Computational Investigation of Binding-induced Conformational Transitions in Proteins: Application of Minimum Energy Path Analysis and Enhanced Sampling Methods

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

Master of Science
in
Computational Natural Sciences By Research

by

Madhur Aggarwal
201564235
madhur.aggarwal@research.iiit.ac.in

Madhur Aggarwal
201564235
madhur.aggarwal@research.iiit.ac.in

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

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Computational Investigation of Binding-induced Conformational Transitions in Proteins: Application of Minimum Energy Path Analysis and Enhanced Sampling Methods**" by **Madhur Aggarwal**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____   _____

Date   Adviser: Prof. Marimuthu Krishnan

To the time ahead

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my mentor and advisor, Prof. Marimuthu Krishnan. I am grateful for him for giving me an opportunity to work on some good scientific problems, which have been compiled in this thesis. He makes sure to have a nice healthy working environment in his research group. He always makes sure to extend his helping hand for guidance and support whenever needed. He also makes sure everyone's contributions to projects and group discussions are truly valued and appreciated. This gives good motivation to everyone in our research group. By working alongside Prof. Krishnan, I have learned to be more punctual and productive. My work ethics have improved significantly and I am now capable of managing my work in better ways. These skills will help me forever.

I would also like to extend a big thank you to all the members of our research group, Mohan sir, Kartheek sir, Abhinandan, Bhagesh, Hemanth, Akshay, Vishal, Shaunak, Nikhil, Keshavan and Aadarsh. I will always cherish the group meetings, late-night research discussions in our lab and all the fun we had while working and collaborating on different projects.

I specifically would like to mention my co-authors, Kartheek sir, Abhinandan and Hemanth for the DNA damage project and Shaunak and Aadarsh for the kinase project. Both of these projects were great learning opportunities for me, and I believe our results published in these research papers will be well-cited and well-discussed in the future. I believe these research papers are tiny droplets we have added to the vast scientific literature but I also believe the right audience will find our methods and results intriguing and insightful. I hope our research papers will inspire researchers to work on similar systems in the future.

Now, I would like to thank wholeheartedly to my friends, Srijan, Kanak, Pritish, Namit, Anam, Apaar and Anirudh. Thank you for always being there for me. College life would have been boring if I hadn't found you guys.

Lastly, I would like to thank my entire family who's constant guidance, love and support has helped me achieve this.

Abstract

The conformational dynamics of proteins is essential for biological functions. The characterization of binding-induced changes in protein conformational dynamics is of paramount importance to understand the fundamental processes of life. Molecular modelling and molecular dynamics simulations have emerged as useful computational tools to probe atomistic details of molecular recognition and binding-induced dynamical responses of proteins. The interactions of proteins with DNA and small drug molecules play crucial roles in regulating various cellular processes. The present thesis reports results of investigation of the functional roles of protein conformational dynamics in the following two important biological systems: (a) a DNA repair protein complexed with a mismatched DNA and (b) an inhibitor drug-bound human c-Src kinase. In the former system, our primary objective is to understand the molecular mechanism of how a specific DNA repair protein (RAD4/XPC) binds to a damaged DNA, recognizes lesions and repair them efficiently using molecular dynamics and enhanced sampling simulations. In the latter system, we have investigated drug-induced conformational transitions in human c-Src kinase to understand its activation mechanism using molecular dynamics simulations and nudged elastic band method.

DNA damage caused by ultraviolet (UV) radiation can lead to a range of genetic skin conditions and cancers. The protein Rad4/XPC recognizes and repairs these types of DNA damage with high accuracy to safeguard the integrity of the genome. However, the precise mechanism of how Rad4/XPC recognizes DNA damage in the crowded cellular milieu is not yet fully understood. The first part of this thesis investigates the mechanism, energetics, dynamics, and molecular basis of Rad4/XPC's recognition with base pair mismatched DNA lesions. In particular, the mechanism of association of RAD4 with the mismatched DNA is investigated using molecular dynamics and umbrella sampling simulations. The Rad4-DNA association free energy surfaces are determined for three Rad4-DNA complexes (CCC/CCC, TTT/TTT, and TAT/TAT) and the sequence-dependent specificity of Rad4 for mismatched DNA is explored. The results reveal that Rad4 exhibits higher specificity for CCC/CCC mismatched DNA compared to the other two mismatches. The key molecular interactions contributing to Rad4-DNA association and specificity are characterized.

Protein kinases act as biological switches that toggle between ON (active) and OFF (inactive) states in response to specific cellular signals to regulate various cellular processes including growth, cell proliferation and differentiation. Since protein kinases are implicated in tumorigenesis and malignant cell proliferation, they have become favourable targets for cancer therapy. The second part of the thesis focuses on drug-induced conformational transition of c-Src kinase between active (drug-bound) and inactive (drug-free) states using the minimum energy path analysis and molecular dynamics simulation. The crucial role of hydration of the binding pocket in the association/dissociation of drug from the kinase is explored. The coupling between the entry and exit of the drug, influx of water molecules into

the binding pocket and critical interactions of key kinase residues near the binding pocket appears to influence the conformational transition between the active and inactive states of the kinase. We believe that the molecular mechanisms elucidated will be useful for rational design of novel therapeutics against cancer and other kinase-related diseases.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

Proteins and nucleic acids are important biomacromolecules that play crucial roles in the structure and function of cells. Deoxyribonucleic acid (DNA) is a chain molecule that carries the genetic information needed for making all the proteins and other molecules that are necessary to drive the cellular machinery. Proteins, on the other hand, are large complex molecules that are involved in virtually all aspects of cellular function, from providing structural support to catalyzing chemical reactions. DNA and proteins interact with each other to give life to cells. Understanding how some of the processes surrounding DNA and proteins occur in biological cells is therefore important for diseases such as cancer.

Proteins are dynamical systems. They exhibit a wide spectrum of dynamics such as fast bond vibrations, side chain motions, loop motions and large domain motions [1–4]. It is important to examine the functional role of fast and slow dynamics of proteins in order to understand how the biological cellular processes are governed by these motions.

The rugged potential energy surface (PES) of proteins consists of multiple energy basins of differing depths and widths separated by barriers of differing heights. The energy basins on PES represent stable configurations of proteins and the energy maxima correspond to unstable (or transition state) protein configurations. The conformational dynamics of proteins can be thought of as transitions between different energy basins via a network of metastable intermediate states. Upon binding to drug molecules or other molecules such as DNA, the protein's PES is altered and consequently the nature of its conformational transitions also changes. Given higher energy barriers on PES, the computational characterisation

of protein conformational transitions is nontrivial and challenging. The conventional molecular dynamics simulations of limited duration fail to capture the entirety of slow protein conformational transitions due to inadequate sampling of the phase space of the system. In such MD simulations, the system tends to be stuck in a stable energy well and the timescales of traditional molecular dynamics techniques deem insufficient to capture barrier crossing events. Enhanced sampling methods like umbrella sampling, metadynamics and path finding techniques like minimum energy path analysis are therefore needed to study binding-induced conformational dynamics of protein and protein-DNA complexes.

In this thesis, we are interested in studying such binding-induced conformational transitions in kinases and protein-DNA systems using minimum energy path analysis and enhanced sampling methods. The two systems of interest are : (a) RAD4/XPC DNA damage sensing protein binding with DNA, and (b) Type-1 drug (Bosutinib)-bound c-Src kinase. In the following sections, we briefly introduce these two model systems and discuss some structural features of the molecular components of these model systems.

1.1 DNA

Deoxyribonucleic acid (DNA) is a long molecule that contains the genetic instructions used in the development and functioning of all known living organisms and many viruses. DNA is a double-stranded molecule that is organized into segments called genes, which are made up of long sequences of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T) (see Figure 1.1). These bases are paired together in specific ways, with A always bonding with T and C always bonding with G, to form the "rungs" of the DNA ladder. This specific sequence of bases in a DNA molecule carries the genetic information that determines the traits and characteristics of an organism [5–7].

The DNA molecule is incredibly stable, but it can also be replicated, or copied, accurately to produce new DNA molecules. This process is essential for the reproduction of cells and for the transmission of genetic information from one generation to the next [8–10]. In addition to its role in carrying genetic information, DNA also plays a crucial role in the functioning of cells. It is involved in the regulation of gene expression, which determines which genes are turned on or off in a particular cell. This helps to control the development and specialization of cells, ensuring that they can perform their specific functions in the body [11]. Overall, DNA is an essential component of life, and our understanding of its structure and function has advanced significantly over the past few decades. This has led to important medical and technological advances, including the development of genetic engineering and personalized medicine [12].

DNA damage refers to any type of change or mutation that occurs in the DNA sequence of a cell. This can happen due to a variety of reasons, including exposure to certain chemicals, radiation, or even as a
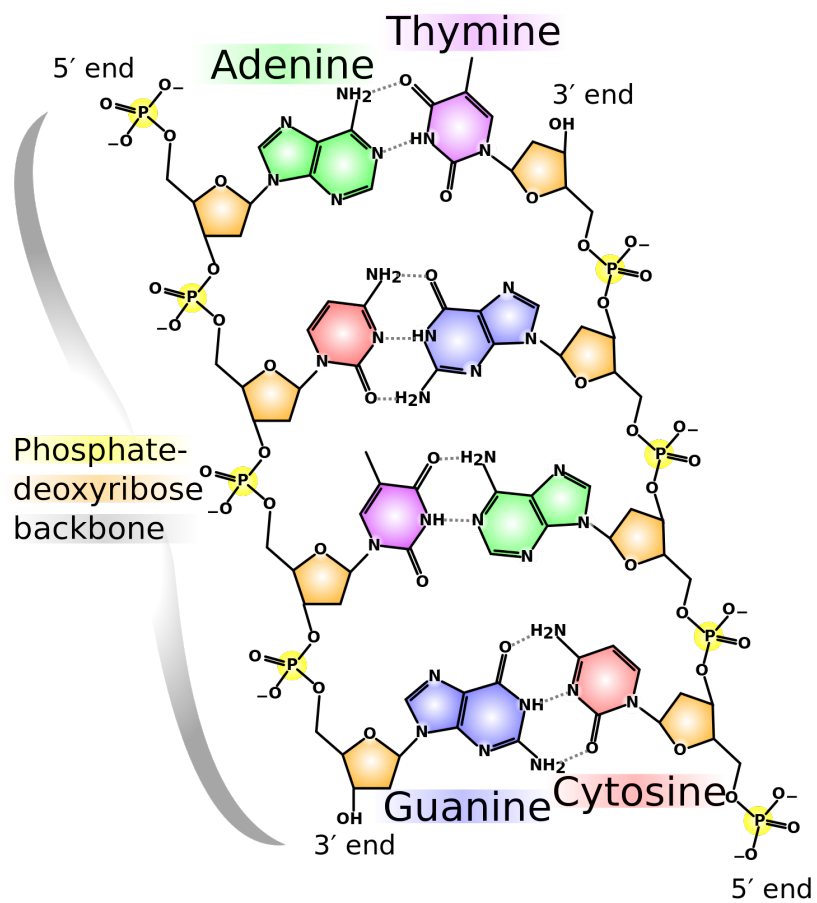
Figure 1.1: Chemical structure of DNA. Hydrogen bonds between DNA base pairs are shown with dotted lines.

result of normal cellular processes. When DNA is damaged, it can lead to various problems, including mutations, cell death, and even cancer.

There are many different types of DNA damage, including single-strand and double-strand breaks, cross-linking of DNA strands, and modifications to the sugar-phosphate backbone of DNA. These damages can be caused by a wide range of factors, including ultraviolet (UV) radiation from the sun, chemical carcinogens, and even normal cellular processes such as replication and repair [13–16]. Single-strand breaks occur when one of the strands of the double-helix structure of DNA is broken. These types of lesions are relatively common and can be caused by a variety of factors, including exposure to UV radiation, certain chemicals, and even normal cellular processes. Double-strand breaks are more severe, as they involve the breaking of both strands of the DNA double-helix [15, 16]. These types of lesions are more difficult to repair, and if left unrepaired, they can lead to mutations, cell death, and even cancer. Cross-linking of DNA strands occurs when two DNA strands become physically linked together. This can happen when certain chemicals, such as formaldehyde, react with DNA, causing the strands to become cross-linked. This can interfere with the normal function of DNA, including replication and transcription. Modifications to the sugar-phosphate backbone of DNA can also cause lesions. These modifications can include the addition of methyl groups to the sugar molecules, which can interfere with the normal function of DNA.

DNA base pair mismatch damage occurs when there is a mismatch between the nucleotide bases that make up the DNA double helix. This can happen during DNA replication, when the DNA polymerase enzyme is copying the genetic information from one strand of DNA to another. If the enzyme encounters a base that does not match its complementary base on the template strand, it can insert the wrong nucleotide, resulting in a mismatch [17, 18]. Mismatches in the DNA sequence can have serious consequences for the cell, as they can cause problems with the proper functioning of proteins. For example, a single base pair mismatch can lead to the creation of a protein with an altered amino acid sequence, which can affect its structure and function. In some cases, this can lead to the development of diseases such as cancer. To repair base pair mismatch damage, the cell has a number of mechanisms in place. One such mechanism is called mismatch repair, which involves specialized enzymes that recognize and remove the incorrect nucleotide from the DNA strand. The DNA polymerase enzyme can then correctly insert the correct nucleotide in its place. The base pair mismatch damage is a potential problem for cells, as it can lead to genetic mutations and other problems.

Each type of DNA lesion can have different effects on the cell and it is important for the body to have mechanisms in place to repair DNA damage in order to maintain the integrity of the genome. DNA lesions in human cells are repaired by a number of mechanisms, depending on the type and severity of the damage. Some of the main mechanisms for repairing DNA lesions in human cells include:
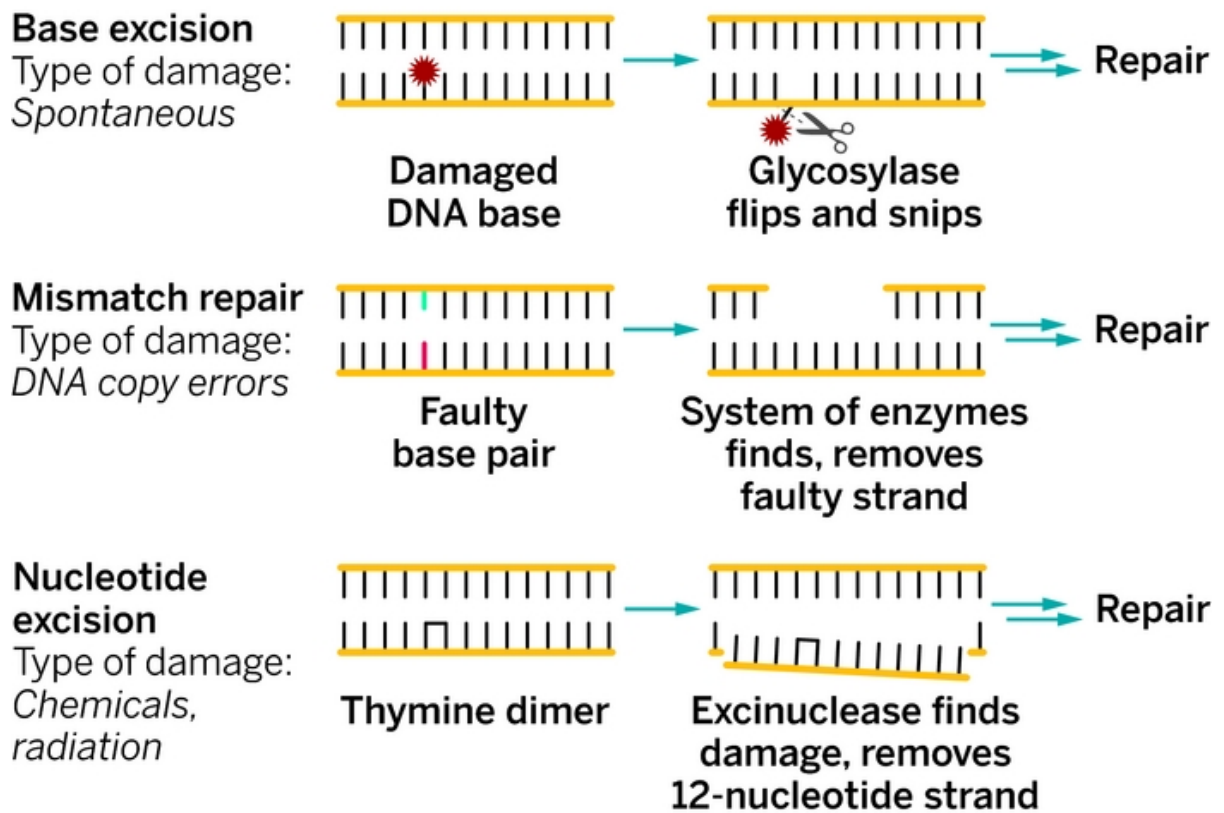
Figure 1.2: A representative image showing few types of DNA damages commonly occuring in human cells.

1. **Base excision repair**: This is a mechanism that is used to repair small, non-helix-distorting lesions in DNA. It involves the removal of the damaged nucleotide by a specific enzyme called a DNA glycosylase, followed by the insertion of a new, correct nucleotide in its place [19].

2. **Nucleotide excision repair**: This mechanism is used to repair larger lesions in DNA that are more helix-distorting. It involves the removal of a larger segment of DNA, including the damaged nucleotide, by a specific enzyme complex. This is followed by the synthesis of a new, undamaged strand of DNA using the intact strand as a template [20].

