the unresolved atoms in the complex. The robust harmonic constraints were kept in place, but the weak constraints were removed in the next phase of energy minimization. No constraints were applied to water molecules or counterions during minimization. This was followed by a 20 ps (picosecond) NVT simulations at 300 K retaining the robust harmonic constraints and then a 2 ns (nanosecond) NPT simulation at 300 K and 1 bar. Again, the entire system was energy minimised, equilibrated in an NVT ensemble at 300 K for 20 ps followed by an NPT equilibration at 300 K for 2 ns such that all restraints were lifted. Characterised by 20,000 steepest descent steps and 20,000 conjugate gradient steps, each minimization was carried out with a convergence tolerance of 10^{-4} kcal mol^{-1} Å^{-1} [108]. A Berendsen barostat [143] was employed

for isobaric conditions of 1 bar with a pressure relaxation time of 1 ps. Langevin thermostat [134] was used to maintain the temperature at 300 K with a collision frequency of 1 ps^{-1} . The velocity Verlet [247] algorithm was used to integrate the equations of motion with a time step of 2 fs.

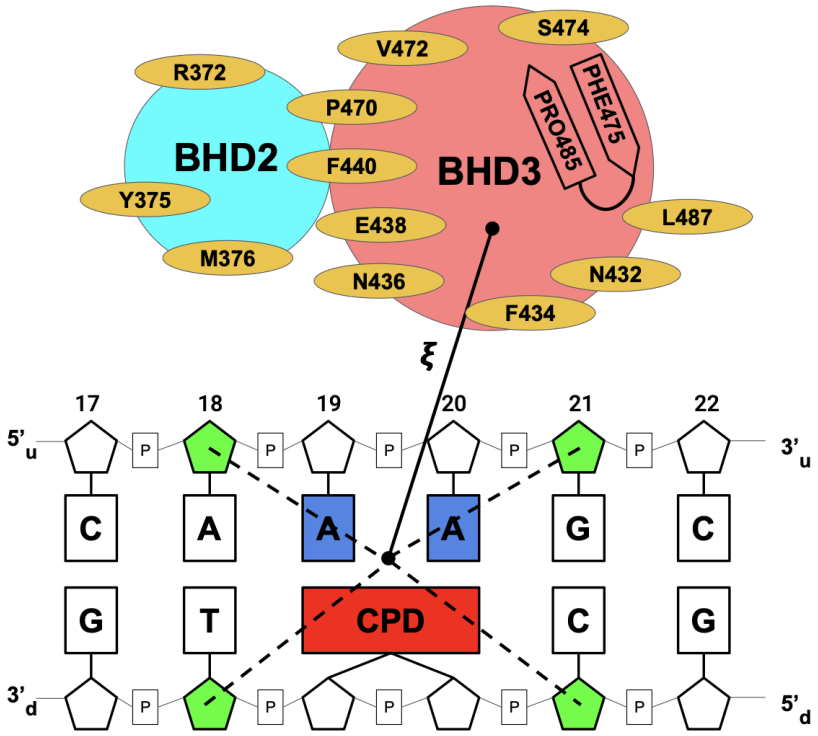


Figure 4.4: **Schematic representation of the Collective Variable used to simulate the association processes of NER.** ξ is the distance between COM of heavy atoms of amino acids(yellow ellipses) and the backbone heavy atoms of BHD3-β-hairpin amino acids(pink loop), and the COM of the sugar rings of the neighbouring bases of CPD and its partners, i.e. ($A18_u$, $G21_u$, $T18_d$, $C21_d$) (in green).

4.2.3 Umbrella Sampling

4.2.3.1 Collective Variable Definition

The association of the BHD2 and BHD3 domains of Rad4 with the damaged DNA was captured by the distance between the COM of the sugar rings of the four neighbouring nucleotides to the lesion site (see Figure 4.1) and the COM of the backbone heavy atoms of the binding pocket residues 372, 375, 376, 432, 434, 436, 438, 440, 470, 472, 474-487 of BHD2 and BHD3 (refer Section 4.2.2), which will be denoted as ξ. ξ was varied from 10 to 30 Å in steps of 0.5 Å, corresponding to a total of 41 windows with the biasing harmonic force constants for the equilibration and production runs were set to 100 kcal mol^{-1} $Å^{-2}$ and 10 kcal mol^{-1} $Å^{-2}$, respectively.

It is uncertain if the CPD and its partner bases remain in their extrahelical states once Rad4 has been successfully dissociated from damaged DNA, whereas both are fully extruded when DNA is bound to Rad4. This called for examination of the dissociation of Rad4 from DNA in two different models of the partner bases: the intra-helical (Model F) and extra-helical (Model A) models(ref Figure 4.3). As mentioned previously, in Model F, the BHD3 β-hairpin is removed from the DNA duplex, while in Model A, it remains inserted.

In theory, the association of Rad4 to DNA (Model F) should force the partner bases and CPD to flip into the DNA duplex. Since this cannot be observed within the duration of our simulations, model A was employed to investigate the dissociation process post-flipping out of the partner bases and CPD from the DNA duplex.

4.2.3.2 Umbrella Sampling Protocol

The final frame of the unbiased MD simulations was used as the starting structure for the umbrella sampling simulations. An umbrella sampling run involved shifting the system to the required window using a harmonic biassing potential with a high spring constant (k_{eq}) in a 200 ps NPT equilibration run (at 300 K and 1 bar) such that the appropriate CV is brought to the centre of this required window. A 6 ns NPT production run succeeded this, where the biassing potential of a weaker spring constant (k_{prod}, considerably smalled than k_{eq}) was placed. The simulation setup of these umbrella sampling simulations was kept the same as that of the unbiased MD runs, with an additional restraint on the following distances: (1) distance between the COMs of bases dA-583 and dT-603 restrained at 6.06 Å using a harmonic bias of 25 kcal mol^{-1} Å^{-2} (2) distance between the COMs of bases dT-586 and dC-601 restrained at 5.85 Åusing a bias of 25 kcal mol^{-1} Å^{-2}. The bases dC-601,dT-603 and dA-583,dT-586 are the neighbouring bases of the CPD lesion and its partner adenines, respectively. These constraints maintain the intra-helical states of these neighbouring bases, which is expected throughout the NER process [248].

4.3 Results and Discussion

The results of umbrella sampling simulations for DNA-Rad4 association are discussed in this section. These results are from the reverse pathway of the NER process, as their initial structure was the crystal structure of the bound complex. In other words, Rad4-DNA dissociation was performed after deinserting the BHD3 β-hairpin from the DNA lesion site and flipping the CPD and its partner bases into an intrahelical conformation.

To solely investigate the effects of Rad4 dissociation from the damaged DNA, a model of the Rad4-DNA complex having both CPD and its partner bases in intrahelicular conformation and a de-inserted BHD3 β-hairpin, a.k.a. model F in Figure 4.3 was considered.

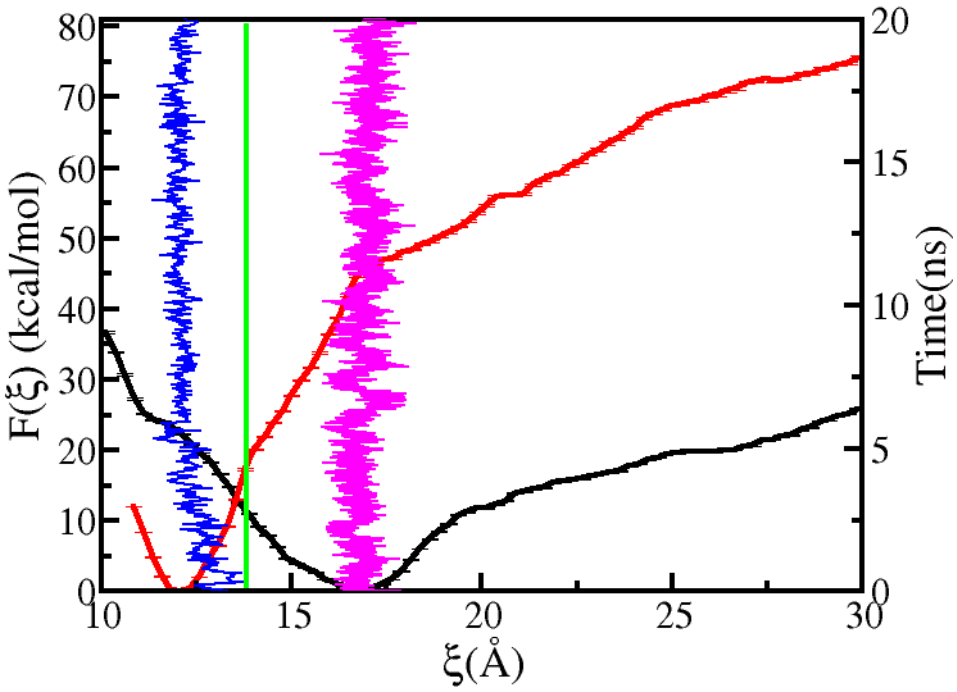


Figure 4.5: PMF for Rad4-DNA dissociation calculated for the crystal structure (red; Model A) and the encounter complex (black; Model F) in which the partner bases and CPD are intrahelical and the BHD3 β-hairpin is deinserted. The time series of the dissociation CV ξ obtained from unbiased MD runs of Model A (blue) and Model F (megenta) are also shown.

A free energy profile ($F(\xi)$) using this model for Rad4-DNA dissociation is shown in Figure 4.5. The point where $F(\xi)$ reaches its lowest free energy, designated as ξ_{min} from now on and positioned at ξ = 16.9 Å, is surrounded by an uneven basin. The sharp surge in $F(\xi)$ occurs when $\xi < \xi_{min}$ due to steric clashes between the DNA and the BHD2/BHD3 domains. Conversely, the gradual attenuation of favourable interactions between the BHD2/BHD3 domains and DNA is the reason behind the elevation of $F(\xi)$ for $\xi > \xi_{min}$. A shift in slope around $\xi = 19.5$Å implies that the critical interactions that once stabilised the bonded complex are nearly lost for $\xi \geq 19.5$Å.

4.3.1 Free Energy Profiles

A free energy profile for the dissociation of Rad4 from DNA for the experimentally resolved crystal structure of the Rad4-DNA bound complex (Model A in Figure 4.3) is shown in Figure 4.5. The lowest free energy point for this curve is at $\xi \approx$ 12Å, closer to the experimental crystal structure value of ξ=13.81 Å and corresponds to the truly associated state of the Rad4-DNA complex.

Comparing the two free energy profiles reveals a notable distinction: displacing Rad4 from DNA by 3 Å (departing from the energy minimum) is less energetically demanding in the former model compared to the latter. The associated energy cost for $\xi - \xi_{min} = 3$ Å is 11.14 kcal mol^{-1} in the former model and 27.67 kcal mol^{-1} in the latter. In the former scenario, an elevation is evident around $\xi \sim 11.5$ Å, indicating a decrease in stability by 23.9 kcal mol^{-1} relative to the corresponding global minimum state. Intriguingly, the position of this elevation roughly corresponds with the location of the energy minimum in the energy profile of the latter model.

A mechanism for the Rad4-DNA association can be inferred from these findings. The intermediate complex is initially formed at ξ = 16.9 Å, where CPD and its partner bases are found to be intrahelical. The shift in energy minimum to ξ = 12 Å is caused by the flipping of CPD and partner bases into an extrahelical conformation. This results in an increased capability of Rad4 to interact with and access deeper regions of the lesion site. The energy expenditure associated with flipping the CPD and partner bases out of the DNA duplex can be approximated using the energy difference ($F(\xi$ = 11.5 Å) - $F(\xi_{min}$ = 16.9 Å) between the global minimum and the elevated state, indicating an approximate cost of 23.9 kcal mol^{-1}. Importantly, the cumulative energy expenses for these flipping events of both partner bases and CPD expressed in other parts of the extended study closely align with this estimated value.

The kinetic gating model proposed by Chen et al. (2015) states that Rad4/XPC needs to stay longer in order to form a stable complex with the DNA site of a smaller helical distortion. This makes the dissociation of Rad4 the most critical point that determines the overall lesion recognition efficiency. To obtain a qualitative assessment of this residential time of Rad4 onto CPD-containing perfectly-matched DNA (as shown in Figure 4.1 the CPD-forming bases are themselves thymines matched with partner adenines on the undamaged strand.), the model A profile was compared with the dissociation profile of a TTT/TTT mismatched DNA of homologous sequence studied in Abhinandan et. al[240]. The energy difference $\Delta F = |F(\xi_{min} + \Delta\xi) - F(\xi_{min})|$ served as a metric for the energy required to dissociate Rad4 from the DNA. Here, ξ_{min} represents the position of the energy minimum, and $\Delta\xi$ denotes the displacement on the high ξ side of the minimum. After fixing $\Delta\xi$ to 3 Å, ΔF enables the measurement of the energy expenditure necessary to displace Rad4/XPC from the energy minimum over a distance of 3 Å. Using this scheme, the PMF profile comparison calculations reveal that the dissociation energy is comparatively higher for the CPD-containing matched DNA in comparison to the mismatched DNA. This implies that Rad4 is likely to stay longer in an associated conformation with a CPD-containing matched DNA than with a TTT/TTT mismatch-only DNA.

4.4 Conclusion

Given the rugged underlying potential energy landscape of this complex system, the NER molecular processes turn out to be necessarily slow, making them particularly difficult to investigate using con-

ventional molecular dynamics simulation. Therefore, there was a need for utilising enhanced sampling methods in addition to classical MD simulations to investigate the mechanism and energetics of Rad4-DNA association and the effect of BHD3 β-hairpin and CPD and its partner bases on the same. The corresponding detailed PMF profiles demonstrated the attribution of a shift in the energy minimum to the switch in the conformations of CPD and its partner bases relative to the DNA helix.

Furthermore, a comparison of these profiles with the dissociation profile of a homologous DNA-sequence base-pair mismatched DNA-Rad4 complex revealed a greater dissociation energy for the CPD-containing matched DNA. Such a deduction suggests that in the case of DNA containing a CPD, Rad4 is more likely to assume an associated conformation for an extended duration compared to a DNA sequence solely consisting of a TTT/TTT mismatch. This implies a greater overall lesion recognition efficiency for the CPD-containing damaged DNA.

Chapter 5

Conclusion

Biologists are increasingly recognising the value of moving beyond static snapshots to explore the dynamics and energetics of complex biomolecular processes. To gain deeper insights into the dynamic and energetic aspects of these intricate systems, researchers have turned to Molecular Dynamics (MD) simulations. The use of MD simulations facilitates exploration of the motion and behaviour of biomolecules over different lengths of time. While empirical methods such as X-ray crystallography and nuclear magnetic resonance (NMR) research have provided vital insights into the makeup of molecules, they fall short of capturing the dynamic nature of biological processes. This examination of the dynamics and energetics of various processes of interest and importance is where MD finds its use, providing a dynamic perspective and enabling the study of conformational changes, ligand binding, and other crucial events with high temporal resolution. However, MD simulations face difficulties when exploring energetically prohibitive regions of free energy landscapes. This is where enhanced sampling methods, such as Umbrella Sampling, come into play. These methods are able to drive the system into states that are typically inaccessible under standard experimental conditions with the application of suitable biassing potentials. The present thesis aims at inspecting structural alterations and the interactions in proteins brought forth by allostery within two biologically significant systems: (b) the conformational changes experienced by a transcription factor, Catabolite Activator Protein, post ligand-binding and subsequent DNA-binding; and (b) understanding how the Rad4/XPC protein binds to a cyclobutane-pyrimidine dimer (CPD) containing DNA while studying the recognition mechanism of the former on the lesion-containing segment of the latter.

Chapters 1 and 2 present how proteins and DNA, two popular and functionally important biomolecules, were discovered and offer an initial look at factors influencing their binding and the computational methods employed to study them. The biological makeup of DNA, proteins, and the complexes formed when they associate with each other is also discussed, delving into the underlying interactions that enable them to perform their roles. Following this, a concise introduction to Molecular Dynamics (MD) and the various strategies implemented by contemporary MD software like GROMACS and AMBER to enhance their efficiency and accuracy is also elaborated.

In Chapter 3, MD simulation and umbrella sampling techniques are utilised to explore the dynamics of cAMP-mediated allostery of CAP and the subsequent binding to DNA. Notably, the cAMP-binding conformational changes effected in the F-helices of CAP and the interaction tendency of CAP-cAMP and CAP-DNA when some meta-stable intermediate conformational states based on cAMP-bound CAP and cAMP-bound CAP complexed with DNA are simulated in an unrestrained environment were examined. Interaction energies between key residues of CAP with cAMP and DNA were observed to ultimately decode allosteric communication pathways between the monomeric subunits of CAP. A set of novel reaction coordinates (η and ξ) were created to examine the conformational changes brought forth in the F-helices of CAP when (a) cAMP molecules are situated inside the binding pocket of each monomeric subunit of CAP, followed by (b) docking the DNA onto the DNA-binding domain of this liganded CAP. The corresponding free energy profiles were created for the unliganded and cAMP-bound complexes using umbrella sampling. These provide a detailed picture of the elasticity imposed on the DNA-binding domain of CAP when cAMP is appropriately situated.

Based on the interaction energy of individual residues of CAP, residue-wise interaction maps are created and used to identify potential pathways between CBD and DBD that facilitate the allosteric transduction of regulatory signals in CAP. Each of the predicted allosteric pathways connects the CBD of one subunit to the DBD of the other subunit and passes through *hotspot* residues of unstructured parts of the C-helices and of the CBD-DBD hinge of CAP. This criss-crossing of the inter-subunit interface offers clues on the microscopic origin of the inter-subunit cooperativity and dimer stability of CAP.