3. **Mismatch repair**: This mechanism is used to repair mismatches in the DNA sequence that occur during DNA replication. It involves the recognition of the mismatch by specific enzymes, followed by the removal of the incorrect nucleotide and the insertion of the correct one [17, 18].

4. **Double-strand break repair**: This mechanism is used to repair double-strand breaks in DNA. There are two main pathways for repairing double-strand breaks: homologous recombination and non-homologous end joining. In homologous recombination, the broken DNA strands are repaired using a homologous (identical) template. In non-homologous end joining, the broken ends are directly joined together without the use of a template [21].

These mechanisms help to maintain the integrity of the genome and prevent mutations and other harmful diseases.

1.2   Protein Kinases

Protein kinases are a family of enzymes that transfer the γ-phosphate group from adenosine triphosphate (ATP) to a protein (specifically to the hydroxyl group of serine, threonine or tyrosine residues of a protein), thereby modifying the protein's function. This process, known as phosphorylation (Figure 1.3), is a critical regulatory mechanism in cells and is involved in many cellular processes, including signal transduction, metabolism, and the cell cycle. Owing to their ability to phosphorylate, protein kinases are also called molecular switches since they act like switches, turning proteins on (also known as activated or upregulated state) or off (also known as deactivated or downregulated state) depending on the needs of the cell [22–24].

Protein kinases are divided into two main classes: serine/threonine kinases and tyrosine kinases. Serine/threonine kinases phosphorylate proteins on the amino acids serine and threonine, while tyrosine kinases phosphorylate proteins on the amino acid tyrosine. Serine/threonine kinases typically consist of two subunits: a catalytic subunit that carries out the phosphorylation reaction, and a regulatory subunit that helps control the activity of the enzyme. There are many different serine/threonine kinases, and they are typically grouped into different families based on their structure and function. Some examples of
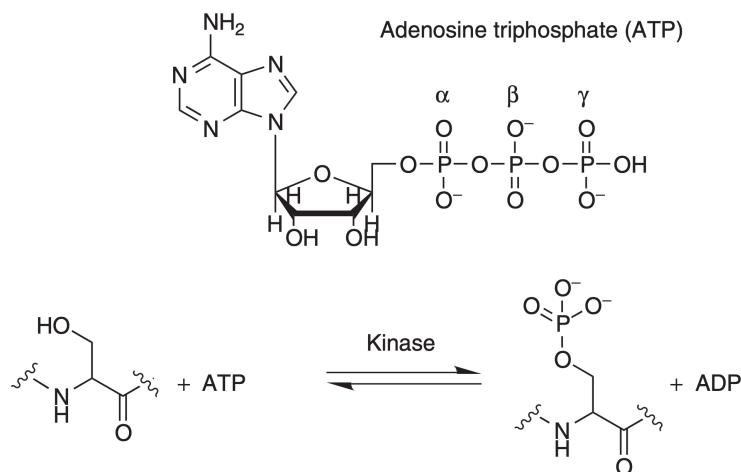
Figure 1.3: The phosphotransfer reaction catalyzed by a protein kinase.

serine/threonine kinase families include the AGC kinase family, the CMGC kinase family, the CAMK kinase family, and the STE kinase family. Each of these families has its own unique set of members, and each member plays a specific role in the cell [25–27].

Tyrosine kinases play a key role in signaling pathways that control a variety of cellular processes. For example, when a growth factor binds to a receptor on the surface of a cell, it activates the receptor's tyrosine kinase activity. This leads to the phosphorylation of specific tyrosine residues on the receptor, which in turn triggers a cascade of downstream signaling events that ultimately result in the cell responding to the growth factor. Tyrosine kinases are also involved in the development of cancer. Many oncogenes, which are genes that can promote cancer, encode for tyrosine kinases. When these oncogenes are activated, they can cause excessive tyrosine kinase activity and lead to the uncontrolled growth and proliferation of cancer cells. As a result, tyrosine kinases are the target of many cancer drugs [28–30].

Non-receptor tyrosine kinases are a subtype of tyrosine kinases that do not function as part of a receptor on the cell surface. Instead, they are found inside the cell, where they phosphorylate tyrosine residues on other proteins to regulate their activity. Unlike receptor tyrosine kinases, which are activated by signaling molecules that bind to the receptor on the cell surface, non-receptor tyrosine kinases are typically activated by other cellular events such as mechanical stress or DNA damage. For example, when cells are stretched or subjected to mechanical stress, non-receptor tyrosine kinases are activated and phosphorylate specific proteins to regulate the cell's response to the stress.

Non-receptor tyrosine kinases are also involved in the development of cancer. Many oncogenes, which are genes that can promote cancer, encode for non-receptor tyrosine kinases. When these oncogenes are activated, they can cause excessive tyrosine kinase activity and lead to the uncontrolled growth

and proliferation of cancer cells. As a result, non-receptor tyrosine kinases are the target of many cancer drugs as well.

The Src family of non-receptor tyrosine kinases includes several closely related proteins including Src, Yes, Fyn, Lyn, Lck, Blk, Hck, Fgr, and Yrk. These proteins share similar structures, functions and regulatory mechanisms. The numerous signal transduction pathways and biological processes depend on these kinases. In addition to signaling downstream DNA synthesis, MAPK activation, cytoskeletal rearrangements, and cell movement, the members interact with a wide variety of transmembrane receptors. Tyrosine residues are phosphorylated in order to control the activity of these kinases. Tyr416 is a regulatory site whose phosphorylation is linked to an increase (upregulation) in kinase activity [31–33]. Another site is Tyr527, whose phosphorylation will suppress (downregulate) the kinase activity [34, 35].

1.3   Structure of Src family Kinases

All of the 9 member proteins of the Src family (Src, Yes, Fyn, Lyn, Hck, Lck, Blk, Fgr, and Yrk) share a similar overall structure, which includes the myristoylated N-terminus domain, SH4 domain, Unique domain, SH3, SH2, linker domain a central kinase domain (also known as SH1 domain), and a C-terminal tail that is involved in protein-protein interactions [36–41].

The myristoylation of the N-terminus refers to the attachment of a myristoyl group, a type of fatty acid, to a specific amino acid at the N-terminus of the protein. This post-translational modification is important for the proper localization and function of Src family kinases within cells. It is located at the beginning of the protein and is composed of a specific sequence of amino acids. Myristoylation helps anchor the protein to the membrane, allowing it to interact with other proteins and perform its functions within the cell.

Src homology 2 (SH2) domain is composed of about 100 amino acid residues. It's role is to bind to phosphorylated tyrosine residues on other proteins, which allows SFKs to interact with and regulate various signaling pathways in response to various stimuli. This leads to a variety of cellular responses, including cell growth, survival, and migration. Similar to SH2, the SH3 domain is composed of about 60 amino acid residues and binds to proline-rich regions on other proteins, thereby playing critical role in regulating various signaling pathways [42, 43].

The linker domain connects SH3 and SH2 domains. It is composed of a variable number of amino acid residues, and its structure and composition can vary depending on the specific SFK. In addition to its role in regulating SFK activity, the linker domain is important for the localization of SFKs within the cell. It has been shown to interact with various proteins and organelles, including the plasma membrane, the cytoskeleton, and the nucleus. This allows SFKs to be properly positioned within the cell to carry out their various functions.
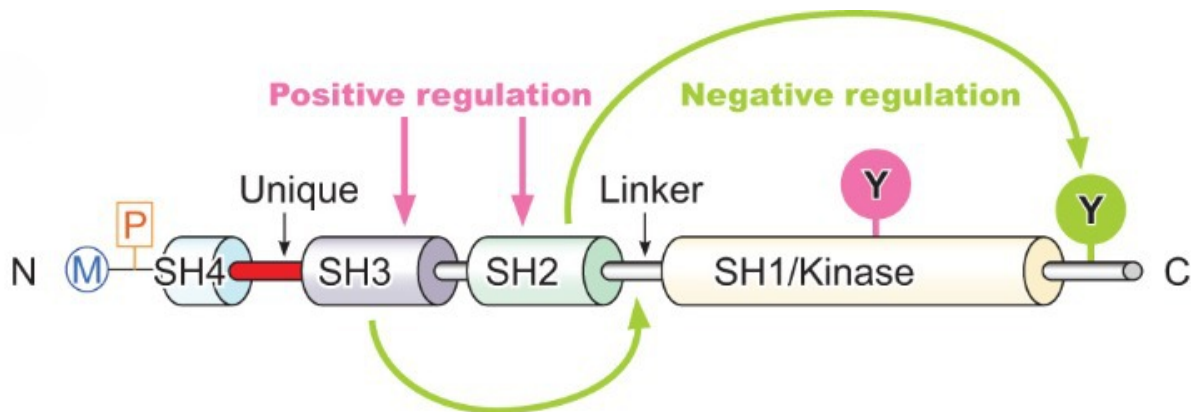
Figure 1.4: A representative diagram showing various structure domains of the Src family Kinases.

The C-terminal tail is located at the end of the SFK. It is composed of a variable number of amino acid residues, and its structure and composition can vary depending on the specific SFK. The C-terminal tail is also responsible for the localization of SFKs within the cell. It also interacts with other proteins inside the cell and makes sure the SFK is properly positioned within the cell to carry out various functions [44].

The central kinase domain performs the catalytic activity of the SFK. It is composed of about 250 amino acid residues, and is responsible for the ability of SFKs to phosphorylate tyrosine residues on other proteins. It contains the active site of the SFK, where the phosphorylation of tyrosine residues takes place. Mutations in the central kinase domain can affect the activity of SFKs and lead to impaired function. It also contains several regulatory sites, including the ATP binding site and the substrate binding site, which allow SFKs to be activated and inhibited in response to various stimuli. We will talk in detail about the kinase domain in the next section.

A schematic representation of the Kinase domain sequence is shown in Figure 1.4. The residue numbering of kinase domains in c-Src kinase is shown in Table 1.1. Crystal structures of the active and inactive forms of c-Src kinase have been shown in Figure 1.5.

Table 1.1: The residue numbering of the domains in c-Src kinase.

Region	SH3	SH2	linker	N lobe	C lobe	C-terminus
Residue no.	$84-153$	$154-245$	$246-259$	$260-341$	$342-521$	$522-533$

1.4 Structural features of the Catalytic Kinase domain of SFKs

The central kinase domain is the most important domain of SFKs. This is responsible for the kinase's enzymatic activity and consists of various regulatory sites, allowing SFKs to be upregulated or
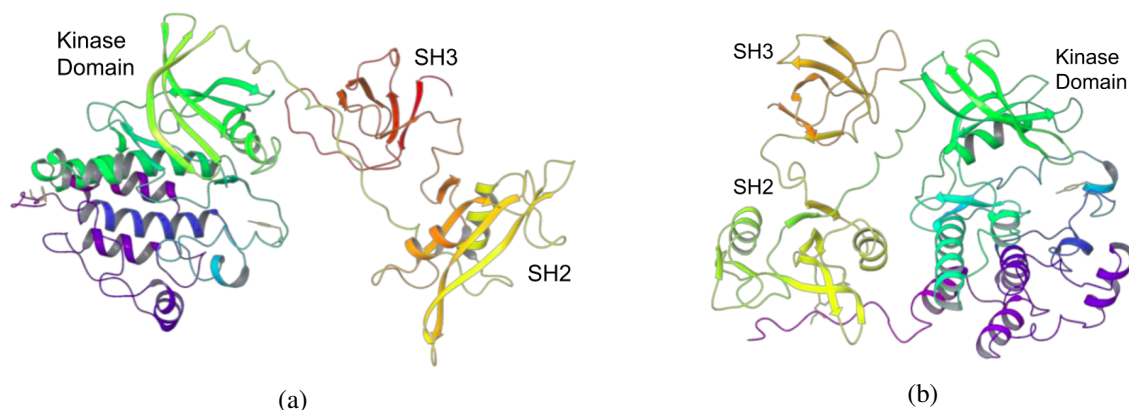
Figure 1.5: Crystal structure of (a) active and (b) inactive form of c-Src kinase along with important domains. Figure adapted from [45].

downregulated in response to various stimuli. The various important structural features of this domain are listed below:

1. **N-lobe**: The N-Lobe is the larger lobe in the catalytic kinase domain. It typically consists of a five-stranded beta-sheet, coupled to a C-helix. It also contains two highly conserved sequence motifs within the Beta-sheet strands, namely, the glycine rich loop and P-loop. Due to the high flexibility of γ-phosphate of ATP, the glycine-rich loop is important to keep the γ-phosphate of ATP in the right position for catalysis.

2. **C-lobe**: C-lobe is the smaller lobe and contains 5 helices (D, E, F, G and H) and one beta-sheet. This helical subdomain is extremely stable and forms the main core of the kinase. The C-lobe also contains the binding site for protein substrates for kinase enzymatic activity. D, E, F, and H helices are well-shielded and buried deep inside the C-lobe. The G-helix is more solvent exposed. The beta-sheet in C-lobe contains the catalytic domains needed for kinase enzymatic activity.

3. **Activation loop**: Activation loop is part of the kinase's catalytic domain. The activation loop helps to position the substrate in the active site and bring the enzyme's catalytic residues into close proximity, facilitating the transfer of the phosphate group.

4. **DFG Motif**: Aspartate (D), Phenylalanine (F), and Glycine (G) amino acids in the activation loop form the DFG motif. It is necessary for the DFG motif to be in "in" conformation, towards the ATP binding site, for the kinase to be in active conformation. The Phe residue in "in" state forms the base which stabilises ATP by forming hydrogen bonds with it.

5. **LYS-GLU Salt Bridge**: Lys295 and Glu310 in N-lobe form an evolutionary salt bridge in the active state and it is conserved across the kinase family. This bridge is broken when the kinase

switches to the inactive state. The opening and closing of this bridge is coupled with water molecules penetrating inside the binding pocket and occupying the ATP binding site.

6. **ATP binding site**: The ATP binding site is the conserved region in between the N and C-lobes. This is the region where ATP binding happens. This region is correctly formed when the kinase is in active conformation. The residues in the ATP binding site stabilise the ATP molecule by forming hydrogen bonds directly and water mediated with the ATP molecule.

7. **Substrate binding site**: Substrate binding site is part of the C-lobe, where the substrate docks and the enzymatic activity happens. The substrate binding site can only form correctly when the kinase is in active conformation.

8. **α-C helix**: This helix is part of the N-lobe and rotates outwards when the kinase goes from active to inactive state.

9. **Catalytic Spline**: The C-spline consists of hydrophobic residues from both the lobes. Binding of ATP is necessary to complete the C-spline since the adenine ring of ATP is a part of this.

10. **Regulatory Spline**: The R-spline consists of 4 non-consecutive hydrophobic residues, 2 each from both the lobes.

The kinase must be in active conformation in order to perform its enzymatic activity. This requires various motifs in the kinase catalytic domain to be oriented in a correct relative geometry to attract ATP and substrate to their respective binding sites. The most important of these is the assembly of the two hydrophobic spines, R-spline and C-spline, in the enzyme's core, which activates the kinase. The C-helix also changes orientation, moving from a position away from the cleft in the down-regulated form to a position closer to the cleft in the active form. This brings a key residue, Glu310, into position to form a catalytically important salt bridge with Lys295. The Activation loop needs to be open and rotated outwards, so as to form enough space for the ATP to bind. In the down-regulated form of the enzyme, the A-loop is folded in a two-helical structure that blocks access to the catalytic site. In the active form, the A-loop extends into a free loop structure, exposing the catalytic site for substrate phosphorylation. The relative orientation of the N-lobe and C-lobe also needs to happen correctly. The distance between the two lobes also increases in the active state, resulting in a more open catalytic cleft.

1.5 Kinase Inhibitors

Kinases, if not regulated properly, can behave abnormally and lead to overexpression, causing various diseases including cancer. Kinase inhibitors are drugs that target specific kinases in order to prevent cancer from spreading by "locking" and stabilising the respective kinase in a particular conformation,
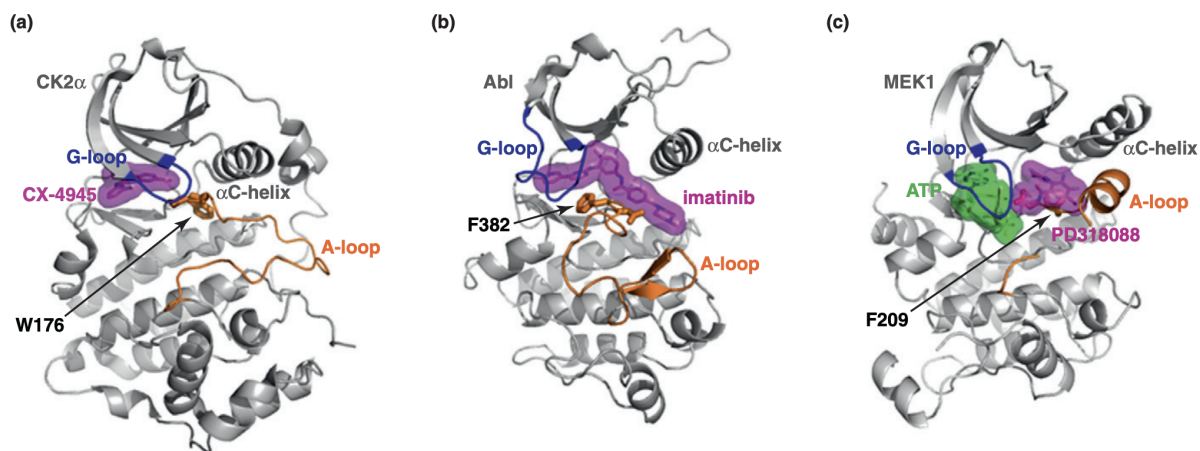
Figure 1.6: Types of kinase inhibitors. The kinase is shown in grey color and inhibitors in their sticks and surface representation are shown in magenta color. Figure adapted from [51].

thereby restricting it from performing its catalytic activity. There are several ways to classify kinase inhibitors based on their mechanism of action [46–50]. One classification divides Src family kinases' inhibitors into three categories: type I, type II, and type III.