Chapter 4 entails the exploration of energetics around the Rad4-DNA interactions using molecular dynamics and umbrella sampling simulations as part of the nucleotide excision repair process of a CPD-containing perfectly matched DNA carried out by Rad4/XPC. The aim is to lay out the first phase: the association of the BHD2 and BHD3 domains of Rad4 onto the lesion site of the DNA. The fact that a kinetic gating mechanism study done in 2015 revealed the importance of the residential time of Rad4 on DNA, which in turn is determined by how efficiently Rad4 has been associated with the DNA, underscores the critical role of this phase of NER. Moreover, the extended version of this study includes studying the other two phases of NER: partner-base pair flipping and BHD3 β-hairpin insertion; hence, there was a need to study the Rad4-DNA association in order to completely layout the NER mechanism of a CPD-containing DNA. The free energy profile corresponding to this event was evaluated using a suitable reaction coordinate.

The free energy profile obtained from this study was compared with another corresponding to the Rad4-DNA association of a DNA having a TTT/TTT base-pair mismatch as the only lesion, covered as part of a previous study. This comparison was performed because the latter is an extensively studied system in terms of the binding efficiency of Rad4 with DNA. The comparison ultimately revealed that Rad4 is more likely to remain in an associated conformation for an extended duration with a CPD-containing matched DNA compared to a TTT/TTT mismatch-only DNA.

Future Work

With regards to the allosteric study performed in Chapter 3, although the allosteric response exhibited by the F-helices of CAP was the highlight, cAMP-induced CAP-DNA binding processes could potentially involve an assortment of various coupled motions involving other key secondary structural units (such as the β-strand-4/β- strand-5 flap segments and the C-helices) and molecular movements beyond the transformations in the F-helices. Future research should focus on understanding the coupling between some of these key molecular movements using novel collective variables and multi-dimensional umbrella sampling methods.

As for Chapter 4, since the present study, including the extended version, only dealt with one-dimensional reaction coordinates, the use of multidimensional enhanced sampling methods involving variation of more than one collective variable is bound to be more realistic in an attempt to capture the intricacies of lesion recognition dynamics and to understand the coupling between the aforementioned molecular events. Additionally, comparing NER of different DNA systems, i.e., DNA having different lesions both in an isolated setting and when multiple lesions exist on the same or different strand, may be needed to gain insights into the similarities, distinctions, and any aspect of lesion specificity across the different lesions and mismatch recognition mechanisms by Rad4.

Related Publications

- Akshay Prabhakant, Abhinandan Panigrahi and Marimuthu Krishnan. "Allosteric Response of DNA Recognition Helices of Catabolite Activator Protein to cAMP and DNA Binding" *Journal of Chemical Information and Modeling*, October 2020.
- Nikhil Jakhar, Akshay Prabhakant and Marimuthu Krishnan. "Mapping the Recognition Pathway of Cyclobutane Pyrimidine Dimer in DNA by Rad4/XPC" *Nucleic Acids Research*, September 2023 (Manuscript under galley proof checks).

Bibliography

[1] The big 4 of biomolecules. https://byjus.com/biology/biomolecules/, 2018. Accessed: 2020-03-16.

[2] K Murray, Victor Rodwell, David Bender, Kathleen M Botham, P Anthony Weil, and Peter J Kennelly. Harper's illustrated biochemistry. 28. Citeseer, 2009.

[3] D.L. Nelson and M.M. Cox. Lehninger Principles of Biochemistry. W. H. Freeman, 2008.

[4] Structure of an amino acid. https://study.com/academy/lesson/what-are-amino-acids-definition-structure-quiz.html, 2015. Accessed: 2020-03-01.

[5] Thomas Burr Osborne. The vegetable proteins. Longmans, Green and Company, 1924.

[6] Robert A. Kyle and David P. Steensma. Jöns jacob berzelius – a father of chemistry. Mayo Clinic Proceedings, 93(5):e53–e54, May 2018.

[7] Joseph S. Fruton. Contrasts in scientific style. emil fischer and franz hofmeister: Their research groups and their theory of protein structure. Proceedings of the American Philosophical Society, 129(4):313–370, 1985.

[8] VV Suresh Babu. One hundred years of peptide chemistry. Resonance, 16(7):640–647, 2011.

[9] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature, 181(4610): 662–666, March 1958.

[10] C. Nick Pace and J. Martin Scholtz. A helix propensity scale based on experimental studies of peptides and proteins. Biophysical Journal, 75(1):422–427, July 1998.

[11] L. Pauling and R. B. Corey. The pleated sheet, a new layer configuration of polypeptide chains. Proceedings of the National Academy of Sciences, 37(5):251–256, May 1951.

[12] Boris Schmidt. Proteins: Structure and function. by david whitford. ChemBioChem, 7(4):702–703, March 2006.

[13] Johann Gregor Mendel. 9.(1866). versuche über plflanzenhybriden verhandlungen des naturforschenden vereines in brünn, bd. iv für das jahr, 1865 abhandlungen: 3-47. for the english translation, see: Druery, ct and william bateson (1901).“experiments in plant hybridization”. Journal of the Royal Horticultural Society, 26:1–32, 1865.

[14] Asbjörn Fölling. Über ausscheidung von phenylbrenztraubensäure in den harn als stoffwechselanomalie in verbindung mit imbezillität. Hoppe-Seyler´s Zeitschrift für physiologische Chemie, 1934.

[15] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in neurospora. Proceedings of the National Academy of Sciences, 27(11):499–506, November 1941.

[16] Fred Griffith. The significance of pneumococcal types. Journal of Hygiene, 27(2):113–159, January 1928.

[17] Schematic comparison of dna with rna. https://www.expii.com/t/dna-vs-rna-differences-similarities-10205#:~:text=While%20both%20DNA%20and%20RNA,adenine%2C%20cytosine%2C%20and%20guanine., 2019. Accessed: 2020-02-27.

[18] Schematic diagram of a nucleotide. https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/nucleotides.html, 2016. Accessed: 2020-02-27.

[19] Phosphodiester bond formation. https://www.pngitem.com/middle/mwiwiJ_image130-phosphodiester-bond-formation-in-dna-hd-png/, 2019. Accessed: 2020-02-27.

[20] John Boyle. Lehninger principles of biochemistry (4th ed.): Nelson, d., and cox, m. Biochemistry and Molecular Biology Education, 33(1):74–75, January 2005.

[21] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type iii. Journal of Experimental Medicine, 79(2):137–158, February 1944.

[22] A. D. Hershey and Martha Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. Journal of General Physiology, 36(1):39–56, September 1952.

[23] Watson crick double helix dna model - b-dna structure. https://geneticeducation.co.in/role-of-alcohol-in-dna-extraction/, 2018. Accessed: 2020-03-04.

[24] Anti-parallel arrangement of dna double helix strands. https://courses.lumenlearning.com/microbiology/chapter/structure-and-function-of-dna/, 2006. Accessed: 2020-03-04.

[25] Erwin Chargaff, Rakoma Lipshitz, and Charlotte Green. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. Journal of Biological Chemistry, 195(1):155–160, 1952.

[26] D. Elson and E. Chargaff. On the desoxyribonucleic acid content of sea urchin gametes. Experientia, 8(4):143–145, April 1952.

[27] MH Wilkins and JT Randall. Crystallinity in sperm heads: molecular structure of nucleoprotein in vivo. Biochimica et biophysica acta, 10(1):192, 1953.

[28] C. A. DEKKER and A. R. TODD. Uracil deoxyriboside. Nature, 166(4222):557–558, September 1950.

[29] A. Holland, B. Lythgoe, and A. R. Todd. 184. experiments on the synthesis of purine nucleosides. part XVIII. a synthesis of 9-d-glucopyranosidoadenine. Journal of the Chemical Society (Resumed), page 965, 1948.

[30] J. D. WATSON and F. H. C. CRICK. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature, 171(4356):737–738, April 1953.

[31] Arrangement of chromosome in cell nucleus. https://courses.lumenlearning.com/suny-wmopen-biology1/chapter/chromosomes-and-dna-packaging/, 2014. Accessed: 2020-02-25.

[32] EJ DuPraw. Macromolecular organization of nuclei and chromosomes: a folded fibre model based on whole-mount electron microscopy. Nature, 206(4982):338–343, 1965.

[33] EJ DuPraw. Evidence for a 'folded-fibre'organization in human chromosomes. Nature, 209 (5023):577–581, 1966.

[34] T. J. Richmond, J. T. Finch, B. Rushton, D. Rhodes, and A. Klug. Structure of the nucleosome core particle at 7 Å resolution. Nature, 311(5986):532–537, October 1984.

[35] K Murray. The basic proteins of cell nuclei. Annual review of biochemistry, 34(1):209–246, 1965.

[36] Dna replication diagramatic representation. https://en.wikipedia.org/wiki/DNA_replication, 2013. Accessed: 2020-03-20.

[37] Francis HC Crick. The biological replication of macromolecules. In Symp. Soc. Exp. Biol, volume 12, pages 138–163, 1958.

[38] John Cairns. The bacterial chromosome and its manner of replication as seen by autoradiography. Journal of molecular biology, 6(3):208–IN5, 1963.

[39] Arthur Kornberg. DNA replication. Freeman, 1980. 574.8732 KOR. CIMMYT.

[40] A. Kornberg. Biologic synthesis of deoxyribonucleic acid. Science, 131(3412):1503–1508, May 1960.

[41] David Owen Morgan. The cell cycle: principles of control. New science press, 2007.

[42] I Robert Lehman, Maurice J Bessman, Ernest S Simms, and Arthur Kornberg. Enzymatic synthesis of deoxyribonucleic acid i. preparation of substrates and partial purification of an enzyme from escherichia coli. Journal of Biological Chemistry, 233(1):163–170, 1958.

[43] Lawrence A. Loeb and Raymond J. Monnat. DNA polymerases and human disease. Nature Reviews Genetics, 9(8):594–604, August 2008.

[44] Manjula Pandey, Salman Syed, Ilker Donmez, Gayatri Patel, Taekjip Ha, and Smita S. Patel. Coordinating DNA replication by means of priming loop and differential synthesis rate. Nature, 462(7275):940–943, November 2009.

[45] R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto, and A. Sugino. Mechanism of DNA chain growth. i. possible discontinuity and unusual secondary structure of newly synthesized chains. Proceedings of the National Academy of Sciences, 59(2):598–605, February 1968.

[46] H Malonga, JF Neault, H Arakawa, and HA Tajmir-Riahi. Dna interaction with human serum albumin studied by affinity capillary electrophoresis and ftir spectroscopy. DNA and cell biology, 25(1):63–68, 2006.

[47] Harry F Noller. Rna structure: reading the ribosome. Science, 309(5740):1508–1514, 2005.

[48] Klemens J Hertel and Brenton R Graveley. Rs domains contact the pre-mrna throughout spliceosome assembly. Trends in biochemical sciences, 30(3):115–118, 2005.

[49] Cristina Iftode, Yaron Daniely, and James A. Borowiec. Replication protein a (RPA): The eukaryotic SSB. Critical Reviews in Biochemistry and Molecular Biology, 34(3):141–180, January 1999.

[50] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature, 389(6648): 251–260, September 1997.

[51] Steve Busby and Richard H Ebright. Transcription activation by catabolite activator protein (CAP). Journal of Molecular Biology, 293(2):199–213, October 1999.

[52] Orlando D Schärer. Nucleotide excision repair in eukaryotes. Cold Spring Harbor perspectives in biology, 5(10):a012609, 2013.

[53] Masayuki Yokoi, Chikahide Masutani, Takafumi Maekawa, Kaoru Sugasawa, Yoshiaki Ohkuma, and Fumio Hanaoka. The xeroderma pigmentosum group c protein complex xpc-hr23b plays an important role in the recruitment of transcription factor iih to damaged dna. Journal of Biological Chemistry, 275(13):9870–9875, 2000.

[54] Jean-Marc Egly and Frédéric Coin. A history of tfiih: two decades of molecular biology on a pivotal transcription/repair factor. DNA repair, 10(7):714–721, 2011.

[55] Jochen Kuper, Cathy Braun, Agnes Elias, Gudrun Michels, Florian Sauer, Dominik R Schmitt, Arnaud Poterszman, Jean-Marc Egly, and Caroline Kisker. In tfiih, xpd helicase is exclusively devoted to dna repair. PLoS Biol, 12(9):e1001954, 2014.

[56] Maxime Louet, Christian Seifert, Ulf Hensen, and Frauke Gräter. Dynamic allostery of the catabolite activator protein revealed by interatomic forces. PLoS Comput. Biol., 11(8):e1004358, August 2015.

[57] Virgil A Rhodius and Stephen JW Busby. Positive activation of gene expression. Curr. Opin. Microbiol., 1(2):152–159, April 1998.

[58] Victor De Lorenzo, Marta Herrero, Fabio Giovannini, and J.B. Neilands. Fur (ferric uptake regulation) protein and cap (catabolite-activator protein) modulate transcription of fur gene in escherichia coli. Eur. J. Biochem., 173(3):537–546, 1988.

[59] A. Kolb, S. Busby, H. Buc, S. Garges, and S. Adhya. Transcriptional regulation by cAMP and its receptor protein. Annu. Rev. Biochem., 62(1):749–797, June 1993.

[60] Catherine L Lawson, David Swigon, Katsuhiko S Murakami, Seth A Darst, Helen M Berman, and Richard H Ebright. Catabolite activator protein: DNA binding and transcription activation. Current Opinion in Structural Biology, 14(1):10–20, February 2004.

[61] W S Reznikoff. Catabolite gene activator protein activation of lac transcription. J. Bacteriol., 174 (3):655–658, February 1992.

[62] Shiou-Ru Tzeng and Charalampos G. Kalodimos. Dynamic activation of an allosteric regulatory protein. Nature, 462(7271):368–372, November 2009.

[63] Aichun Dong, Jedrzej M. Malecki, Lucy Lee, John F. Carpenter, and J. Ching Lee. Ligand-induced conformational and structural dynamics changes inEscherichia coliCyclic AMP receptor protein†. Biochemistry, 41(21):6660–6667, May 2002.

[64] Hyung-Sik Won, T. Yamazaki, Tae-Woo Lee, Mi-Kyung Yoon, Sang-Ho Park, Y. Kyogoku, and Bong-Jin Lee. Structural understanding of the allosteric conformational change of cyclic AMP receptor protein by cyclic AMP binding†. Biochemistry, 39(45):13953–13962, November 2000.

[65] James G. Harman. Allosteric regulation of the cAMP receptor protein. Biochim. Biophys. Acta, 1547(1):1–17, May 2001.

[66] Otto G. Berg and Peter H. von Hippel. Selection of DNA binding sites by regulatory proteins. Trends Biochem. Sci., 13(6):207–211, June 1988.

[67] Erica A. Pyles and J. Ching Lee. Mode of selectivity in cyclic AMP receptor protein-dependent promoters inEscherichia coli†. Biochemistry, 35(4):1162–1172, January 1996.

[68] Dietmar Porschke. Allosteric control of cAMP receptor binding dynamics. Biochemistry, 51 (19):4028–4034, May 2012.

[69] Nataliya Popovych, Shiou-Ru Tzeng, Marco Tonelli, Richard H Ebright, and Charalampos G Kalodimos. Structural basis for camp-mediated allosteric control of the catabolite activator protein. Proc. Natl. Acad. Sci. U.S.A., 106(17):6927–6932, 2009.

[70] David B McKay and Thomas A Steitz. Structure of catabolite gene activator protein at 2.9 å resolution suggests binding to left-handed b-dna. Nature, 290(5809):744, 1981.

[71] JM Passner, SC Schultz, and TA Steitz. Modeling the camp-induced allosteric transition using the crystal structure of cap-camp at 2.1 å resolution. J. Mol. Biol., 304(5):847–859, 2000.

[72] Steve C Schultz, George C Shields, and Thomas A Steitz. Crystal structure of a cap-dna complex: the dna is bent by 90 degrees. Science, 253(5023):1001–1007, 1991.

[73] Gary Parkinson, Christopher Wilson, Angelo Gunasekera, Yon W Ebright, Richard E Ebright, and Helen M Berman. Structure of the cap-dna complex at 2.5 å resolution: a complete picture of the protein-dna interface. J. Mol. Biol., 260(3):395–408, 1996.

[74] Andrew A Napoli, Catherine L Lawson, Richard H Ebright, and Helen M Berman. Indirect read-out of dna sequence at the primary-kink site in the cap–dna complex: Recognition of pyrimidine-purine and purine-purine steps. J. Mol. Biol., 357(1):173–183, 2006.

[75] Ewelina Fic, Agnieszka Polit, and Zygmunt Wasylewski. Kinetic and structural studies of the allosteric conformational changes induced by binding of cAMP to the cAMP receptor protein fromEscherichia coli†. Biochemistry, 45(2):373–380, January 2006.

[76] Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. Dynamically driven protein allostery. Nat. Struct. Mol. Biol., 13(9):831, 2006.

[77] R. B. Setlow. The wavelengths in sunlight effective in producing skin cancer: A theoretical analysis. Proceedings of the National Academy of Sciences, 71(9):3363–3366, September 1974.

[78] Kenneth H Kraemer. Sunlight and skin cancer: another link revealed. Proceedings of the National Academy of Sciences, 94(1):11–14, 1997.

[79] John J DiGiovanna and Kenneth H Kraemer. Shining a light on xeroderma pigmentosum. Journal of investigative dermatology, 132(3):785–796, 2012.

[80] Luís F.Z. Batista, Bernd Kaina, Rogério Meneghini, and Carlos F.M. Menck. How DNA lesions are turned into powerful killing structures: Insights from UV-induced apoptosis. Mutation Research/Reviews in Mutation Research, 681(2-3):197–208, March 2009.