Type I inhibitors bind to the active form of a kinase in the ATP binding site. These inhibitors are therefore similar in size and share similar structural characteristics with the ATP molecule. Similar to ATP, type I inhibitors also form covalent contacts with kinase residues in the ATP binding pocket and with the water molecules trapped inside. They thus prevent the ATP molecule from binding with the kinase even when the kinase is in active-like conformation [46–48]. Since the ATP binding pocket is conserved across the kinase family, these drugs are generally not specific to a particular kinase, but rather effective for the family of kinases. This is a drawback since the behaviour of other kinases would also be impacted upon the usage of such drugs. Some examples of FDA approved type I drugs are:

1. Imatinib (Gleevec) - A medication used for the treatment of chronic myeloid leukemia and gastrointestinal stromal tumors. It works by targeting the BCR-ABL tyrosine kinase, a protein involved in the growth and spread of these types of cancer.

2. Sunitinib (Sutent) - A medication used for the treatment of several types of solid tumors, including kidney, pancreatic, and gastrointestinal cancers. It targets a range of receptor tyrosine kinases, including VEGFR, PDGFR, and KIT.

3. Erlotinib (Tarceva) - A medication used for the treatment of non-small cell lung cancer and pancreatic cancer. It works by targeting the epidermal growth factor receptor (EGFR), which is often overexpressed in these types of cancer.

4. Lapatinib (Tykerb) - A medication used for the treatment of breast cancer. It works by targeting the human epidermal growth factor receptor 2 (HER2) and the epidermal growth factor receptor (EGFR).

5. Dasatinib (Sprycel) - A medication used for the treatment of chronic myeloid leukemia and acute lymphoblastic leukemia. It works by targeting the BCR-ABL tyrosine kinase, as well as a number of other receptor tyrosine kinases.

Type II kinase inhibitors are small molecule kinase inhibitors that bind to inactive or "off" kinase conformation. They are typically characterized by their ability to bind to the Asp-Phe-Gly (DFG)-OUT conformation of a kinase. They bind to the ATP binding conserved site and extend to the allosteric site ahead. This allosteric is formed only when the kinase is in DFG-out inactive conformation and when the Activation loop is in an inactive-like conformation. Therefore, Type II inhibitors often have slower dissociation rates, resulting in significantly increased affinities over type I inhibitors. Some examples of FDA approved type II drugs are:

1. Nilotinib (Tasigna) - A medication used for the treatment of chronic myeloid leukemia. It works by targeting the BCR-ABL tyrosine kinase, which is involved in the development and progression of this type of cancer.

2. Ponatinib (Iclusig) - A medication used for the treatment of chronic myeloid leukemia and acute lymphoblastic leukemia. It works by targeting the BCR-ABL tyrosine kinase, as well as a number of other receptor tyrosine kinases.

Type III kinase inhibitors, also known as allosteric inhibitors, are a class of small molecule protein kinase inhibitors that do not compete with ATP for binding to the active site of the kinase. Instead, they bind to a different site on the kinase, often referred to as the allosteric site, which can affect the activity of the enzyme. Allosteric inhibitors can alter the conformation of the kinase, making it more or less likely to bind to ATP or other substrates. They can also interfere with the ability of the kinase to transfer the phosphate from ATP to the substrate. Some examples of FDA approved type III drugs are:

1. Palbociclib (Ibrance) - A medication used for the treatment of breast cancer. It works by targeting cyclin-dependent kinases 4 and 6 (CDK4/6), which are involved in the regulation of the cell cycle.

2. AZD9291 (Osimertinib) - A medication used for the treatment of non-small cell lung cancer. It works by targeting the epidermal growth factor receptor (EGFR), which is often overexpressed in this type of cancer.

3. Alectinib (Alecensa) - A medication used for the treatment of non-small cell lung cancer. It works by targeting the anaplastic lymphoma kinase (ALK) receptor tyrosine kinase, which is often overexpressed in this type of cancer.

1.6   Research Focus

1.6.1   Sequence Specificity, Energetics, and Mechanism of Mismatch Recognition by DNA Damage Sensing Protein Rad4/XPC

At the beginning of this chapter, we discussed about various DNA lesions that can harm the cells, and different damage identification and repair mechanisms the cells deploy to cope with these DNA lesions.

In the present work, we have studied how the DNA damage sensing enzyme, Rad4, recognises base-pair mismatches in yeast cells and the mechanism and energetics surrounding this process. The three key events that occur during this recognition process undertaken by Rad4 are as follows: (a) the association of Rad4 with the damaged base-pair site, (b) the flipping of the damaged bases and their partner bases and (c) the insertion of β-hairpin of Rad4 into the damaged site. The exact mechanism, ordering and energetics of these events remain elusive. We have employed the umbrella sampling technique, along with molecular dynamics simulations to understand the association of RAD4 enzyme with the damaged DNA site in detail. We have studied 3 different base-pair mismatches in this work, namely, TTT/TTT, TAT/TAT, CCC/CCC. We studied and observed how the energetics surrounding these 3 events also varies with the exact DNA base-pair mismatch sequence.

1.6.2   Minimum Energy Path Analysis of Drug-induced Conformational Transitions of c-Src Kinase

Despite the wealth of structural knowledge about the active and inactive states of protein kinases, much remains to be understood about the precise mechanism of conformational transitions between the active and inactive states. The effect of various inhibitors on protein kinase activity and associated conformational transitions is yet to be fully understood. Although the rotation of α-C helix and the opening/closing of activation loop appear to important events in kinase activation/inactivation, there are still ambiguities in the order in which these events occur (whether the folding of the A-loop precedes or succeeds the rotation of the α-C helix) during the inactivation process. While some studies have suggested that the α-C helix rotation occurs first and is followed by the folding of the A-loop, others have suggested the opposite order of events. Moreover, there is a lack of structural information about the intermediates along the path of the active-inactive conformational transition. Another area of uncertainty is the function of the Src regulatory domains and how they impact the transition of the kinase domain. Most studies done so far to understand activation/deactivation mechanism and energetics of kinases have focused primarily on the kinase catalytic domain and its conformational changes. Only a few studies have also taken the impact of drugs on these conformational mechanisms into account.

In the present study, we have answered some of these questions. We have employed the use of Minimum Energy Paths (MEPs) to study the conformational transition between the drug-bound active

and drug-free inactive states. The crucial role of the drug and water molecules at the binding site in the mechanism of this conformational transition and drug binding/unbinding process is also studied. We believe that the results on the mechanism and energetics of drug-induced conformational transition in c-Src kinase will be useful for rational design of novel drugs for therapeutics.

Chapter 2

Computational Methods

Contents

2.1   Introduction

Computer simulations and molecular modelling have emerged as valuable tools to study complex biological systems at the molecular level. They enable us to better understand the mechanics, kinetics and energetics that govern molecular systems. Computer simulations are typically based on mathematical models that represent the behavior of a system. These models can be based on existing theories or on data collected from observations of the real-world system. Once a model has been developed, it can be used to generate virtual experiments or scenarios that can be studied in detail. This allows researchers to study the behavior of a system under a wide range of conditions, and to test how the system will respond to different inputs or perturbations.

A key advantage of computer simulations is that they can be performed at multiple times, allowing researchers to study the behavior of a system over a wide range of possible scenarios. This can provide a more comprehensive view of the system than what could be obtained from a single experiment in the real world. In addition, computer simulations can be run much faster than real-world experiments, allowing researchers to study the behavior of a system on much shorter timescales.
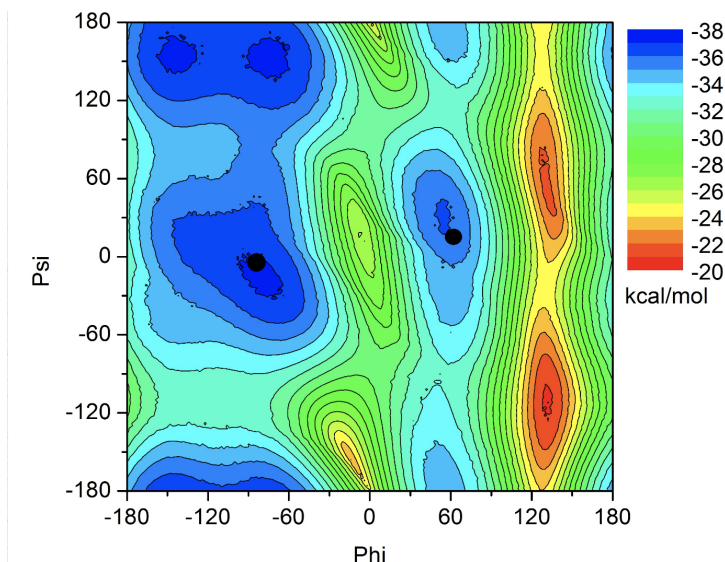
Figure 2.1: Potential energy surface of alanine dipeptide, in terms of its dihedral angles. Figure adapted from `http://ambermd.org/tutorials/advanced/tutorial5_amber11/section2.htm`

2.2 Potential Energy Surfaces

The potential energy surface describes how the potential energy of a system changes as the positions of its constituent atoms (described by the configuration of the system ($\{\mathbf{r}\}$)) are modified. The potential energy of any system is therefore a function of 3N dimensions, where N is the number of atoms in the system, considering x, y and z dimensions for each constituent atom. The derivative of potential energy $U(\{\mathbf{r}\})$ of a system with respect to the atomic coordinates calculated at a given configuration (defined by its position on the 3N dimensional potential energy surface) gives us the forces on atoms, which governs the dynamics of the system [52]. It is not feasible to plot and visualise potential energy surfaces for large complex systems, as it spreads in 3N dimensions. However, potential energy surfaces for small systems like alanine dipeptide have been studied extensively and used as model systems for developing scientific research. The potential energy surface for alanine dipeptide is a function of its two torsion dihedral angles, Φ and Ψ, as can be seen in Figure 2.1.

The potential energy surface is rugged, full of maximum and minimum, for large complex systems. Studying the mechanisms of conformational transitions between different stable states, represented by energy basins on the potential energy surface, is therefore nontrivial. Developing algorithms for finding paths that the system might take for conformational transitions on the energy landscape is an ongoing area of research. We seek to understand and use one such algorithm in our current work.

2.3 Force Fields

Force fields are mathematical models that are used to describe the interactions between atoms in a molecular system. They are used in computer simulations to predict the behavior of molecules, including their structures, energies, and dynamics. They are essential for understanding and predicting the properties of molecular systems by helping us understand the behaviour of such systems using computer simulations. Force fields are typically divided into two categories: empirical force fields and ab initio force fields. Empirical force fields are based on fitting parameters to experimental data and are generally more efficient and easier to use than ab initio force fields. They are commonly used for studying large molecular systems and for performing simulations over long time scales. However, they are limited by their reliance on experimental data and may not accurately describe systems that are significantly different from those used to fit the parameters [52, 53].

Ab initio force fields, on the other hand, are based on first principles, such as quantum mechanics. They provide a more accurate description of the interactions between atoms, but are generally more computationally intensive and less efficient than empirical force fields. They are typically used for studying small molecular systems or for studying systems for which experimental data is not available. There are many different force fields available, each with its own set of parameters and assumptions. Choosing the appropriate force field for a particular simulation depends on the nature of the system being studied and the desired level of accuracy [54–58].

Most standard empirical force fields can be represented using the following general equation:

$$
\begin{aligned}
U(\mathbf{r}) = &\frac{1}{2} \sum_{\text{bonds}} k_b (b - b_0)^2 \\
&+ \frac{1}{2} \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \\
&+ \frac{1}{2} \sum_{\text{torsions}} k_\phi (1 + \cos{(n\phi - \delta)}) \\
&+ \sum_{\text{coulomb,ij}} \frac{1}{4\pi\varepsilon} \frac{q_i q_j}{r_{ij}} + \sum_{\text{vdW,ij}} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]
\end{aligned} \tag{2.1}
$$

The first and second terms in Equation (2.1) describe bond-stretching and angle-bending, which are treated harmonically to keep bonds and angles close to their equilibrium values [55, 56, 59–62]. The bond length and angle are represented by b_0 and θ_0, respectively, and the force constants for these terms are K_b and K_θ. The third term represents dihedral angles and includes the dihedral force constant k_ϕ, the multiplicity of the function n, the dihedral angle ϕ, and the phase shift δ [52, 63].

The last term in Equation (2.1) describes the nonbonded interactions between pairs of atoms (i,j), including van der Waals and electrostatic interactions. The van der Waals interaction is described by a Lennard-Jones potential, which consists of a short-range repulsion and a long-range attraction. The distance between the interacting atoms is represented by r_{ij}, and the LJ well depth and distance at which the minimum LJ energy occurs are represented by ϵ_{ij} and σ_{ij}, respectively. These LJ parameters are calculated using the Lorentz-Berthelodt combination rules, which involve taking the geometric mean of ϵ_{ii} and ϵ_{jj} for the individual atom types and the arithmetic mean of σ_{ii} and σ_{jj} for the individual atom types. The electrostatic interaction between the partial atomic charges q_i and q_j is calculated using the Coulombic potential with a dielectric constant of ϵ [64].

2.4 Energy Minimization

Energy minimization on a potential energy surface refers to the process of finding the configuration of a system that corresponds to the minimum potential energy. This can be achieved by using optimization algorithms such as steepest descent or conjugate gradient. In these algorithms, the positions and orientations of the particles are adjusted iteratively until the potential energy reaches a local minimum.

Any configuration that is an energy minimum on the potential energy surface should satisfy the following conditions :

$$\frac{\partial U}{\partial x_i} = 0 \qquad \text{and} \qquad \frac{\partial^2 U}{\partial x_i^2} > 0$$

where U is the potential energy function of the system.

Energy minimisation techniques are used to find stable states for the system on the potential energy landscape. These represent more stable energy basins, where the system is more likely to be found.

2.4.1 Steepest Descent Algorithm

Steepest Descent energy minimisation is a type of gradient descent algorithm and it adjusts the coordinates of the particles in the direction of the negative gradient of the potential energy surface.

The gradient of the potential energy surface gives the direction of steepest descent, and the particles are moved in this direction until the potential energy reaches a local minimum [65, 66]. The algorithm works by iteratively updating the coordinates of the particles according to the following formula:

$$x_{n+1} = x_n - \alpha_n \nabla U(x_n) \qquad (2.2)$$

where x is the coordinate of the particle, α is a step size parameter, and $\nabla U(x)$ is the gradient of the potential energy U with respect to the coordinate x.

The steepest descent algorithm is simple to implement and easy to parallelize, but it can be slow to converge and may get stuck in local minima. If the step size parameter alpha is too small, the algorithm will converge slowly because the particles will be moved only a small distance at each iteration. If the step size parameter is too large, the algorithm may oscillate or diverge because the particles will be moved too far in each iteration. It is therefore important to choose an appropriate step size parameter to ensure α good convergence.

2.4.2 Conjugate Gradient Algorithm

The conjugate gradient algorithm is another gradient descent algorithm for minimizing the potential energy of a system. It converges faster than the steepest descent algorithm. This is because the conjugate gradient algorithm uses a search direction that is conjugate to the previous search directions, which helps to avoid oscillations and ensure that the algorithm converges quickly [67].

The conjugate gradient algorithm works by iteratively updating the coordinates of the particles according to the following formula:

$$x_{n+1} = x_n - \alpha d_n \tag{2.3}$$

where x is the coordinate of the particle, α is a step size parameter, and d is the search direction. The search direction is updated at each iteration according to the following formula:

$$d_n = \nabla U(x_n) + \beta d_{n-1} \tag{2.4}$$

where $\nabla U(x)$ is the gradient of the potential energy with respect to the coordinate x, and β is a coefficient that ensures that the search direction is conjugate to the previous search direction.

The conjugate gradient algorithm is easy to implement and parallelize, and it has good convergence properties. However, it can be sensitive to the choice of step size parameter alpha, and it may fail to converge if the potential energy surface has multiple local minima.

2.5 Molecular Dynamics (MD) Theory

Molecular dynamics (MD) simulations are computer-based modelling techniques that use classical mechanics to simulate the motion of a system of particles. These simulations are used to study the behavior of molecules at the atomic level, including their structural, dynamical, and thermodynamic properties. In an MD simulation, the positions and velocities of the particles in a system are calculated at each time step using Newton's laws of motion. The interactions between the particles are typically described using force fields of our choice. As studied above as well, these force fields can either be empirical, based on experimental data or theoretical, based on quantum mechanical calculations.

The simulations are performed using algorithms that solve the equations of motion for the system of particles. The most common algorithm is the Verlet algorithm, which is based on the second order Taylor expansion of the equations of motion. One of the main advantages of MD simulations is that they can be used to study systems that are too large or too complex to be studied experimentally or theoretically. They can also be used to study systems that are difficult to study experimentally because they are too small, too fast, or too reactive.