[81] Vivek Anantharaman, Eugene V Koonin, and L Aravind. Peptide-n-glycanases and dna repair proteins, xp-c/rad4, are, respectively, active and inactivated enzymes sharing a common transglutaminase fold. Human molecular genetics, 10(16):1627–1630, 2001.

[82] J.-H. Min and N.P. Pavletich. Crystal structure of the rad4-rad23 complex, October 2007.

[83] Xuejing Chen, Yogambigai Velmurugu, Guanqun Zheng, Beomseok Park, Yoonjung Shim, Youngchang Kim, Lili Liu, Bennett Van Houten, Chuan He, Anjum Ansari, et al. Kinetic gating mechanism of dna damage recognition by rad4/xpc. Nature communications, 6(1):5849, 2015.

[84] Josep Gelpi, Adam Hospital, Ramón Goñi, and Modesto Orozco. Molecular dynamics simulations: advances and applications. Advances and Applications in Bioinformatics and Chemistry, page 37, November 2015.

[85] Atanas G. Atanasov, Birgit Waltenberger, Eva-Maria Pferschy-Wenzig, Thomas Linder, Christoph Wawrosch, Pavel Uhrin, Veronika Temml, Limei Wang, Stefan Schwaiger, Elke H. Heiss, Judith M. Rollinger, Daniela Schuster, Johannes M. Breuss, Valery Bochkov, Marko D. Mihovilovic, Brigitte Kopp, Rudolf Bauer, Verena M. Dirsch, and Hermann Stuppner. Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotechnology Advances, 33(8):1582–1614, December 2015.

[86] Ankur Gupta and James B. Rawlings. Comparison of parameter estimation methods in stochastic chemical kinetic models: Examples in systems biology. AIChE Journal, 60(4):1253–1268, March 2014.

[87] Nicolas Le Novère. Quantitative and logic modelling of molecular and gene networks. Nature Reviews Genetics, 16(3):146–158, February 2015.

[88] Mahmoud A.A. Ibrahim, Alaa H.M. Abdelrahman, and Alaa M.A. Hassan. Identification of novel plasmodium falciparum PI4kb inhibitors as potential anti-malarial drugs: Homology modeling, molecular docking and molecular dynamics simulations. Computational Biology and Chemistry, 80:79–89, June 2019.

[89] Subhomoi Borkotoky, Debajit Dey, Zaved Hazarika, Amit Joshi, and Keshawanand Tripathi. Unravelling viral dynamics through molecular dynamics simulations - a brief overview. Biophysical Chemistry, 291:106908, December 2022.

[90] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. Journal of Chemical Information and Modeling, 51(1):69–82, November 2010.

[91] Weitong Ren, Hisham M. Dokainish, Ai Shinobu, Hiraku Oshima, and Yuji Sugita. Unraveling the coupling between conformational changes and ligand binding in ribose binding protein using multiscale molecular dynamics and free-energy calculations. The Journal of Physical Chemistry B, 125(11):2898–2909, March 2021.

[92] Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W. Lowe. Computational methods in drug discovery. Pharmacological Reviews, 66(1):334–395, December 2013.

[93] Andrés Pinzón, Emiliano Barreto, Adriana Bernal, Luke Achenie, Andres F González Barrios, Raúl Isea, and Silvia Restrepo. Computational models in plant-pathogen interactions: the case of phytophthora infestans. Theoretical Biology and Medical Modelling, 6(1), November 2009.

[94] Konstantinos Georgiadis, Selina Wray, Sébastien Ourselin, Jason D. Warren, and Marc Modat. Computational modelling of pathogenic protein spread in neurodegenerative diseases. PLOS ONE, 13(2):e0192518, February 2018.

[95] Sergei L Kosakovsky Pond, Steven Weaver, Andrew J Leigh Brown, and Joel O Wertheim. HIV-TRACE (TRAnsmission cluster engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Molecular Biology and Evolution, 35(7):1812–1819, January 2018.

[96] V. B. Sulimov, E. V. Katkova, I. V. Oferkin, A. V. Sulimov, A. N. Romanov, A. I. Roschin, I. B. Beloglazova, O. S. Plekhanova, V. A. Tkachuk, and V. A. Sadovnichiy. Application of molecular modeling to urokinase inhibitors development. BioMed Research International, 2014:1–15, 2014.

[97] Ahmed A. Al-Karmalawy and Muhammad Khattab. Molecular modelling of mebendazole polymorphs as a potential colchicine binding site inhibitor. New Journal of Chemistry, 44(33):13990–13996, 2020.

[98] Mahreen Arooj, Songmi Kim, Sugunadevi Sakkiah, Guang Ping Cao, Yuno Lee, and Keun Woo Lee. Molecular modeling study for inhibition mechanism of human chymase and its application in inhibitor design. PLoS ONE, 8(4):e62740, April 2013.

[99] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd – visual molecular dynamics. Journal of Molecular Graphics, 14:33–38, 1996.

[100] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. This is reference to PyMoL., November 2015.

[101] Marcus D Hanwell, Donald E Curtis, David C Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R Hutchison. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. Journal of Cheminformatics, 4(1), August 2012.

[102] Zheng Yang, Keren Lasker, Dina Schneidman-Duhovny, Ben Webb, Conrad C. Huang, Eric F. Pettersen, Thomas D. Goddard, Elaine C. Meng, Andrej Sali, and Thomas E. Ferrin. UCSF chimera, MODELLER, and IMP: An integrated modeling system. Journal of Structural Biology, 179(3):269–278, September 2012.

[103] Sadi Carnot. Reflections on the motive power of fire: And other papers on the second law of thermodynamics. Courier Corporation, 2012.

[104] Ludwig Boltzmann. Lectures on gas theory. Courier Corporation, 2012.

[105] Daan Frenkel and Berend Smit. Understanding molecular simulation: from algorithms to applications, volume 1. Elsevier, 2001.

[106] Ludwig Boltzmann. On certain questions of the theory of gases. Nature, 51(1322):413–415, 1895.

[107] K. Huang. Statistical Mechanics. Wiley, 1987. ISBN 9780471815181.

[108] David Case, Robin M. Betz, D.S. Cerutti, Thomas Cheatham, Thomas Darden, Robert Duke, T.J. Giese, Holger Gohlke, Andreas Götz, Nadine Homeyer, Saeed Izadi, Pawel Janowski, J Kaus,

Andriy Kovalenko, Tai-Sung Lee, S LeGrand, P Li, C Lin, Tyler Luchko, and Peter A. Kollman. Amber 16, university of california, san francisco., April 2016.

[109] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. Journal of Computational Chemistry, pages NA–NA, 2009.

[110] Wilfred F. van Gunsteren, Xavier Daura, and Alan E. Mark. GROMOS force field, September 1998.

[111] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. Journal of the American Chemical Society, 118(45):11225–11236, November 1996.

[112] Stewart A Adcock and J Andrew McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. Chemical reviews, 106(5):1589–1615, 2006.

[113] Andrew R Leach and Andrew R Leach. Molecular modelling: principles and applications. Pearson education, 2001.

[114] Frank L. Somer. Molecular modelling for beginners (alan hinchliffe). Journal of Chemical Education, 81(11):1573, November 2004.

[115] H. Bernhard Schlegel. Optimization of equilibrium geometries and transition structures. Journal of Computational Chemistry, 3(2):214–218, 1982.

[116] B. J. Alder and T. E. Wainwright. Phase transition for a hard sphere system. The Journal of Chemical Physics, 27(5):1208–1209, November 1957.

[117] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. i. general method. The Journal of Chemical Physics, 31(2):459–466, August 1959.

[118] A. Rahman. Correlations in the motion of atoms in liquid argon. Physical Review, 136(2A): A405–A411, October 1964.

[119] Frank H. Stillinger and Aneesur Rahman. Improved simulation of liquid water by molecular dynamics. The Journal of Chemical Physics, 60(4):1545–1557, February 1974.

[120] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. Nature, 267(5612):585–590, June 1977.

[121] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. The Journal of Chemical Physics, 76(1):637–649, January 1982.

[122] Mark Abraham, Andrey Alekseenko, Cathrine Bergh, Christian Blau, Eliane Briand, Mahesh Doijade, Stefan Fleischmann, Vytautas Gapsys, Gaurav Garg, Sergey Gorelov, Gilles Gouaillardet, Alan Gray, M. Eric Irrgang, Farzaneh Jalalypour, Joe Jordan, Christoph Junghans, Prashanth Kanduri, Sebastian Keller, Carsten Kutzner, Justin A. Lemkul, Magnus Lundborg, Pascal Merz, Vedran Miletić, Dmitry Morozov, Szilárd Páll, Roland Schulz, Michael Shirts, Alexey Shvetsov, Bálint Soproni, David Van Der Spoel, Philip Turner, Carsten Uphoff, Alessandra Villa, Sebastian Wingbermühle, Artem Zhmurov, Paul Bauer, Berk Hess, and Erik Lindahl. GROMACS 2023.2 Manual, 2023.

[123] James C. Phillips, David J. Hardy, Julio D. C. Maia, John E. Stone, João V. Ribeiro, Rafael C. Bernardi, Ronak Buch, Giacomo Fiorin, Jérôme Hénin, Wei Jiang, Ryan McGreevy, Marcelo C. R. Melo, Brian K. Radak, Robert D. Skeel, Abhishek Singharoy, Yi Wang, Benoît Roux, Aleksei Aksimentiev, Zaida Luthey-Schulten, Laxmikant V. Kalé, Klaus Schulten, Christophe Chipot, and Emad Tajkhorshid. Scalable molecular dynamics on CPU and GPU architectures with NAMD. The Journal of Chemical Physics, 153(4), July 2020.

[124] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. Journal of Computational Chemistry, 29(11): 1859–1865, March 2008.

[125] John G. Kirkwood. Statistical mechanics of fluid mixtures. The Journal of Chemical Physics, 3 (5):300–313, May 1935.

[126] IR McDonald and K Singer. Machine calculation of thermodynamic properties of a simple fluid at supercritical temperatures. J. Chem. Phys., 47(11):4766–4772, 1967.

[127] IR McDonald and K Singer. Examination of the adequacy of the 12–6 potential for liquid argon by means of monte carlo calculations. J. Chem. Phys., 50(6):2308–2315, 1969.

[128] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. J. Comput. Phys., 23(2):187–199, 1977.

[129] Glenn M Torrie and John P Valleau. Monte carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid. Chem. Phys. Lett., 28(4):578–581, 1974.

[130] Qinghua Liao. Enhanced sampling and free energy calculations for protein simulations. In Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly, pages 177–213. Elsevier, 2020.

[131] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. Journal of Computational Chemistry, 13(8):1011–1021, October 1992.

[132] Alan Grossfield. Wham: the weighted histogram analysis method, 2012. , version 2.0.7, http://membrane.urmc.rochester.edu/wordpress/?page_id=126.

[133] Johannes Kästner. Umbrella sampling. Wiley Interdisciplinary Reviews: Computational Molecular Science, 1(6):932–942, May 2011.

[134] Richard J Loncharich, Bernard R Brooks, and Richard W Pastor. Langevin dynamics of peptides: The frictional dependence of isomerization rates of n-acetylalanyl-n′-methylamide. Biopolymers: Original Research on Biomolecules, 32(5):523–535, 1992.

[135] W.F. van Gunsteren and H.J.C. Berendsen. Algorithms for brownian dynamics. Molecular Physics, 45(3):637–647, February 1982.

[136] Richard W Pastor, Bernard R Brooks, and Attila Szabo. An analysis of the accuracy of langevin and molecular dynamics algorithms. Molecular Physics, 65(6):1409–1419, 1988.

[137] Hans C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. The Journal of chemical physics, 72(4):2384–2393, 1980.

[138] JM Haile and HW Graben. Molecular dynamics simulations extended to various ensembles. i. equilibrium properties in the isoenthalpic–isobaric ensemble. The Journal of Chemical Physics, 73(5):2412–2419, 1980.

[139] Jean-Paul Ryckaert and Giovanni Ciccotti. Introduction of andersen's demon in the molecular dynamics of systems with constraints. The Journal of Chemical Physics, 78(12):7368–7374, 1983.

[140] M Parrinello and A Rahman. Strain fluctuations and elastic constants. The Journal of Chemical Physics, 76(5):2662–2666, 1982.

[141] M Parrinello, A Rahman, and P Vashishta. Structural transitions in superionic conductors. Physical review letters, 50(14):1073, 1983.

[142] Shuichi Nosé and ML Klein. Constant pressure molecular dynamics for molecular systems. Molecular Physics, 50(5):1055–1076, 1983.

[143] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. J. Chem. Phys., 81(8):3684–3690, 1984.

[144] Shein-shion Wang and JA Krumhansl. Superposition assumption. ii. high density fluid argon. The Journal of Chemical Physics, 56(9):4287–4290, 1972.

[145] David J Adams. Computer simulation of ionic systems: The distorting effects of the boundary conditions. Chemical Physics Letters, 62(2):329–332, 1979.

[146] Max Berkowitz and J Andrew McCammon. Molecular dynamics with stochastic boundary conditions. Chemical Physics Letters, 90(3):215–217, 1982.

[147] C. L. Brooks, Axel Brünger, and M. Karplus. Active site dynamics in protein molecules: A stochastic boundary molecular-dynamics approach. Biopolymers, 24(5):843–865, May 1985.

[148] Alan C Belch and M Berkowitz. Molecular dynamics simulations of tips2 water restricted by a spherical hydrophobic boundary. Chemical physics letters, 113(3):278–282, 1985.

[149] Dmitrii Beglov and Benoit Roux. Finite representation of an infinite bulk system: solvent boundary potential for computer simulations. The Journal of chemical physics, 100(12):9050–9063, 1994.

[150] R Dickman, G Stell, G Hummer, and L Pratt. Treatment of electrostatic interactions in computer simulations of condensed media. In Proceedings of the Conference, Santa Fe, 1999.

[151] Dennis C Rapaport. The art of molecular dynamics simulation. Cambridge university press, 2004.

[152] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. Physical Review, 159(1):98–103, July 1967.

[153] P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. Annalen der Physik, 369(3):253–287, 1921.

[154] WF Van Gunsteren and Herman JC Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. Molecular Physics, 34(5):1311–1327, 1977.

[155] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of computational physics, 23(3):327–341, 1977.

[156] Hans C Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. Journal of computational Physics, 52(1):24–34, 1983.

[157] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: a linear constraint solver for molecular simulations. Journal of computational chemistry, 18(12):1463–1472, 1997.

[158] Ramzi Kutteh. New methods for incorporating nonholonomic constraints into molecular dynamics simulations. The Journal of chemical physics, 111(4):1394–1406, 1999.

[159] Thomas Kelly. Evidence. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[160] Richard Feldman. Evidentialism, higher-order evidence, and disagreement. Episteme, 6(3): 294–312, 2009.

[161] David Christensen. Higher-order evidence1. Philosophy and Phenomenological Research, 81(1): 185–215, May 2010.

[162] Harvey Lodish, Arnold Berk, Chris A Kaiser, Monty Krieger, Matthew P Scott, Anthony Bretscher, Hidde Ploegh, and Paul Matsudaira. Molecular Cell Biology, 6E. Macmillan, 2008.

[163] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. Nature, 489(7414):101–108, 2012.

[164] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Essential Cell Biology. Garland Science, 2013.

[165] A. Oeckinghaus and S. Ghosh. The NF- b family of transcription factors and its regulation. Cold Spring Harbor Perspect. Biol., 1(4):a000034–a000034, September 2009.

[166] W Brown, A Ralston, and K Shaw. Positive transcription control: The glucose effect. Nature Education, 1(1):202, 2008.

[167] O. Hobert. Gene regulation by transcription factors and MicroRNAs. Science, 319(5871):1785–1786, March 2008.

[168] Agustino Martínez-Antonio and Julio Collado-Vides. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr. Opin. Microbiol., 6(5):482–489, October 2003.

[169] Venkata Rajesh Yella, Devesh Bhimsaria, Debostuti Ghoshdastidar, José A Rodríguez-Martínez, Aseem Z Ansari, and Manju Bansal. Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. Nucleic Acids Res., 46(22):11883–11897, November 2018.

[170] M. A. Cleary, P. S. Pendergrast, and W. Herr. Structural flexibility in transcription complex formation revealed by protein-DNA photocrosslinking. Proc. Natl. Acad. Sci. U.S.A., 94(16): 8450–8455, August 1997.

[171] Søren Lindemose, Peter Eigil Nielsen, Poul Valentin-Hansen, and Niels Erik Møllegaard. A novel indirect sequence readout component in theE. coliCyclic AMP receptor protein operator. ACS Chemical Biology, 9(3):752–760, January 2014.

[172] Murat Tuğrul, Tiago Paixão, Nicholas H. Barton, and Gašper Tkačik. Dynamics of transcription factor binding site evolution. PLos Genet., 11(11):e1005639, November 2015.

[173] Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Griffin Chure, Tal Einav, Mitchell Lewis, and Rob Phillips. Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction. Cell Systems, 6(4):456–469.e10, April 2018.

[174] Yongping Pan, Chung-Jung Tsai, Buyong Ma, and Ruth Nussinov. How do transcription factors select specific binding sites in the genome? Nat. Struct. Mol., 16(11):1118–1120, November 2009.

[175] David West, Roy Williams, Virgil Rhodius, Andrew Bell, Naveen Sharma, Chao Zou, Nobuyuki Fujita, Akira Ishihama, and Stephen Busby. Interactions between the escherichia coli cyclic AMP receptor protein and RNA polymerase at class II promoters. Mol. Microbiol., 10(4):789–797, November 1993.

[176] Wei Niu, Younggyu Kim, Gregory Tau, Tomasz Heyduk, and Richard H Ebright. Transcription activation at class II CAP-dependent promoters: Two interactions between CAP and RNA polymerase. Cell, 87(6):1123–1134, December 1996.