The microscopic state of a system, in terms of phase-space, can be characterized by the positions and momenta of the atoms that make up the system. The total energy of the system, known as the Hamiltonian and denoted by H, is composed of the kinetic energy (the energy associated with the motion of the atoms) and the potential energy (the energy associated with the interactions between the atoms). The Hamiltonian can be written as the sum of these two energies for a system consisting of N atoms using the following equation :

$$H(\mathbf{r}, \mathbf{p}) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}^N) \tag{2.5}$$

where

The coordinates of the atoms in the system are represented by a set of coordinates, $\mathbf{r} = (r_1, r_2, ..., r_N)$, and the momenta of the atoms are represented by a set of momenta, $\mathbf{p} = (p_1, p_2, ..., p_N)$. The mass of the ith atom is represented by m_i, and the potential energy of the system is given by $U(\mathbf{r}^N)$. The force acting on each atom is calculated by taking the negative gradient of the potential energy with respect to the atom's position, resulting in the equation $F_i = -\Delta_i U(\mathbf{r}^N)$. The time evolution of the system is governed by Hamilton's equations of motion:

$$\dot{\mathbf{p}}_i = -\frac{\partial H}{\partial \mathbf{r}_i} = -\frac{\partial U}{\partial \mathbf{r}_i} = \mathbf{F}_i \tag{2.6}$$

$$\dot{\mathbf{r}}_i = \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} \tag{2.7}$$

In the above expression, r_i, p_i, $\dot{r}_i$ and $\dot{p}_i$ represent the position, momentum and their time-derivatives, respectively for the ith atom in the system.

The equations of motion are integrated to obtain the dynamics of the system. If $\mathbf{r}_i(t)$ and $\mathbf{p}_i(t)$ represent the position and momentum of ith particle at time t, then its position and momentum after a timestep Δt are given by

To study the dynamics of the system, we integrate the above equations of motion. If $\mathbf{r}_i(t)$ and $\mathbf{p}_i(t)$ represent the position and momentum of the ith particle at time t, then its position and momentum at time t+Δt can be expressed as :

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{1}{2m_i}\mathbf{F}_i(t)(\Delta t)^2 \tag{2.8}$$

$$\mathbf{p}_i(t + \Delta t) = \mathbf{p}_i(t) + \frac{1}{2m_i}[\mathbf{F}_i(t) + \mathbf{F}_i(t + \Delta t)]\Delta t \tag{2.9}$$

By repeatedly applying this procedure for a large number of time steps, we can determine the trajectory of the system, or the path that the system follows over time. This can give us insight into the structural, dynamical, and thermodynamic properties of the system [68–71].

2.5.1 The Velocity Verlet Integrator

The velocity Verlet equation is a mathematical equation that is used in molecular dynamics simulations to calculate the movement of atoms and molecules. It is based on the principles of classical mechanics, is a second-order integrator, and it is used to predict the position and velocity of a particle at each step of the simulation [72].

The update equation is given by:

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{1}{2}\left[\mathbf{a}(t) + \mathbf{a}(t + \Delta t)\right]\Delta t \tag{2.10}$$

where $\mathbf{v}$ is the velocity of the particle, $\mathbf{a}$ is the acceleration of the particle, and Δt is the time step.

The acceleration of the particle at time $t+\Delta t$ can be calculated from the potential energy of the system using the equation:

$$\mathbf{a}(t + \Delta t) = -\frac{\nabla U(\mathbf{r}(t + \Delta t))}{m} \tag{2.11}$$

where $U(\mathbf{r})$ is the potential energy of the system as a function of the positions of the particles, $\mathbf{r}$ is the position of the particle, and m is the mass of the particle.

The position of the particle at time $t+\Delta t$ can then be updated using the velocity at time $t+\Delta t$:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t + \Delta t)\Delta t \tag{2.12}$$

This completes one time step of the velocity Verlet algorithm. The algorithm can then be repeated for a specified number of time steps to simulate the dynamics of the system.

2.6 The Umbrella Sampling Method

The umbrella sampling method is a computer simulation technique that is used to calculate the free energy of a system as a function of a collective variable. A collective variable is a macroscopic quantity that represents the state of the system, such as the distance between two atoms or the dihedral angle between four atoms. The free energy of the system is a measure of the thermodynamic stability of the system and can provide insights into the behavior of the system.

In umbrella sampling, the system is artificially constrained to explore a specific range of the collective variable. The constraint is applied using a biasing potential, which is a force that acts to keep the system within a certain range of the collective variable. The biasing potential can be represented by a harmonic potential or a more complex function, depending on the needs of the simulation [73–79].

As the simulation runs, the system is allowed to move around within the constrained range of the collective variable, and the energy of the system is recorded at regular intervals. The resulting energy data is used to calculate the free energy of the system as a function of the collective variable. This method can be used to calculate the free energy of the system over a range of values of the collective variable, allowing for a detailed understanding of the thermodynamic stability of the system. It is particularly useful for studying systems that have multiple minima or multiple conformations, as it can provide information about the relative stability of different conformations [73–79].

In mathematical terms,

The free energy can be calculated by integrating the distribution function over the range of the collective variable:

$$F(x) = -kT \ln(g(x)) = -kT \ln\left(\int g(x) \, dx\right) \tag{2.13}$$

where x is the collective variable, k is the Boltzmann constant, T is the temperature, and $g(x)$ is the distribution function. The distribution function $g(x)$ can be calculated from the umbrella sampling simulations as follows:

$$g(x) = <e^{-\beta U(x)}> \tag{2.14}$$

where $U(x)$ is the potential energy of the system at a given value of the collective variable, and $\langle\rangle$ denotes an average over all configurations sampled in the simulation. The parameter β is related to the temperature and is defined as $\beta = 1/(kT)$.

2.7 Minimum Energy Path

A path in conformational space is defined as a set of conformations, represented by points. The Minimum Energy Path (referred commonly as MEP) on the Potential Energy Landscape is a set of conformations/points such that a movement in a direction perpendicular to the MEP would result in increase in the potential energy of the system. So, if a point lying on the MEP is displaced in any direction perpendicular to the path, the force due to potential energy (i.e. Perpendicular component of Natural Force) should bring the system back to its initial conformation. Therefore, a MEP is a path is such that any point on the path is at an energy minimum in all directions perpendicular to the path. Alternatively, it can be understood as the union of steepest descent paths from the saddle points to the

minima [80]. The two conformational states chosen to find MEP are generally two stable states on the potential energy surface representing the system of interest, and they are generally called the end states [81–89].

2.8 Nudged Elastic Band Method

The Nudged Elastic Band (NEB) algorithm is a method for finding the minimum energy pathway (MEP) between two stable conformational states (also known as the start and end conformations) in a system. The algorithm begins by taking a series of intermediate conformations, known as replicas, between the start and end conformations as an initial guess for the MEP. These replicas are then minimized on the Potential Energy Surface (PES). To prevent the replicas from collapsing towards the end states, they are connected to one another with artificial springs, known as nudging forces. Reparameterization, or the adjustment of the spring forces along the tangent to the MEP, is performed to maintain an equal distance between the replicas in order to efficiently sample the conformational space between the start and end conformations. However, the component of the nudging forces that is perpendicular to the MEP should not be considered in the convergence of the MEP, as these forces are artificially introduced and do not contribute to the actual convergence. Similarly, the component of the natural forces, or potential energy forces, that is parallel to the MEP is not taken into account since moving along the minimum energy path does not contribute to its convergence [81–91].

An advantage of the NEB method described above over other advanced sampling methods is it does not restrict any degrees of freedom and hence the system is free to undergo changes in its conformations naturally.

A set of initial configurations are "guessed" or interpolated, either linearly or by some guessing strategy, from the given two end states. The number of these intermediate configurations should be high, in order to effectively sample the rugged configuration space connecting the end states. These guesses, or images, are attached by Hooke springs to the neighbouring replicas. These images are then minimized over the potential energy landscape of the system following the equation:

$$\mathbf{x}^i = \mathbf{x}^i - \eta \mathbf{F}_i \qquad (2.15)$$

where $\mathbf{x}^i$ denotes the 3N-dimensional position vector of the ith image. With a standard minimization technique such as the steepest descent, the force on each image would be equivalent to the negative gradient of the potential energy $[\mathbf{F}_i = -\nabla U(\mathbf{x}_i)]$. In the nudged elastic band method, however, since replicas are connected to each other with the help of artificial springs (known as nudging), the force acting on image i is a resultant of the spring force and the negative gradient. A tangent at image i is defined, which is considered as the angle bisector of the two tangents in the original implementation:

$$\boldsymbol{\tau_i} = \frac{\mathbf{x_i} - \mathbf{x_{i-1}}}{|\mathbf{x_i} - \mathbf{x_{i-1}}|} + \frac{\mathbf{x_{i+1}} - \mathbf{x_i}}{|\mathbf{x_{i+1}} - \mathbf{x_i}|} \tag{2.16}$$

Spring forces along the tangent to the MEP are considered for reparameterization, i.e. to have the replicas equidistant from each other and to thus efficiently sample the conformation space between the start and end replica. Similarly, the component of the gradient perpendicular to the tangent is considered for convergence of elastic band towards the MEP.

Hence, the Force $\mathbf{F_i}$ on any NEB image becomes,

$$
\begin{aligned}
\mathbf{F_i} &= \mathbf{F_i^s}||| + \mathbf{F_i}|\perp \\
\mathbf{F_i}||| &= [k\{(\mathbf{x_{i+1}} - \mathbf{x_i}) - (\mathbf{x_i} - \mathbf{x_{i-1}})\}.\hat{\boldsymbol{\tau}}_i]\hat{\boldsymbol{\tau}}_i \\
\mathbf{F_i}|\perp &= -\nabla U(\mathbf{x_i}) + \{\nabla U(\mathbf{x_i}).\hat{\boldsymbol{\tau}}_i\}\hat{\boldsymbol{\tau}}_i
\end{aligned}
\tag{2.17}
$$

where $\hat{\boldsymbol{\tau}}_i$ is the normalized tangent ($\hat{\boldsymbol{\tau}}_i = \boldsymbol{\tau}_i/|\boldsymbol{\tau}_i|$) along the image. The above equations assume that all the springs have the same spring constant, i.e. $k_0, k_1... = k$.

The exclusion of the perpendicular component of the spring force is done to avoid the corner-cutting of the MEP along the regions where it is curved, thereby avoiding the true saddle point. The parallel component of the true force $[-\nabla U(\mathbf{x_i}).\hat{\boldsymbol{\tau}}_i]$ forces the images to slide down, thereby reducing the resolution around the saddle point [92].

The angle-bisector definition of the tangent led to kinks in the Minimum Energy path. A revised tangent definition was introduced as an alternative solution to eliminate kinks. This tangent was defined as a unit vector pointing to the neighbouring image with the higher potential energy. Thus,

$$
\boldsymbol{\tau_i} = \begin{cases} \boldsymbol{\tau_i^+} = (\mathbf{x_{i+1}} - \mathbf{x_i}) & V_{i+1} > V_i > V_{i-1} \\ \boldsymbol{\tau_i^-} = (\mathbf{x_i} - \mathbf{x_{i-1}}) & V_{i+1} < V_i < V_{i-1} \end{cases}
\tag{2.18}
$$

If the image is at a minimum $V_{i+1} > V_i < V_{i-1}$ or a maximum $V_{i+1} > V_i < V_{i-1}$, then:

$$
\boldsymbol{\tau_i} = \begin{cases} \boldsymbol{\tau_i^+}\Delta V_i^{max} + \boldsymbol{\tau_i^-}\Delta V_i^{min} & V_{i+1} > V_{i-1} \\ \boldsymbol{\tau_i^+}\Delta V_i^{min} + \boldsymbol{\tau_i^-}\Delta V_i^{max} & V_{i+1} \leq V_{i-1} \end{cases}
\tag{2.19}
$$

where

$$\Delta V_i^{max} = max(|V_{i+1} - V_i|, |V_{i-1} - V_i|) \quad \text{and} \quad \Delta V_i^{min} = min(|V_{i+1} - V_i|, |V_{i-1} - V_i|) \tag{2.20}$$

The study of the convergence of the elastic band to the MEP can be divided into two subproblems through the use of Nudging:
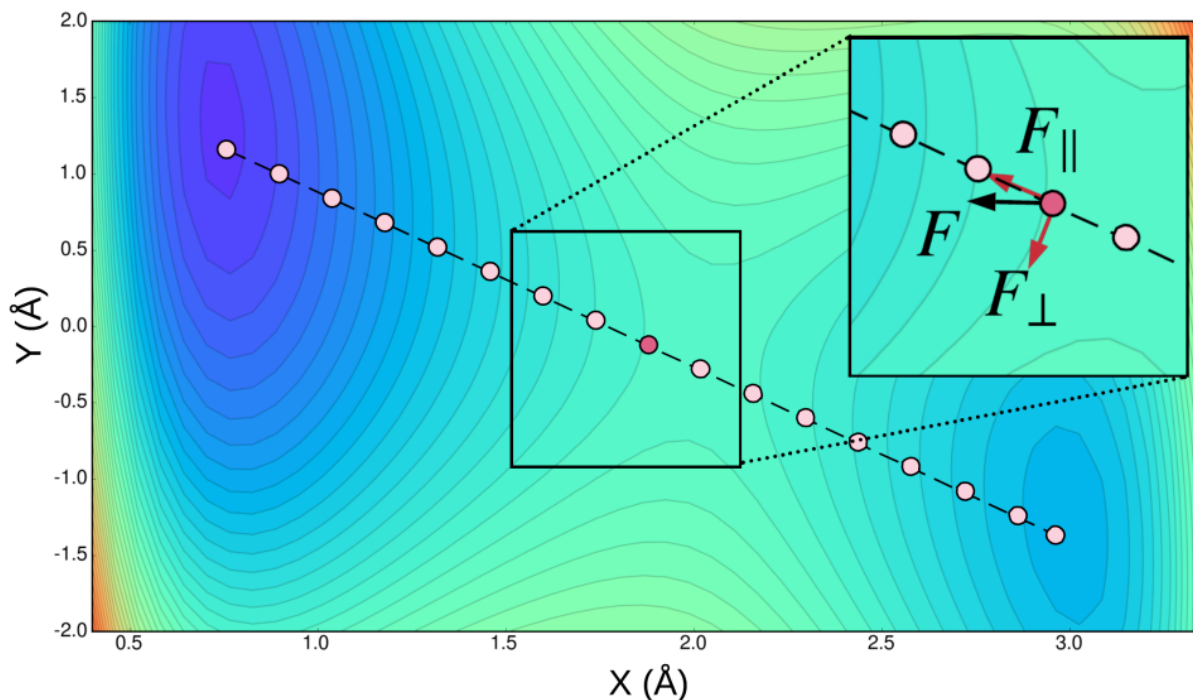
Figure 2.2: A representative image with the initial guess of a Minimum Energy Path on the two-dimensional potential energy surface of LEPS potential coupled to a harmonic oscillator. The inner box shows how forces are decoupled parallel and perpendicular to the elastic band. Figure adapted from [93].

- The natural potential force, which aids in convergence and is perpendicular to the position of the replicas on the elastic band.

- The spring force, which maintains the equidistant spacing of the replicas and allows for a larger conformational space to be sampled during the transitional pathway. The spring constant chosen for this process is arbitrary, as it does not impact either the convergence of the elastic band or the position of the replicas relative to each other.

2.8.1 Convergence Criteria for Nudged Elastic Band Algorithm

As described above, the movement of the replicas on the nudged elastic band is controlled by both spring forces and natural potential forces. While the movement of the replicas along the nudged elastic band does not affect the nudged elastic band itself, it does impact the resolution of the MEP. On the other hand, movement of the replicas in a perpendicular direction to the nudged elastic band does change the MEP. The nudged elastic band is considered to have converged to the MEP when the replicas do not move any further in the perpendicular direction. In other words, the nudged elastic band has reached the

MEP when the replicas are no longer able to move along the perpendicular direction [81, 82, 88–90]. This convergence can be described mathematically as:

$$\mathbf{F}_{i,natural}^{\perp} = -\boldsymbol{\nabla}V(\mathbf{R_i})_{\perp} = 0 \tag{2.21}$$

$\forall\ i \in [0..N\text{-}1]$, where N is the total number of images.

For practical implementations in simulation softwares, the convergence criteria can be said to have met if for each image, the magnitude of the total NEB force acting on the image is less than a specified tolerance force value, $\mathbf{F}_{max}$ [80]:

$$\sqrt{\mathbf{F}_i^{NEB}.\mathbf{F}_i^{NEB}} < \mathbf{F}_{max} \tag{2.22}$$

2.8.2 Simulated Annealing Protocol in Amber MD

To find the global minimum energy pathway (MEP), we follow a simulated annealing protocol in AMBER MD. This process involves dividing the number of images (N) in half, with one half being given the conformation of one end state and the other half being given the conformation of the other end state. These images are then heated at a low temperature with a low spring constant, allowing them to explore different conformations freely, similar to an independent normal MD simulation. The low spring constant value also makes sure the system does not blow up since the images are stacked together at the start. As the process continues, the images are heated at higher temperatures and cooled, allowing them to sample a wider range of conformations and eventually converge on the global MEP. During this process, a higher spring constant is used to evenly space the images and improve the resolution of the global MEP. Finally, the images are cooled to 0K and subjected to quenched MD to make the system's kinetic energy disappear when the nudged elastic band reaches the MEP. The simulated annealing protocol has two main benefits: 1. It allows for a wider exploration of conformational space, increasing the chances of finding the global MEP rather than a local MEP close to the initial guess, and 2. It allows for multiple different MEPs to be found by running the simulation multiple times [93, 94].

2.8.3 Partial NEB implementation in Amber MD to accomodate large explicitly solvated systems

The Simulated Annealing Protocol described above to find MEPs had certain limitations in the types of systems that could be simulated due to the requirement of applying NEB forces to all atoms in the system. This limitation was because of the implementation of the NEB algorithm in Amber MD considered all atoms in the system. This typically made it computationally inefficient to simulate larger systems and also restricted the use of NEB to systems with implicit solvent, as the inclusion of explicit water molecules could introduce artifacts by forcing them to move between the endpoints without allowing them to equilibrate in response to the conformational change. Moreover, in previous implementations of

the NEB method in Amber MD, all images distributed along the path were treated as a single simulation. Running long simulations for large complex systems with multiple images to have enough sampling thus became tricky. Also, the coordinate file included the coordinates for all images, which were later split into discrete structures at the end of the NEB simulation. The parameter file contained a separate copy of the system for each bead, while the nonbonded exclusion list, which was the same for all images and not copied, was used to prevent interactions between images. However, the large size of the nonbonded exclusion list made it impractical to use the particle mesh Ewald method, which would have required distance calculations with PME for each bead in the list in order to remove their interaction in reciprocal space.

These limitations made it difficult to apply NEB to large systems and also resulted in a single reciprocal space grid being used for the entire set of beads, which could limit scaling due to fast Fourier transform calculations.

A new implementation of the nudged elastic band (NEB) method, known as Partial Nudged Elastic Band (PNEB), has been developed in Amber MD, to address the limitations of the previous implementations. This new version, PNEB, allows users to specify which parts of the system the NEB forces should be applied to, such as only the solute and not the solvent atoms. The code has also been rewritten and optimized for greater parallel efficiency and the ability to handle periodic systems. The multisander functionality of the Amber MD engine (SANDER), which runs multiple SANDER jobs simultaneously under a single Message Passing Interface (MPI) program, has been applied to the NEB method. This allows each image to remain a discrete simulation and enables the use of the particle mesh Ewald method with no exclusion list beyond the standard system. Each image also has its own reciprocal space calculation and writes output and trajectory information during its own MD simulation. These changes make it possible to simulate larger systems by spreading the simulations of different NEB images across multiple processors [95].

PNEB (particle mesh Ewald molecular dynamics) was later integrated into the PMEMD (particle mesh Ewald molecular dynamics) module of Amber and further optimized by implementing it with CUDA on NVIDIA GPUs. To minimize data transfer through the message passing interface (MPI) between GPUs, a shuttle transfer was implemented. We have used this PNEB implementation in Amber MD for our MEP calculations [93].

2.8.4 PNEB calculations for Alanine Dipeptide Conformational Transition: Test system

The above defined PNEB algorithm and its implementation in Amber MD program was tested using Alanine Dipeptide as model system. This model was chosen because it is well-studied in the past by various research groups and its potential energy surface is well-explored and well-known to us.

The potential energy surface for alanine dipeptide is a function of its 2 dihedral torsional angles, Phi (Φ) and Psi (Ψ). Phi (Φ) torsional angle is defined as the angle between the plane formed by the atoms
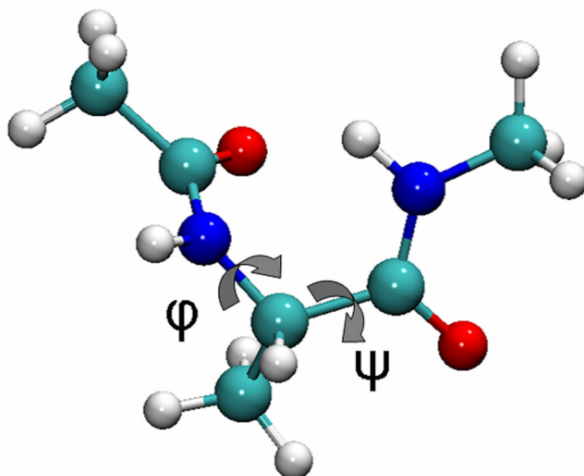
Figure 2.3: Phi (Φ) and Psi (Ψ) torsional angles of Alanine Dipeptide. Figure adapted from [95].

N-Ca-Cb and the plane formed by the atoms C-N-Ca, as seen in Figure 2.3. Psi (Ψ) torsional angle is defined as the angle between the plane formed by the atoms Ca-Cb-C and the plane formed by the atoms N-Ca-Cb, as can be seen in Figure 2.3.

Alanine dipeptide can exist in two stable conformations, which can be seen in the phi-psi contour diagram (Figure 2.4). These conformations are represented by A (phi=-76°, psi=-16°) and B (phi=68°, psi=8°). The diagram shows two energy basins, indicating that there are two distinct conformational states that the molecule can adopt. We intend to find the Minimum Energy Path connecting the states A and B.

Simulated Annealing discussed in the above section was used to find the MEP between Alanine dipeptide isomers. In the first step of simulated annealing protocol, the system was heated to 300K in a duration of 40ps with a Langevin collision frequency of 30ps^{-1}. Springs were used with force constant of $10\text{kcalmol}^{-1}\text{Å}^{-2}$. The second step involved equilibrating the images at 300K for 80ps with langevin collision frequency of 60ps^{-1}. Springs used in this step had a force constant of $50\text{kcalmol}^{-1}\text{Å}^{-2}$. Same spring constant is used for the next step as well. The third step involved heating the images from 300K to 500K and bringing the system back to 300K. This annealing of the system was performed stepwise over 100ps with langevin collision frequency of 70ps^{-1}. In the fourth step, the system was cooled further to 0K over 120ps with langevin collision frequency of 70ps^{-1} and force constant of $10\text{kcalmol}^{-1}\text{Å}^{-2}$ to remove kinetic energy. In the last step, Quenched MD was performed for 100ps with langevin collision frequency of 240ps^{-1} and force constant of $10\text{kcalmol}^{-1}\text{Å}^{-2}$. In all of the above 5 steps, NEB algorithm was applied to all the alanine-dipeptide atoms. The NEB algorithm was not applied to the solvent atoms. The resulting conformation of the images along the MEP is shown on the phi psi contour diagram in Figure 2.4 in black color. The actual MEP, as known previously from past studies [95?], is also shown in white color.

Figure 2.4: Phi psi contour plot of Alanine Dipeptide. The actual MEP for the conformational change of alanine dipeptide is shown with white color, white filled squares representing the images along the MEP. The MEP obtained using Simulated Annealing Protocol in AMBER MD is shown in black color, black filled circles representing the images. Color coding for the Potential Energy Landscape is shown on the right side. Figure adapted from [95].

Simulated Annealing discussed in the above section was used to find the MEP between Alanine dipeptide isomers. In the first step, the system was heated to 300K in a duration of $40ps$ with a Langevin collision frequency of $1000ps^{-1}$. Springs were used with force constant of $10kcalmol^{-1}Å^{-2}$. The second step involved equilibrating the images at 300K for $100ps$. Springs used in this step had a force constant of $50kcalmol^{-1}Å^{-2}$. Same spring constant was used for the next 3 steps as well. The third step involved heating the images from 300K to 500K and bringing the system back to 300K. This annealing of the system was performed stepwise over $300ps$. In the fourth step, the system was cooled further to 0K over $120ps$ to remove kinetic energy. In the last step, Quenched MD was performed for $200ps$. In all of the above 5 steps, NEB algorithm was applied to all the alanine-dipeptide atoms. The NEB algorithm was not applied to the solvent atoms. The resulting conformation of the images along the MEP is shown on the phi psi contour diagram in Figure 2.4 in black. The actual MEP, as known previously from past studies [93, 95], is also shown in white color.
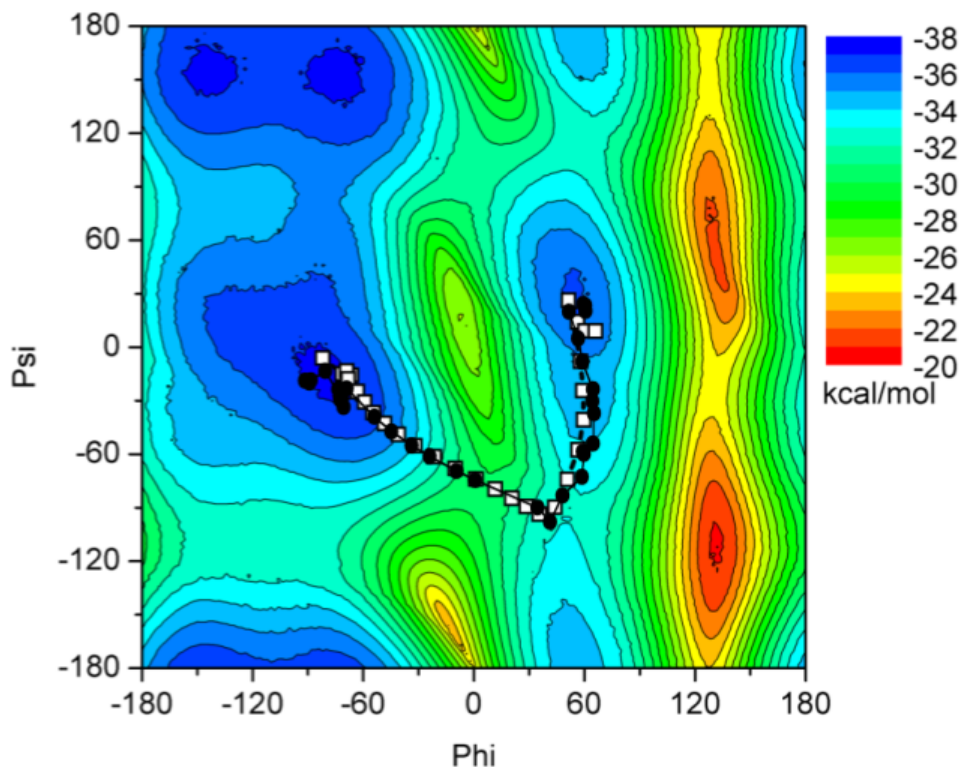
We observe the MEP obtained for the conformational transition between alanine dipeptide isomers via PNEB implementation of NEB method in Amber is almost identical with previously obtained results.

Chapter 3

Molecular Dynamics Simulations of Association of Mismatched DNA and Damage Sensing Protein Rad4/XPC

Contents

3.1 Introduction

DNA repair proteins detect and repair DNA damage in cells and protect the integrity of the genome [96–101]. Given the numerous nucleotide base pairs found in genomes, it remains unclear as to how these proteins search for, interrogate, identify, and accurately repair particular DNA lesions in a crowded cellular environment [102–105]. Deciphering the molecular mechanism behind these proteins' mode of action during DNA repair remains a nontrivial challenge in this field of study [106–108].

The most frequent UV radiation-induced DNA lesion, known as a cyclobutane pyrimidine dimer (CPD), is responsible for many common hereditary skin illnesses and malignancies in humans [13, 14]. Using the nucleotide excision repair (NER) mechanism, the protein xeroderma pigmentosum C (XPC) detects and treats CPD and other lesions of this type [109–111]. Mismatches, tiny and bulky lesions, including CPD, are eliminated with varying efficiency, according to previous investigations on the sensitivity of diverse DNA lesions to the human excision mechanism [112].

Understanding the molecular basis of XPC-mediated DNA damage repair in humans is severely constrained by the lack of crystal or solution structure of the lesion-containing DNA-XPC complex [113]. The yeast orthologue of XPC, Radiation Sensitive 4 (Rad4), is a crucial protein involved in identifying numerous DNA damage types in yeast cells [114, 115]. Rad4 is a valuable model to study
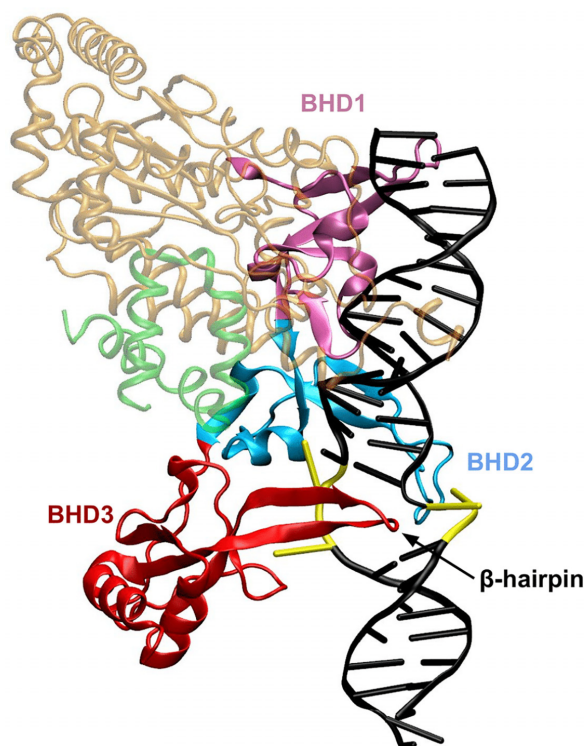
Figure 3.1: The crystal structure of the Rad4-DNA complex with TTT/TTT mismatch. The image was generated using the POV-Ray renderer (www.povray.org) in VMD [116]. DNA strands are represented in black and mismatched basepairs are represented in yellow color. BHD3 domain of Rad4 is colored in red and BHD2 domain is colored in blue.

XPC-mediated DNA damage repair in humans because of its structural and functional similarities to XPC and because the crystal structure of DNA-Rad4 complex is available.

Analysis of crystal structure of the mismatch or CPD-containing DNA-Rad4 complex reveals the DNA duplex is distorted near the mismatch or lesion and the two mismatched base pairs are completely flipped out of the double helix so that they are exposed to water, while the partner bases are safely caught by Rad4 [114]. Additionally, the structure demonstrates that the three β-hairpin domains (BHDs) of Rad4—BHD1, BHD2, and BHD3—are essential for detecting and repairing DNA damage, as shown in Figure 4.2. BHD1 binds to an 11-bp stretch of the matched or undamaged section of the DNA, whereas BHD2 and BHD3 specifically bind to a 4-bp DNA region that contains the mismatches or CPD lesion [114, 115]. The Rad4-DNA complexation process is only statically depicted by the crystal structure, but it nevertheless provides important hints on the following three crucial molecular processes that are most likely to control the dynamics of damage recognition and repair at the initial stage: (a) the association of DNA and Rad4, (b) the insertion of a β-hairpin in BHD3 into the DNA duplex at the mismatch or lesion site, and (c) the flipping out of the pair of nucleotide bases at the mismatch or lesion site [117–119]. Uncertainty persists on the precise process and timing of these events.

The conformational capture or passive mechanism and the associated motion or active mechanism are the two most plausible methods suggested for Rad4/XPC to detect DNA lesions [118–121]. Before Rad4/XPC binds to DNA in the first mechanism, one or two bases opposing the lesion spontaneously flip out. Rad4/XPC searches the damaged DNA and detects and captures these extruded bases. The lesion is then forced to leave the duplex as a result of the BHD3 β-hairpin's insertion [120]. However, in the latter pathway, the BHD3 β-hairpin insertion after Rad4-DNA binding is followed by or takes place during the base flipping and lesion eviction [121].

While noting that the order of these events is probably lesion dependent, Broyde et al. presented findings from a computational investigation that supported conformation capture of the bases [118, 122]. In addition, they investigated the sequence of events for Rad4 binding to 6-4PP and contrasted the excision efficiency between numerous structurally various DNA lesions by resolving the crystal structure of pyrimidine-pyrimidone (6-4) photoproduct (6-4PP)-containing DNA bound to Rad4 [123, 124]. These computer experiments showed that it may be possible to distinguish between poorly and well-recognized lesions based on the initial contact of Rad4's BHD2 with DNA via the minor groove and the subsequent insertion of BHD3's β-hairpin into the DNA duplex.

In a recent fluorescence lifetime study, conformational heterogeneity, inherent deformability, and Rad4-binding selectivity of a set of 3-bp mismatched DNA were compared [122, 125]. The outcomes showed that Rad4 recognizes mismatched DNA with excellent specificity even when it is highly dynamic and conformationally diverse and has higher deformability. That is, the specificity of Rad4 increases with the structural heterogeneity and deformability of mismatched DNA. Rad4 in particular showed strong, medium, and low specificity for the CCC/CCC, TTT/TTT, and TAT/TAT mismatches, respectively, among all the 3-bp mismatched DNA tested.

We investigate the structure, dynamics, and energetics of three mismatched DNA-Rad4 complexes (CCC/CCC, TTT/TTT, and TAT/TAT) using molecular dynamics and umbrella sampling simulations to better understand the mechanism and molecular basis of specificity of mismatch recognition by Rad4/XPC [126]. We used these DNA-Rad4 complexes as model systems to investigate the mechanism of Rad4's mismatch recognition because it is known that Rad4/XPC binds to these mismatched DNA despite the fact that it is not removed by the NER machinery. These three unique mismatch sequences were chosen in consideration of their capacity to span the whole specificity spectrum and the availability of experimental FRET data [125].

This study's main goal is to understand the effects of sequence-dependent specificity and related energetics surrounding the process of Rad4 association with damaged DNA. Rad4's affinity for association with different sequence mismatches is also discussed here.

3.2 Simulation Details

3.2.1 Models

The Rad4-DNA complex with a TTT/TTT mismatch, whose X-ray crystal structure [114] has been resolved at 2.8 Å (PDB ID: 2QSH) was used as a template to build CCC/CCC and TAT/TAT mismatched DNA-Rad4 complexes using the swapna module of the UCSF Chimera software package [127]. There were two regions that were missing in these structures. The first missing region was the loop G(518)R(519) P(520)K(521)G(522)E(523)A(524)E(525) in the BHD2 domain of Rad4, which was modelled using MODELLER v9.22 software. The second missing region was the two disordered thymine bases in the damaged strand of DNA, which were modelled in their extra-helical conformational states using the PSFGEN module of VMD [116, 128]. To solvate each model system, a TIP3P water box of appropriate dimensions was used, to which 15 Na^+ ions were added to neutralize the system [129, 130].