[177] Ruth Nussinov, Chung-Jung Tsai, and Jin Liu. Principles of allosteric interactions in cell signaling. J. Am. Chem. Soc., 136(51):17692–17701, December 2014.

[178] Ruth Nussinov and Chung-Jung Tsai. Allostery in disease and in drug discovery. Cell, 153(2): 293–305, April 2013.

[179] Meng-Xi Zhao, Yong-Liang Jiang, Yong-Xing He, Yi-Fei Chen, Yan-Bin Teng, Yuxing Chen, Cheng-Cai Zhang, and Cong-Zhao Zhou. Structural basis for the allosteric control of the global transcription factor NtcA by the nitrogen starvation signal 2-oxoglutarate. Proc. Natl. Acad. Sci. U.S.A., 107(28):12487–12492, June 2010.

[180] Ron O. Dror, Hillary F. Green, Celine Valant, David W. Borhani, James R. Valcourt, Albert C. Pan, Daniel H. Arlow, Meritxell Canals, J. Robert Lane, Raphaël Rahmani, Jonathan B. Baell,

Patrick M. Sexton, Arthur Christopoulos, and David E. Shaw. Structural basis for modulation of a g-protein-coupled receptor by allosteric drugs. Nature, 503(7475):295–299, October 2013.

[181] Gregory R Bowman, Eric R Bolin, Kathryn M Hart, Brendan C Maguire, and Susan Marqusee. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. Proc. Natl. Acad. Sci. U.S.A., 112(9):2734–2739, 2015.

[182] Amy Hauck Newman, Thijs Beuming, Ashwini K Banala, Prashant Donthamsetti, Katherine Pongetti, Alex LaBounty, Benjamin Levy, Jianjing Cao, Mayako Michino, Robert R Luedtke, et al. Molecular determinants of selectivity and efficacy at the dopamine d3 receptor. J. Med. Chem., 55(15):6689–6699, 2012.

[183] Julie R Schames, Richard H Henchman, Jay S Siegel, Christoph A Sotriffer, Haihong Ni, and J Andrew McCammon. Discovery of a novel binding trench in hiv integrase. J. Med. Chem., 47 (8):1879–1881, 2004.

[184] Yaw Sing Tan, Paweł Śledź, Steffen Lang, Christopher J Stubbs, David R Spring, Chris Abell, and Robert B Best. Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1. Angew. Chem. Int. Ed., 51(40):10078–10081, 2012.

[185] Marco Berrera, Sergio Pantano, and Paolo Carloni. Catabolite activator protein in aqueous solution: a molecular simulation study. J. Phys. Chem. B, 111(6):1496–1501, February 2007.

[186] Ewa Heyduk, Tomasz Heyduk, and James C. Lee. Intersubunit communications in escherichia coli cyclic AMP receptor protein: studies of the ligand binding domain. Biochemistry, 31(14): 3682–3688, April 1992.

[187] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol., 234(3):779–815, 1993.

[188] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. J. Chem. Theory Comput., 11(8):3696–3713, 2015.

[189] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther, Pau Andrio, Ramon Goñi, and Alexandra Balaceanu. Parmbsc1: a refined force field for dna simulations. Nat. Methods, 13(1):55, 2016.

[190] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys., 79(2):926–935, 1983.

[191] Béla Voß, Reinhard Seifert, U Benjamin Kaupp, and Helmut Grubmüller. A quantitative model for camp binding to the binding domain of mlok1. Biophys. J., 111(8):1668–1678, 2016.

[192] Abhinandan Panigrahi. A computational study of protein-dna binding during dna damage repair and gene activation. Master's thesis, International Institute of Information Technology, Hyderabad., 2020. Accessed: 2020-01-31.

[193] Daniel R. Roe and Thomas E. Cheatham. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. Journal of Chemical Theory and Computation, 9 (7):3084–3095, June 2013.

[194] Susan S. Taylor, Choel Kim, Cecilia Y. Cheng, Simon H.J. Brown, Jian Wu, and Natarajan Kannan. Signaling through cAMP and cAMP-dependent protein kinase: Diverse strategies for drug design. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1784(1):16–26, January 2008.

[195] Y Matsushita, T Murakawa, K Shimamura, M Oishi, T Ohyama, and N Kurita. Specific interactions between dna and regulatory protein controlled by ligand-binding: Ab initio molecular simulation. In AIP Conference Proceedings, volume 1649, pages 121–129. American Institute of Physics, 2015. 1.

[196] Marie Christine Vaney, Gary L. Gilliland, James G. Harman, Alan Peterkofsky, and Irene T. Weber. Crystal structure of a cAMP-independent form of catabolite gene activator protein with adenosine substituted in one of two cAMP-binding sites. Biochemistry, 28(11):4568–4574, May 1989.

[197] E. Fic, P. Bonarek, A. Gorecki, S. Kedracka-Krok, J. Mikolajczak, A. Polit, M. Tworzydlo, M. Dziedzicka-Wasylewska, and Z. Wasylewski. cAMP receptor protein from escherichia coli as a model of signal transduction in proteins & a review. Journal of Molecular Microbiology and Biotechnology, 17(1):1–11, 2009.

[198] Alexander Hollaender. Effect of long ultraviolet and short visible radiation (3500 to 4900Å) on escherichia coli. Journal of Bacteriology, 46(6):531–541, 1943.

[199] H. Park, K. Zhang, Y. Ren, S. Nadji, N. Sinha, J.-S. Taylor, and C. Kang. Crystal structure of a DNA decamer containing a cis-syn thymine dimer. Proceedings of the National Academy of Sciences, 99(25):15965–15970, November 2002.

[200] I. Husain, J. Griffith, and A. Sancar. Thymine dimers bend DNA. Proceedings of the National Academy of Sciences, 85(8):2558–2562, April 1988.

[201] David A. Pearlman, Stephen R. Holbrook, David H. Pirkle, and Sung-Hou Kim. Molecular models for DNA damaged by photoreaction. Science, 227(4692):1304–1308, March 1985.

[202] Errol C Friedberg, Graham C Walker, Wolfram Siede, and Richard D Wood. DNA repair and mutagenesis. American Society for Microbiology Press, 2005.

[203] Renata MA Costa, Vanessa Chigançcas, Rodrigo da Silva Galhardo, Helotonio Carvalho, and Carlos FM Menck. The eukaryotic nucleotide excision repair pathway. Biochimie, 85(11):1083–1099, 2003.

[204] Luca Proietti De Santis, Claudia Lorenti Garcia, Adayabalam S Balajee, Paolo Latini, Pietro Pichierri, Osamu Nikaido, Miria Stefanini, and Fabrizio Palitti. Transcription coupled repair efficiency determines the cell cycle progression and apoptosis after uv exposure in hamster cells. DNA repair, 1(3):209–223, 2002.

[205] James M Ford. Regulation of dna damage recognition and nucleotide excision repair: another role for p53. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 577 (1-2):195–202, 2005.

[206] RB Setlow and WL Carrier. The disappearance of thymine dimers from dna: an error-correcting mechanism. Proceedings of the National Academy of Sciences of the United States of America, 51(2):226, 1964.

[207] Richard P Boyce and Paul Howard-Flanders. Release of ultraviolet light-induced thymine dimers from dna in e. coli k-12. Proceedings of the National Academy of Sciences of the United States of America, 51(2):293, 1964.

[208] David Pettijohn and Philip Hanawalt. Evidence for repair-replication of ultraviolet damaged dna in bacteria. Journal of molecular biology, 9(2):395–410, 1964.

[209] M. T. Hess, U. Schwitter, M. Petretta, B. Giese, and H. Naegeli. Bipartite substrate discrimination by human nucleotide excision repair. Proceedings of the National Academy of Sciences, 94(13): 6664–6669, June 1997.

[210] Kaoru Sugasawa, Tomoko Okamoto, Yuichiro Shimizu, Chikahide Masutani, Shigenori Iwai, and Fumio Hanaoka. A multistep damage recognition mechanism for global genomic nucleotide excision repair. Genes & Development, 15(5):507–521, March 2001.

[211] Kaoru Sugasawa, Jessica M.Y Ng, Chikahide Masutani, Shigenori Iwai, Peter J van der Spek, André P.M Eker, Fumio Hanaoka, Dirk Bootsma, and Jan H.J Hoeijmakers. Xeroderma pigmentosum group c protein complex is the initiator of global genome nucleotide excision repair. Molecular Cell, 2(2):223–232, August 1998.

[212] Marcel Volker, Martijn J Moné, Parimal Karmakar, Anneke van Hoffen, Wouter Schul, Wim Vermeulen, Jan H.J Hoeijmakers, Roel van Driel, Albert A van Zeeland, and Leon H.F Mullenders. Sequential assembly of the nucleotide excision repair factors in vivo. Molecular Cell, 8(1): 213–224, July 2001.

[213] T. Riedl. The comings and goings of nucleotide excision repair factors on damaged DNA. The EMBO Journal, 22(19):5293–5303, October 2003.

[214] T. Nouspikel. DNA repair in mammalian cells. Cellular and Molecular Life Sciences, 66(6): 994–1009, January 2009.

[215] Ulrike Camenisch, Daniel Träutlein, Flurina C Clement, Jia Fei, Alfred Leitenstorfer, Elisa Ferrando-May, and Hanspeter Naegeli. Two-stage dynamic dna quality check by xeroderma pigmentosum group c protein. The EMBO journal, 28(16):2387–2399, 2009.

[216] Muwen Kong, Lili Liu, Xuejing Chen, Katherine I. Driscoll, Peng Mao, Stefanie Böhm, Neil M. Kad, Simon C. Watkins, Kara A. Bernstein, John J. Wyrick, Jung-Hyun Min, and Bennett Van Houten. Single-molecule imaging reveals that rad4 employs a dynamic dna damage recognition process. Molecular cell, 64(2):376–387, 2016.

[217] Stephen E Halford and John F Marko. How do site-specific dna-binding proteins find their targets? Nucleic acids research, 32(10):3040–3052, 2004.

[218] Otto G Berg, Robert B Winter, and Peter H Von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. models and theory. Biochemistry, 20(24):6929–6948, 1981.

[219] Paul C Blainey, Antoine M van Oijen, Anirban Banerjee, Gregory L Verdine, and X Sunney Xie. A base-excision dna-repair protein finds intrahelical lesion bases by fast sliding in contact with dna. Proceedings of the National Academy of Sciences, 103(15):5752–5757, 2006.

[220] Yogambigai Velmurugu, Xuejing Chen, Phillip Slogoff Sevilla, Jung-Hyun Min, and Anjum Ansari. Twist-open mechanism of dna damage recognition by the rad4/xpc nucleotide excision repair complex. Proceedings of the National Academy of Sciences, 113(16):E2296–E2305, 2016.

[221] Patrick Calsou and Bernard Salles. Properties of damage-dependent dna incision by nucleotide excision repair in human cell-free extracts. Nucleic acids research, 22(23):4937–4942, 1994.

[222] Elizabeth Evans, Jane Fellows, Arnold Coffer, and Richard D Wood. Open complex formation around a lesion during nucleotide excision repair provides a structure for cleavage by human xpg protein. The EMBO journal, 16(3):625–638, 1997.

[223] Elizabeth Evans, Jonathan G Moggs, Jae R Hwang, Jean-Marc Egly, and Richard D Wood. Mechanism of open complex and dual incision formation by human nucleotide excision repair factors. The EMBO journal, 16(21):6559–6573, 1997.

[224] David Mu, Mitsuo Wakasugi, David S Hsu, and Aziz Sancar. Characterization of reaction intermediates of human excision repair nuclease. Journal of Biological Chemistry, 272(46):28971–28979, 1997.

[225] Fr'ed'eric Coin, Valentyn Oksenych, and Jean-Marc Egly. Distinct roles for the xpb/p52 and xpd/p44 subcomplexes of tfiih in damaged dna opening during nucleotide excision repair. Molecular cell, 26(2):245–256, 2007.

[226] John Bradsher, Frederic Coin, and Jean-Marc Egly. Distinct roles for the helicases of tfiih in transcript initiation and promoter escape. Journal of Biological Chemistry, 275(4):2532–2538, 2000.

[227] G Sebastiaan Winkler, Sofia J Araújo, Ulrike Fiedler, Wim Vermeulen, Frederic Coin, Jean-Marc Egly, Jan HJ Hoeijmakers, Richard D Wood, H Th Marc Timmers, and Geert Weeda. Tfiih with inactive xpd helicase functions in transcription initiation but is defective in dna repair. Journal of Biological Chemistry, 275(6):4258–4266, 2000.

[228] Li Fan, Jill O Fuss, Quen J Cheng, Andrew S Arvai, Michal Hammel, Victoria A Roberts, Priscilla K Cooper, and John A Tainer. Xpd helicase structures and activities: insights into the cancer and aging phenotypes from xpd mutations. Cell, 133(5):789–800, 2008.

[229] Huanting Liu, Jana Rudolf, Kenneth A Johnson, Stephen A McMahon, Muse Oke, Lester Carter, Anne-Marie McRobbie, Sara E Brown, James H Naismith, and Malcolm F White. Structure of the dna repair helicase xpd. Cell, 133(5):801–812, 2008.

[230] Stefanie C Wolski, Jochen Kuper, Petra Hänzelmann, James J Truglio, Deborah L Croteau, Bennett Van Houten, and Caroline Kisker. Crystal structure of the fes cluster–containing nucleotide excision repair helicase xpd. PLoS biology, 6(6):e149, 2008.

[231] Nadine Mathieu, Nina Kaczmarek, and Hanspeter Naegeli. Strand-and site-specific dna lesion demarcation by the xeroderma pigmentosum group d helicase. Proceedings of the National Academy of Sciences, 107(41):17545–17550, 2010.

[232] Jochen Kuper, Stefanie C Wolski, Gudrun Michels, and Caroline Kisker. Functional and structural studies of the nucleotide excision repair helicase xpd suggest a polarity for dna translocation. The EMBO journal, 31(2):494–502, 2012.

[233] Robert A Pugh, Colin G Wu, and Maria Spies. Regulation of translocation polarity by helicase domain 1 in sf2b helicases. The EMBO journal, 31(2):503–514, 2012.

[234] Juch-Chin Huang, Daniel L Svoboda, Joyce T Reardon, and Aziz Sancar. Human nucleotide excision nuclease removes thymine dimers from dna by incising the 22nd phosphodiester bond 5'and the 6th phosphodiester bond 3'to the photodimer. Proceedings of the National Academy of Sciences, 89(8):3664–3668, 1992.

[235] Jonathan G Moggs, Kevin J Yarema, John M Essigmann, and Richard D Wood. Analysis of incision sites produced by human cell extracts and purified proteins during nucleotide excision repair of a 1, 3-intrastrand d(gptpg)-cisplatin adduct (*). Journal of Biological Chemistry, 271 (12):7177–7186, 1996.

[236] Anne O'Donovan, Adelina A Davies, Jonathan G Moggs, Stephen C West, and Richard D Wood. Xpg endonuclease makes the 3' incision in human dna nucleotide excision repair. Nature, 371 (6496):432–435, 1994.

[237] Angelos Constantinou, Daniela Gunz, Elizabeth Evans, Philippe Lalle, Paul A Bates, Richard D Wood, and Stuart G Clarkson. Conserved residues of human xpg protein important for nuclease activity and function in nucleotide excision repair. Journal of Biological Chemistry, 274(9): 5637–5648, 1999.

[238] Odilia Popanda and Heinz Walter Thielmann. The function of dna polymerases in dna repair synthesis of ultraviolet-irradiated human fibroblasts. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 1129(2):155–160, 1992.

[239] Silvano Nocentini. Rejoining kinetics of DNA single- and double-strand breaks in normal and DNA ligase-deficient cells after exposure to ultraviolet c and gamma radiation: An evaluation of ligating activities involved in different DNA repair processes. Radiation Research, 151(4):423, April 1999.

[240] Abhinandan Panigrahi, Hemanth Vemuri, Madhur Aggarwal, Kartheek Pitta, and Marimuthu Krishnan. Sequence specificity, energetics and mechanism of mismatch recognition by DNA damage sensing protein rad4/XPC. Nucleic Acids Research, 48(5):2246–2257, February 2020.

[241] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF chimera?a visualization system for exploratory research and analysis. Journal of Computational Chemistry, 25(13):1605–1612, 2004.

[242] Thomas J. Macke and David A. Case. Modeling unusual nucleic acid structures. In ACS Symposium Series, pages 379–393. American Chemical Society, July 1997.

[243] Jong-Ki Kim, Dinshaw Patel, and Byong-Seok Choi. Contrasting structural impacts induced by cis-syn cyclobutane dimer and (6–4) adduct in dna duplex decamers: implication in mutagenesis and repair activity. Photochemistry and photobiology, 62(1):44–50, 1995.

[244] Joon-Hwa Lee, Chin-Ju Park, Jae-Sun Shin, Takahisa Ikegami, Hideo Akutsu, and Byong-Seok Choi. Nmr structure of the dna decamer duplex containing double t· g mismatches of cis–syn cyclobutane pyrimidine dimer: implications for dna damage recognition by the xpc–hhr23b complex. Nucleic Acids Research, 32(8):2474–2481, 2004.

[245] Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. Journal of Chemical Theory and Computation, 9(9):3878–3888, August 2013.

[246] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n· log (n) method for ewald sums in large systems. J. Chem. Phys., 98(12):10089–10092, 1993.

[247] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. J. Chem. Phys., 76(1):637–649, 1982.

[248] Jung-Hyun Min and Nikola P. Pavletich. Recognition of DNA damage by the rad4 nucleotide excision repair protein. Nature, 449(7162):570–575, September 2007.