3.2.2 Molecular Dynamics

MD (molecular dynamics) simulations were conducted using the GROMACS 2016.3 simulation package, patched with PLUMED 2.4, with the amber99bsc1 force field for Rad4-free and Rad4-bound DNA [56, 131–134]. The equations of motion were integrated using the Velocity-Verlet algorithm with a time step of 2 fs, and hydrogen bond lengths were constrained using the LINear Constraint Solver (LINCS) algorithm to allow for a slightly larger integration time step of 2 fs [135]. Periodic boundary conditions were applied for all directions using a periodic box with dimensions of 120Å × 95Å × 95 Å in the x, y, and z directions, respectively. Long-range electrostatic interactions were calculated using the particle mesh ewald (PME) method with a direct space cut-off of 14 Å [136, 137]. Pressure was set at 1 bar using the Parrinello-Rahman method with a time constant of 1.0 ps and temperature was set at 300 K using a velocity-rescale thermostat with a time constant of 0.1 ps [138].

In order to minimize the energy of the systems, we followed a two-step process. First, we used the steepest descent algorithm for 50000 steps. Then, we used the conjugate gradient algorithm for an additional 50000 steps [67, 139]. To ensure convergence, we set a criterion such that the absolute value of the maximum force was less than 0.239 kcal mol^{-1} Å^{-1}. After minimizing the energy, we performed three stages of equilibration. In the first stage, we used the NVT ensemble and gradually increased the temperature from 0 K to 300 K over a period of 140 ps, while applying position restraints (with a spring constant of kcal mol^{-1} Å^{-2}) on the heavy atoms of DNA and Rad4. The velocities of the atoms were assigned based on a randomly-seeded Maxwell velocity distribution at 300 K. In the second stage, we used the NPT ensemble at 300 K and 1 bar pressure, with the position restraints still in place, for 2

ns. Finally, we removed all position restraints and performed an additional 3 ns of NPT equilibration followed by a 50 ns NPT production run.

3.2.3 Umbrella Sampling

For studying the association of BHD2/BHD3 domains of Rad4 with the lesion site, a distance based collective variable, ξ, was chosen in accordance with previous studies on the subject [118]. ξ denotes the distance between the COM of the backbone heavy atoms of the two mismatched paired bases of interest and the COM of the backbone heavy atoms of the residues (R494, Y497, M498, N554, F556, N558, E560, F562, P592, V594, S596, F597, P607 and L609) of Rad4 that constitute the binding pocket for the flipped partner bases. The range of ξ (from 5 Å to 21 Å) was divided into 40 bins of size 0.4 Å each. Each US window for ξ was sampled for 3 ns. Since our simulations started off with Rad4 already bound to the DNA, the association of Rad4 and DNA is modelled here as a dissociation process.

Using the reaction coordinate or collective variable (CV) ξ (shown in Figure 3.2) to describe the Rad4-DNA association, umbrella sampling (US) simulations were performed to compute the potential of mean force (PMF) for Rad4-DNA association in the chosen complexes. The conformations obtained from the unbiased simulations have been used as the starting structure for biased simulations for each mismatched complex. The Weighted Histogram Analysis Method (WHAM) (membrane.urmc.rochester.edu/?page_id=126) was used to determine PMFs from CV trajectories obtained from the US simulations [140]. Each US simulation (3 ns trajectory per window) was performed in the NPT ensemble, under the same conditions as those of unbiased MD runs. The spring constants used for biasing potential was 15.0 kcal mol^{-1} Å^{-2}.
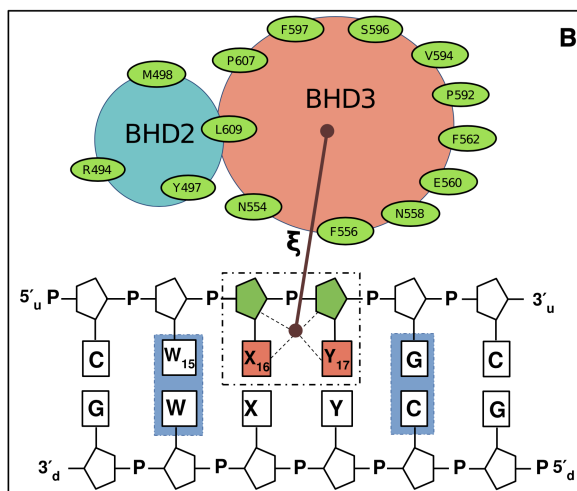

Figure 3.2: ξ is the distance between the DNA damage site (dotted box) and the center of mass (COM) of key residues (green ellipses) in the BHD2/3 domains of Rad4.
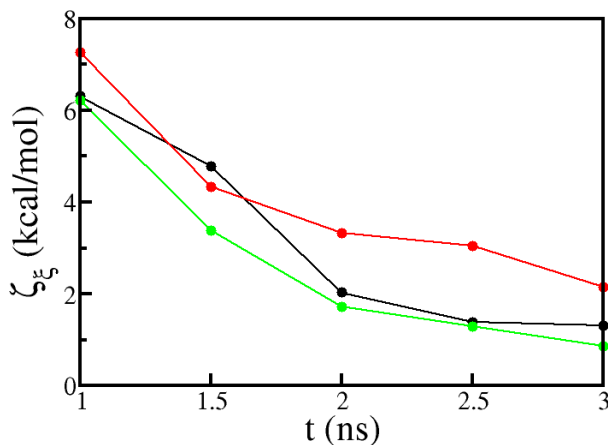
Figure 3.3: Time evolution of ζ

3.2.3.1 Convergence Analysis

To ensure the accuracy of the obtained potential of mean force (PMF) plots from umbrella sampling simulations, a thorough convergence analysis was conducted. The force constants used to bring the system to the center of the simulation window were carefully chosen after testing various values in the range of $0.239 - 45$ kcal/(mol Å^2) and selecting the lowest value that quickly moved the system to the center of the window (in less than 100 ps). The relatively large force constant for the association umbrella sampling studies is due to the separation of the large groups being considered in this analysis.

To assess the convergence of the free energy profiles, the time evolution of a convergence quantity called ζ was monitored. ζ was defined as follows:

$$\zeta = \sqrt{\frac{1}{N}\sum_{D=1}^{N}\Big(F(D, t+\Delta t) - F(D, t)\Big)^2} \tag{3.1}$$

In the analysis, N bins were used to divide the space being considered (D), and the free energy profiles at two adjacent time windows, t and $t + \Delta t$, were calculated. Δt was set to 0.5 ns, so the free energy profiles were recorded at regular intervals of 0.5 ns along the simulation trajectories for each window. The value of ζ was then monitored as a function of time t (where $t = \Delta t$, $2\Delta t$, $3\Delta t$, ... , $n\Delta t$, and n is an integer such that $n\Delta t$ corresponds to the total simulation trajectory length for each window). As the simulation progresses, ζ decreases and approaches zero, and the conformational sampling is considered complete when ζ is less than a certain threshold value. The convergence of the free energy profiles can be seen in the decreasing values of ζ over time, as shown in Figure 3.3.

An evaluation of the probability distributions of the collective variable ξ calculated from the umbrella sampling trajectories showed that there is sufficient overlap between the distributions of the adjacent windows. This overlap is necessary in order to obtain accurate and converged estimates of the free energy from the simulations. The overlap between the distributions can be seen in Figure 3.4.
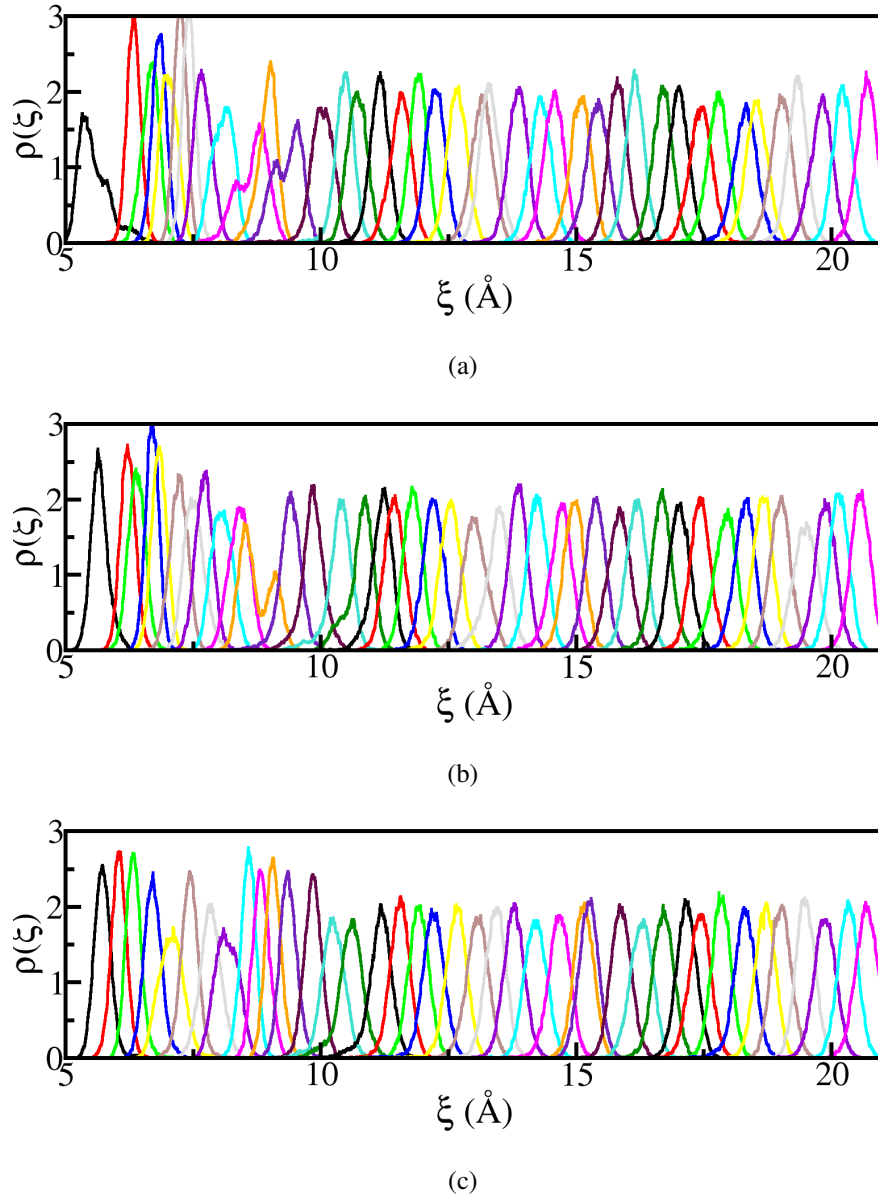
Figure 3.4: The probability distributions of the association collective variable ξ for the umbrella sampling simulation of the Rad4-DNA complex with (a) TTT/TTT, (b) TAT/TAT, and (c) CCC/CCC mismatches.
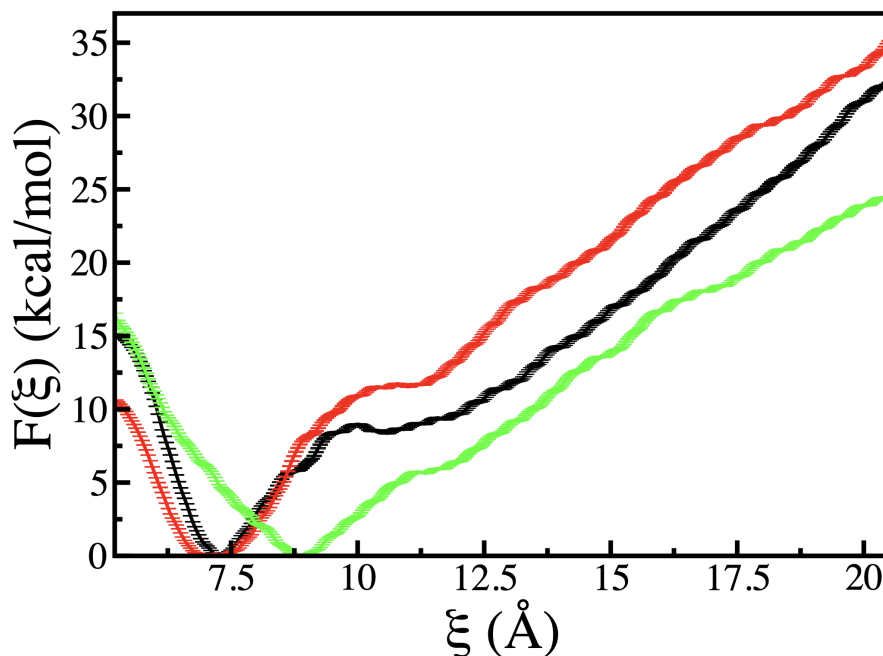
Figure 3.5: Free energy profiles of ξ for TTT/TTT (black), TAT/TAT (red) and CCC/CCC (green).

3.3   Results and Discussion

3.3.1   Free Energy Profiles

Figure 3.5 shows the association free energy profiles $F(\xi)$ obtained from umbrella sampling simulations. The structure of the Rad-DNA complex corresponding to high-energy regions of $F(\xi)$ ($\xi > 20.5$ Å) for the TTT/TTT mismatch is shown in Figure 3.10. For all three systems studied, $F(\xi)$ exhibits an asymmetric parabolic well around the global energy minimum at ξ_{min}, followed by a plateau and a linear increase at higher ξ. The three regimes of interest in the study of Rad4-DNA association are: (1) the low-ξ regime (regime-I) in which Rad4 makes the closest approach to DNA (2) the intermediate-ξ regime (regime-II) in which Rad4 experiences interactions with both water and DNA. (3) the high-ξ regime (referred to as regime-III) in which Rad4 is distantly separated from DNA and it interacts primarily with solvent molecules and only weakly with the DNA.

The interplay between Rad4-water, DNA-water and Rad4-DNA interactions and the ξ-dependent variation in their relative contributions to the overall stability of the complex determine the nature of $F(\xi)$. Owing to close proximity of Rad4 and DNA, and their favourable interactions, the global minimum in $F(\xi)$ is expected to lie in regime-I. The width of the energy well around the global minimum quantifies the "tightness" of Rad4-DNA association.

We examined the variation in the number of hydrogen bonds between Rad4 and DNA (Figure 3.8) and the Rad4-DNA interaction energy as a function of ξ (Figure 3.9). There is a general decrease in the number of hydrogen bonds and a weakening of interaction energies as the BHD2/3 domains move away from the lesion site. The slope change observed in regime-II is a consequence of this.

In regime-III, when ξ is high, the DNA and Rad4 are not associated and are mainly surrounded by water molecules. Therefore, the Rad4-water and DNA-water interactions dominate over the Rad4-DNA interactions in this regime. Since the DNA and Rad4 are homogeneously surrounded by water, $F(\xi)$ is expected to be more or less insensitive to the variation in ξ in this regime. Although complete DNA-Rad4 dissociation would occur for $\xi > 25$ Å, the present study investigates the nature of $F(\xi)$ only for $\xi < 20$ Å.

The location (ξ_{min}) of the global minimum is different for different sequences studied; in particular, ξ_{min} for CCC/CCC is significantly different from that for the TTT/TTT and TAT/TAT mismatches. CCC/CCC docks with the respective domains on Rad4 at a farther distance than TTT/TTT and TAT/TAT, hinting at the high specificity the protein has for it. This corresponds well with the results from the study on β-hairpin insertion, and the farther apart that Rad4 BHD domains can dock on the bases, the more space the β-hairpin has to insert itself into the lesion site and flip the bases out.

We propose that location of the energy minimum might possibly be affected by the presence of favourable interactions between Rad4 and the active site. High specificity of the CCC/CCC mismatch might signal that Rad4 experiences these favourable interactions much earlier than in the case of TAT/TAT mismatch or TTT/TTT mismatch.

3.3.2   Error Estimate of Free Energy Profiles

To further assess the accuracy of our free energy profiles, the Potential of Mean Force (PMF) calculations were bootstrapped using 4 Monte Carlo trials in the WHAM 2.0.9 software. This process helped us to estimate the error in calculating the free energy and produced smaller error bars, which indicates that the convergence of the free energy profiles (as shown in 3.6) is reliable.
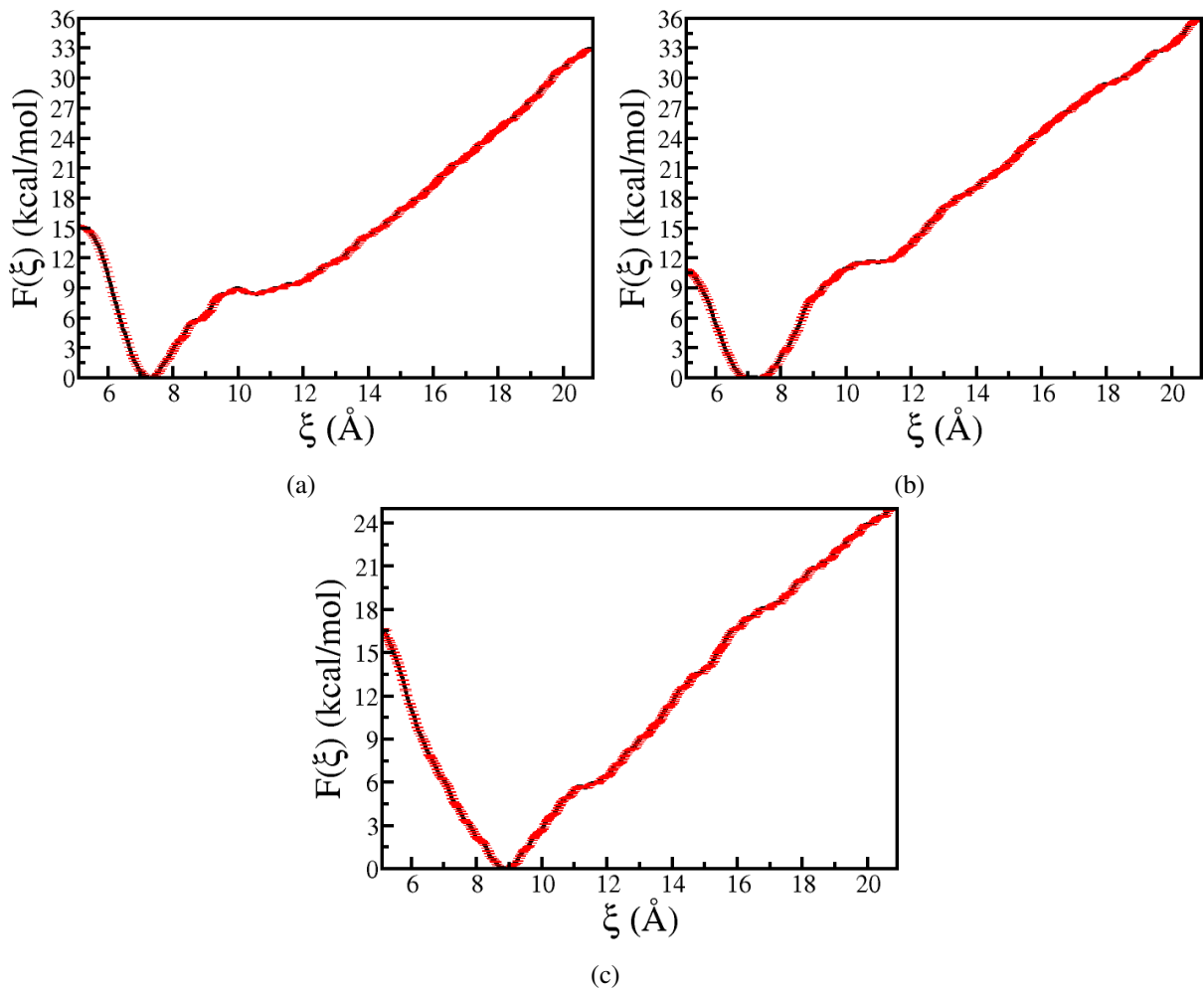
Figure 3.6: The free energy profiles for association (ξ) (black) and WHAM error estimates (red) for (a) TTT/TTT, (b) TAT/TAT and (c) CCC/CCC mismatch.

3.3.3 Free Energy Profile from Molecular Dynamics

We also evaluated the free energy profiles for the association collective variable, ξ, for the three mismatched sequences from the unbiased MD runs. This was done to compare and validate our free energy calculations from US simulations and to validate consistency across our results. Probability distribution of the association CV ξ was calculated from the unbiased MD time-series data(Figure 3.4). This time-series data is obtained by calculating the value of ξ for every frame in the trajectory. The free energy profile can then be obtained from the probability distribution $P(x)$ from the following formula:

$$F(x) = -k_B T \ln P(x) \tag{3.2}$$

Where k_B is the Boltzmann constant and T is the temperature.

40

The free energy profiles obtained from MD trajectory are presented in Figure 3.7. The free energy profiles calculated from the unbiased molecular dynamics (MD) simulations (shown in green) align well with the free energy profiles obtained from the US runs (shown in black) in the area surrounding the global energy minimum basin. This suggests that the unbiased MD simulations accurately capture the energy landscape in this region.
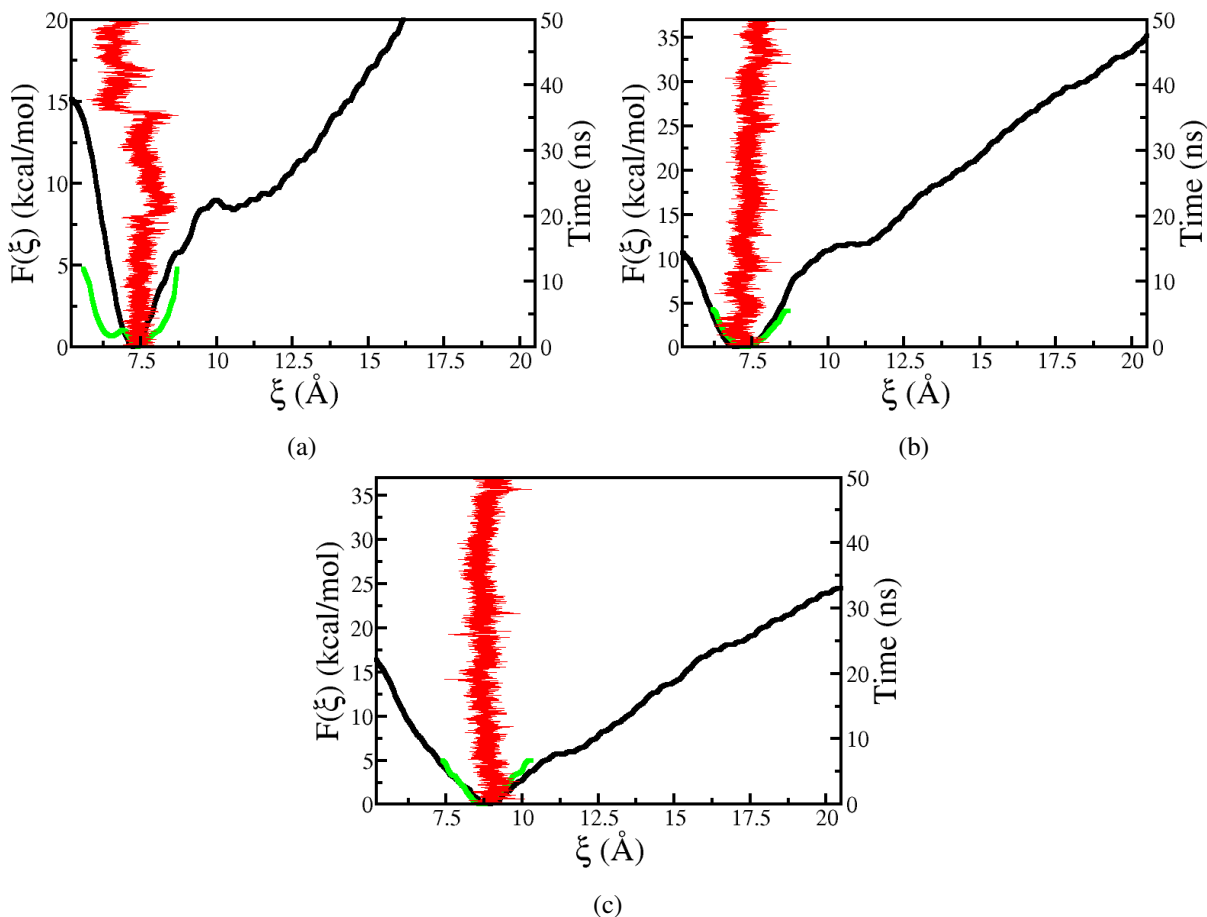
Figure 3.7: The free energy profiles of ξ obtained from US simulation (black) and unbiased MD simulation (green) are shown for (a) TTT/TTT, (b) TAT/TAT and (c) CCC/CCC mismatches. The time series of the collective variable ξ obtained from the unbiased MD is shown in red and the time axis is shown on right.

3.3.4 Hydrogen Bond Analysis

As the Rad4 enzyme starts to dissociate and move away from its bound position, where it is docked with the damaged DNA site, the number of hydrogen bonds between the Rad4-DNA complex is expected to reduce. The hydrogen bond calculator module in VMD was used to calculate the average number of hydrogen bonds between Rad4 and the DNA lesion region as a function of the collective
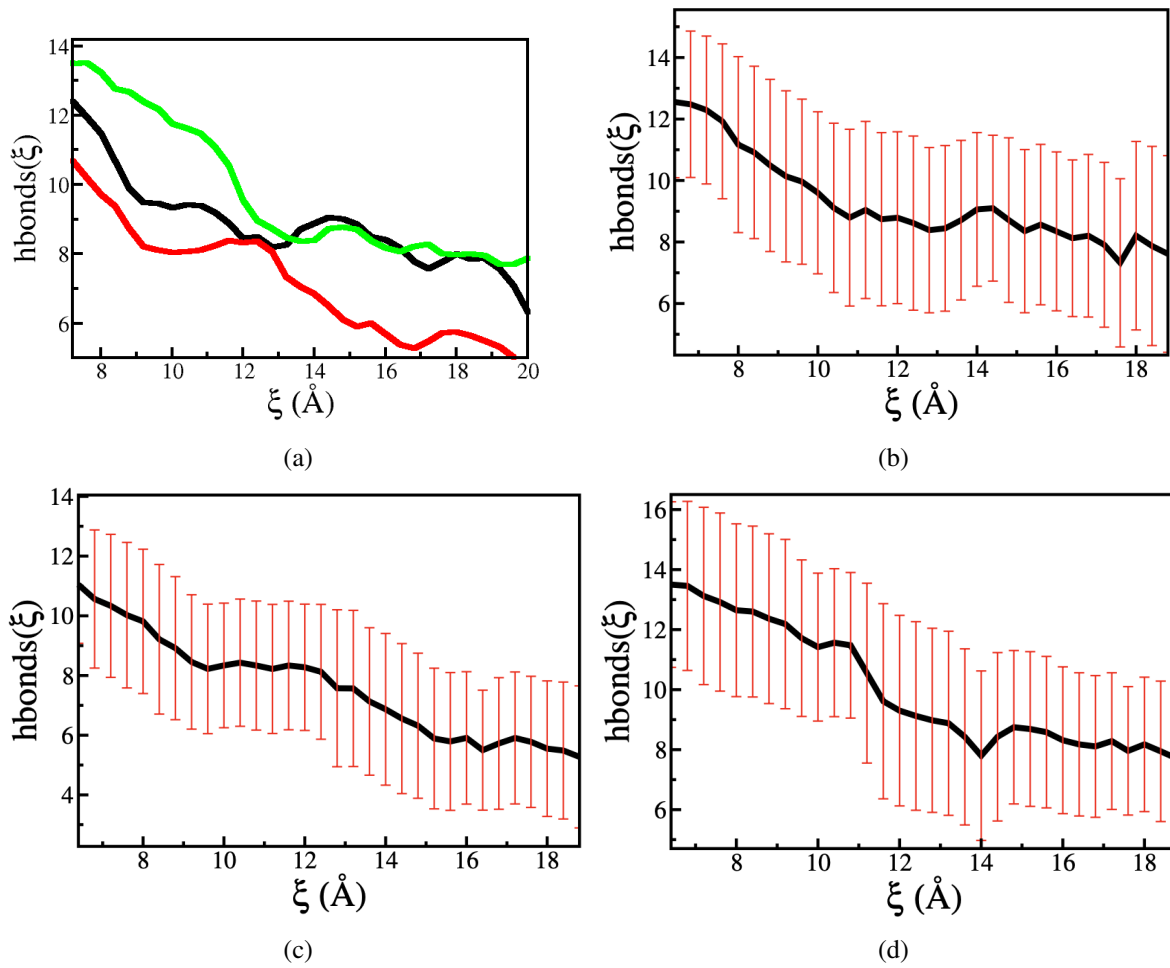
Figure 3.8: The average number of hydrogen bonds between Rad4 and the DNA damaged site as a function of ξ for TTT/TTT (black), TAT/TAT (red) and CCC/CCC (green) derived from US simulations is shown in (a). Panels (b), (c) and (d) also show the associated error with these calculations for the TTT/TTT, TAT/TAT and CCC/CCC systems respectively.

variable ξ, using a distance and angle of 3.5 Å and 35°, respectively. As shown in 3.8, the number of hydrogen bonds decreases as ξ increases, indicating that the dissociation of Rad4 away from the damaged site leads to a decrease in the interaction energies between Rad4 and DNA.

3.3.5   Interaction Energy of Rad4-DNA from US simulations

The umbrella sampling simulations of Rad4-DNA association events were used to calculate the total interaction energy (including van der Waals and electrostatic interactions) between Rad4 and DNA. The mean interaction energy for each window was calculated from the stored trajectories at a time interval of 100 ps. Figure 3.9 shows how the mean interaction energy varies with the collective variable ξ. The data indicates that Rad4 interacts most strongly with DNA in case of TTT/TTT mismatch, followed by CCC/CCC and TAT/TAT mismatches.
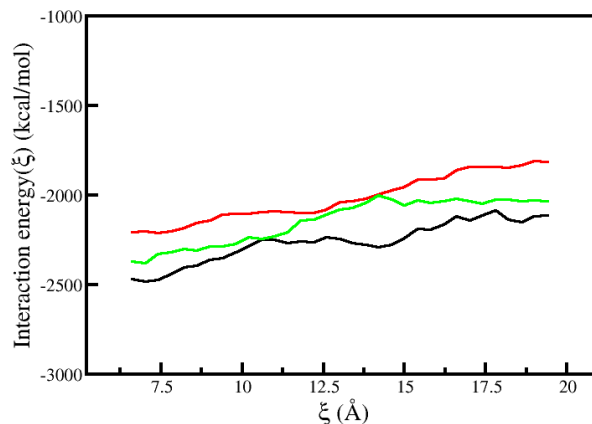
Figure 3.9: The mean interaction energies calculated from the biased US simulation of Rad4-DNA association (bias on ξ) for TTT/TTT (black), TAT/TAT (red) and CCC/CCC (green). The associated errors with these calculations are also shown as vertical bars.

3.3.6   Maximum Energy Structures of Rad4-DNA Complexes

The structure of the Rad4-DNA complex with the highest energy in the high-energy regions of the BHD2/BHD3 association free energy profiles for the TTT/TTT mismatch is shown here. This structure was generated using VMD, and a high resolution image is provided.
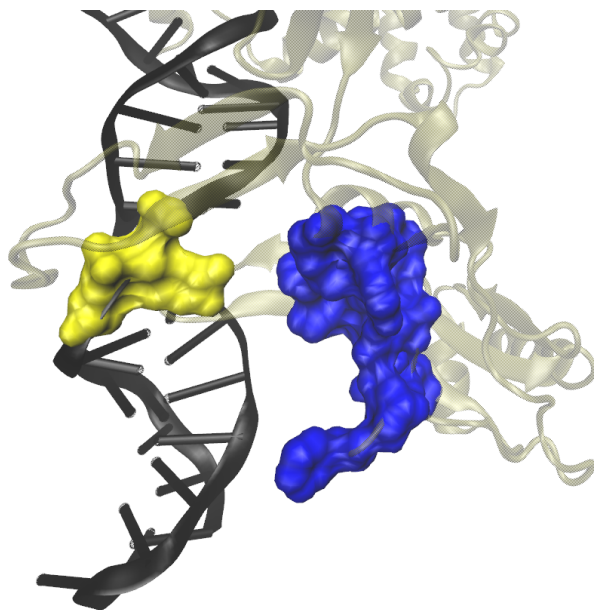
Figure 3.10: The maximum-energy structure of the Rad-DNA complex corresponding to high-energy regions of BHD2/BHD3 association (BHD2/BHD3 domains in blue, and the lesion pocket is shown in yellow) free energy profile for the TTT/TTT mismatch. The DNA is shown in black, and Rad4 in translucent gray.

3.4 Conclusion

In the present study, we have investigated the association of the DNA-damage sensing enzyme Rad4 with different sequence mismatched DNA substrates. We have understood the mechanism, the underlying dynamics and energetics surrounding the association of Rad4 with the DNA damaged site, which happens during the process of damage recognition. From the free energy profiles obtained from umbrella sampling simulations, we see how the docking position, which represents the local minimum on the free energy profiles, also varies with different sequence mismatches. By looking at the free energy profiles, we observed that CCC/CCC mismatch was the first one to be recognised and the Rad4 enzyme associates with this sequence damage before than the other two sequences. We conclude Rad4 exhibiting high specificity for CCC/CCC mismatch, docking further apart and giving enough space for β-hairpin to insert into the damaged site. We also highlight three different regimes in the free energy profiles, where different interactions dominate, depending upon the value of ξ.

It is important to note that in the current study, we have considered only the distance of Rad4 from the DNA damaged site for defining the collective variable for association of the enzyme. In principle, the appropriate angle of approach of the enzyme also needs to be taken into account in defining the suitable collective variables for studying such complex systems. Moreover, the association of the enzyme and

insertion of β-hairpin from the BDH3 domain are actually interdependent events happening during the damage recognition process. Therefore, coupling between these events should be taken into account in order to make more realistic observations for such dynamical processes. We hope to address these issues in future research.

Chapter 4

Minimum Energy Path Analysis of Drug-induced Conformational Transitions of c-Src Kinase

Contents

4.1 Introduction

Protein kinases act as biological switches that toggle between ON (active) and OFF (inactive) states in response to specific cellular signals to regulate various cellular processes including growth, cell proliferation and differentiation. Since protein kinases are implicated in tumorigenesis and malignant cell proliferation, they serve as attractive targets for cancer therapy. Thus, elucidation of the detailed molecular mechanisms of drug-induced inactive→active transition and vice versa exhibited by protein kinases is of paramount importance for rational design of novel therapeutics against cancer and other diseases [22, 24, 141–143].

The c-Src kinase, a member of the Src family of cytoplasmic tyrosine kinases, serves as an excellent model system to investigate mechanistic details of kinase activation and regulation due to availability of extensive experimental data including the crystal structures of both the active and inactive states [36–41, 144–146]. The crystal structures reveal that the two key structural determinants that differentiate active and inactive states of c-Src kinase are the orientation of the αC-helix (residues 304-319) [147] in the amino-terminal lobe and the activation loop (A-loop, residues 404-424) [147, 148] in the carboxy-terminal lobe. The A-loop plays a significant role in recognising substrates for catalytic activity of the kinase. Switching of the kinase conformation between active and inactive states happens via phosphorylation at different sites. Phosphorylation at Tyr416 in the A-loop of kinase domain activates the

kinase, while phosphorylation at Tyr527 in carboxy-terminal tail deactivates it. The A-loop is closed and folded in the inactive state to prevent substrate access into the active site with unphosphorylated Tyr416 residue. In the active state, Tyr416 is phosphorylated and the A-loop opens up to form the binding pocket for substrate binding. The αC-helix has to rotate inwards towards the binding pocket as well for the kinase to be in active state, in order to form the evolutionary Lys295-Glu310 salt bridge, along with other key protein-protein interactions to make the kinase catalytically active. As the αC-helix rotates inwards towards the binding pocket, Glu310 switches its hydrogen bonding partners from Arg409 to Lys295 as the kinase goes from inactive to active state. The evolutionary Lys295-Glu310 salt-bridge is preserved across the kinase family and is necessary to stabilise active kinases. The DFG motif (Asp404, Phe405 and Gly406) is an important structural motif, which is part of the A-loop and is necessary to be in DFG-in conformation for kinase to be in active state. Asp404 of the DFG motif faces outwards from the binding pocket and Phe405 faces inwards towards the binding pocket when the DFG motif is in an "in" conformation. In case of DFG-out conformation, both of these residues flip opposite to their respective DFG-in conformations.

The gatekeeper residue is a crucial component of the ATP binding site. It's relative size, positioning and chemistry controls access of inhibitors to the adjacent binding cavity. In case of type-1 inhibitors, which occupy the ATP binding pocket and do not extend into the adjacent binding cavity, there is generally a presence of conserved water molecules. These water molecules contribute to the binding affinity of type-1 inhibitors by forming water mediated hydrogen bonds, in addition to their contact with the gatekeeper residue itself. Mutations at the gatekeeper site can impact the type-1 inhibitor's binding affinity if it leads to the gatekeeper residue extending into the binding cavity and dislodging the conserved water molecules outside from it.

It is necessary for all structural motifs in the kinase domain to orient properly in order for kinase to be in active state. Whereas, inactive kinase exhibits conformational plasticity [149]. The kinase has to go through a network of conformationally plastic inactive states in order to become catalytically and conformationally active. Although significant research has been carried out to understand the activation mechanism of c-Src kinase, and it being well known that the A-loop has to open outwards first, followed by the inward rotation of αC helix [147, 148], the precise mechanism of how the binding of a drug (for example, bosutinib) to c-Src kinase induces the inactive → active conformational transition via key structural intermediates along the transition pathway are still need to be explored in depth. However, the experimental and computational characterization of these conformational transitions is nontrivial for the following reasons. Earlier studies have indicated that Type-1 inhibitors (which occupy the ATP binding pocket) on c-Src kinase have residence times of the order of seconds or more, which are beyond the time scales accessible to traditional molecular dynamics simulations [150–153]. Moreover, these transitions involve barrier crossing events on the underlying rugged potential energy surface and the time scales of such rare events are also too long to be captured in conventional MD simulations. Thus, most of

the modern computational studies on drug-bound c-Src kinases rely on enhanced sampling techniques and/or coarse-grained models to probe these transitions. The biased sampling methods rely on a limited set of collective variables (CVs) and seek to achieve uniform sampling of a reduced space defined by these CVs. The requirement of predefined CVs limits the usage of such methods to systems with known CVs. However, it is non-trivial to apply them to complex systems with unknown CVs. Moreover, the identification of suitable CVs for biomolecular processes involving multitude of coupled degrees of freedom is a challenging task.

Previous studies on inhibitor-bound Src kinases have highlighted key conformational changes in the kinase domain as the inhibitor associates/dissociates from the binding pocket [154]. A recent study on dasatinib-bound chicken c-Src kinase revealed that the Lys295-Glu310 salt bridge first becomes water mediated and then it merely acts as a switch for drug unbinding to proceed. This eventually leads to penetration of water molecules into the binding pocket and dissociation of the drug after breakage of key interactions with kinase residues, which is the actual rate-determining step of the process. Thus, water appears to be an active participant in the drug binding-induced kinase conformational changes [155, 155–162]. However, it is not clear whether the initial penetration of water into the drug binding pocket to make the Lys295-Glu310 salt bridge interactions water mediated precedes or succeeds the disruption of the Lys295–Glu310 salt bridge and much remains to be understood about the order of these molecular events and the correlation/coupling between them [154].

Nudged Elastic Band (NEB) method is a useful computational technique to determine the minimum energy path (MEP) between two stable states of a system [81–91]. It works by simulating multiple replicas, known as intermediates, and converging them to a path on the potential energy landscape with the least overall activation energy cost for the transition. Unlike many biased sampling methods that rely on one or two collective variables, the MEP obtained via NEB does not require the knowledge of any apriori collective variables and hence it reveals a true sense of the conformational transition on the 3N-dimensional configurational space (here N is the total number of atoms in the system). In this study, we employ the NEB method to determine the MEP between the drug-bound active state and drug-free inactive state of human c-Src kinase. The characterization of change in the number of water molecules in the binding pocket and the structural and energetic changes in the protein as it transits between the end states via the identified MEP provides us with key insights into the drug-induced c-Src kinase activation mechanism.

4.2 Modelling and Simulation Details

4.2.1 Models

The active (PDB ID: 4MXO) and inactive (PDB ID: 2SRC) X-ray crystal structures of the human c-Src kinase were used to generate two end states for performing NEB simulations to find the MEP

[144, 145]. Only the catalytic domain residues comprising of residues TRP260–THR521 (262 residues in total) are included in this study. Since we wanted to study the process and mechanism of how the drug binds/unbinds as the kinase transitions between active and inactive states, we have considered the active state crystal structure 4MXO which contains type-1 inhibitor bound inside the ATP-binding pocket of the active kinase conformation as one of the NEB end states. The corresponding inactive kinase + drug end state consists of kinase conformation from PDB ID 2SRC with drug kept away from the binding pocket. Missing kinase residues were modelled using Charmm-GUI [163]. Both the end states were solvated in a cubic TIP3P water box with 10 Å as the water-padding cutoff distance using tleap software in AmberTools. 3 K+ ions were added to neutralise the system.

The X-ray crystal structures of the entire catalytic domain (residues TRP260–THR521) of the human c-Src kinase in the active (PDB ID: 4MXO) and inactive (PDB ID: 2SRC) states were chosen as starting structures to generate end states for NEB calculations [144, 145]. The active state model consisted of a type-1 inhibitor (Bosutinib) bound at the ATP-binding pocket, whereas the inactive state model was generated by taking kinase atoms from PDB ID: 2SRC and keeping the inhibitor away from the binding pocket. The NEB-derived minimum energy path between these two end states enables us to probe the mechanism of how the drug binds/unbinds from the c-Src kinase as it undergoes transitions between active and inactive states. All the missing residues were modelled using CHARMM-GUI [163]. Each end state was solvated in a cubic TIP3P water box with 10 Å of water padding using tleap software in AmberTools. Three K^+ ions were added to neutralise the systems.

4.2.2 Simulation Details

Initially, the minimization of the total potential energy of the end states was carried out in two stages: in the first stage, 5000 steps of steepest descent followed by 5000 conjugate gradient steps were carried out with position restraints (with a harmonic spring constant of 500 kcal mol^{-1} Å^{-2}) on the protein atoms [67, 164]; in the second stage, 25000 steps of steepest descent followed by 25000 conjugate gradient steps were carried out without any position restraints [67, 139, 164]. The energy minimized structures were then subjected to temperature annealing from 0 K to 300 K under NVT conditions for 20 ps with weak position restraints using harmonic spring constant of 10 kcal mol^{-1} Å^{-2} on protein and drug atoms. Subsequently, the systems were equilibrated in the NPT ensemble at 300 K and 1 bar pressure for 200 ps with weak positional restraints (harmonic spring constant of 10 kcal mol^{-1} Å^{-2}) on drug atoms only.

4.2.3 NEB Simulations

Simulated Annealing protocol has become a popular method to find MEPs using NEB algorithm for conformational transitions in large-scale complex systems [153, 165–170]. We have used Simulated
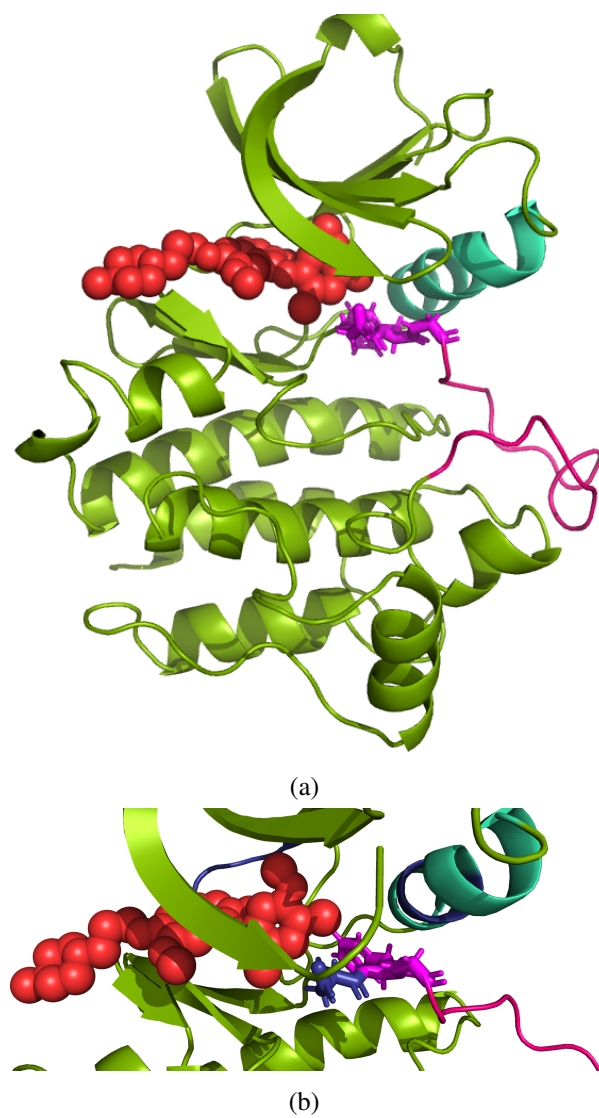
49

(a)

(b)

Figure 4.1: Crystal structure of (a) active (drug-bound) c-Src kinase and (b) Zoomed in portion of drug bosutinib in the binding pocket of active kinase.
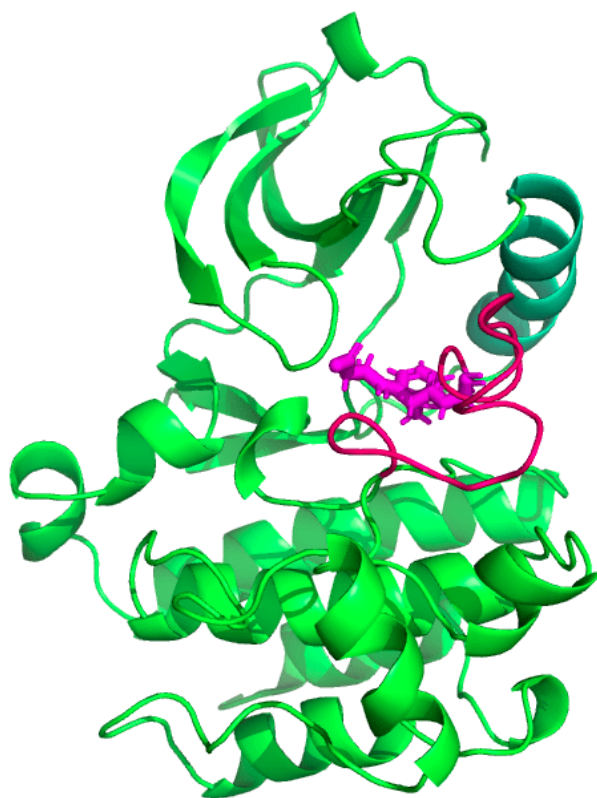
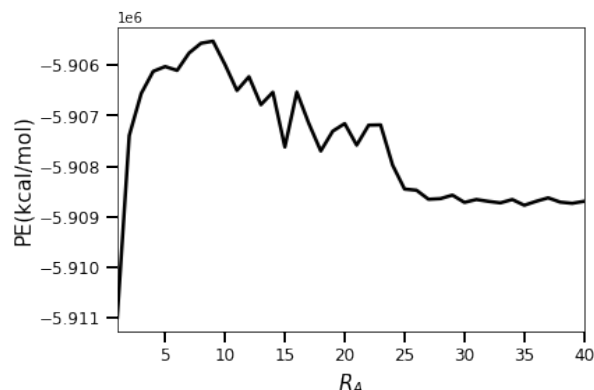Figure 4.2: Crystal structure of inactive form of c-Src kinase

Figure 4.3: Variation of sum of potential energies of all images as 5COOL2 NEB step progresses

Annealing protocol for our simulations, which consists of 5 steps. For running NEB simulations, we consider 40 images, 20 starting from one end state and 20 starting from other end state. These end states are the post-equilibrated configurations obtained via above method.

In the first step of simulated annealing protocol, the system was heated to 300K in a duration of 80ps with a Langevin collision frequency of 30ps^{-1}. Springs were used with force constant of $10\text{kcalmol}^{-1}\text{Å}^{-2}$. The second step involved equilibrating the images at 300K for 200ps with langevin collision frequency of 60ps^{-1}. Springs used in this step had a force constant of $50\text{kcalmol}^{-1}\text{Å}^{-2}$. Same spring constant was used for the next step as well. The third step involved heating the images from 300K to 500K and bringing the system back to 300K. This annealing of the system was performed stepwise over 100ps with langevin collision frequency of 70ps^{-1}. In the fourth step, the system was cooled to 0K over 120ps with langevin collision frequency of 70ps^{-1} and force constant of $10\text{kcalmol}^{-1}\text{Å}^{-2}$ to remove kinetic energy. In the last step, Quenched MD was performed for 100ps with langevin collision frequency of 240ps^{-1} and force constant of $10\text{kcalmol}^{-1}\text{Å}^{-2}$. In all of the above 5 steps, NEB algorithm was applied to all protein and drug heavy atoms. The NEB algorithm was not applied to the solvent atoms.

4.2.4 Convergence calculations along the MEP

In order to ensure convergence of MEP, we examined the evolution of the sum of potential energies of all the images (Figure 4.3) and also the potential energy of an arbitrarily chosen image (Figure 4.4) as the last step in the NEB run progresses. It is observed that both of these energies remain more or less constant beyond certain number of steps indicating convergence of the NEB results.
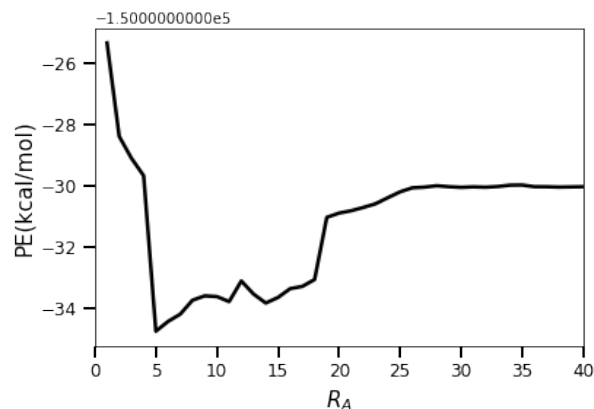
Figure 4.4: Variation of potential energy of a representative NEB image as 5COOL2 NEB step progresses

4.3 Results and Discussion

4.3.1 Structural changes along MEP

The minimum energy path for the conformational transition of Human c-Src kinase between the bosutinib-bound active state and bosutinib-free inactive state was obtained using the protocol described above. The MEP is represented on two-dimensional planes defined by pairs of structural parameters that quantify variations in some key structural motifs in the kinase domain, as seen in fig. 4.5. Following ref. [147, 148], the present work considers the two-dimensional planes defined by the following structural parameters to showcase the calculated MEP: (a) the distance (D_1) between atom NZ of Lys295 and atom CD of Glu310, (b) the distance (D_2) between atom CD of Glu310 and atom CZ of Arg409, (c) the root mean square deviation of the activation loop comprising of residues 404-424 (R_A) with respect to its inactive state conformation, (d) the root mean square deviation of the αC-helix comprising of residues 304-319 (R_C) with respect to its inactive state conformation, and (e) the root mean square deviation of the DFG motif comprising of residues 404-406 (R_{DFG}) with respect to its active state conformation.

The calculated MEP reveals key structural changes exhibited the kinase domain during drug-induced active-to-inactive state transitions. In Figure 4.5a, D_2-D_1 shows significant variation along MEP as Glu310 changes its hydrogen-bonding partners between Lys295 and Arg409, which lead to the breakage of the evolutionary Lys295-Glu310 salt bridge and the crucial interactions that stabilized it. The Lys295-Glu310 salt bridge is preserved in the active state across the kinase family and the breakage of Lys295-Glu310 salt bridge along with consequential changes in interactions with Arg409 and in other structural motifs appear to be a key step for the kinase deactivation.
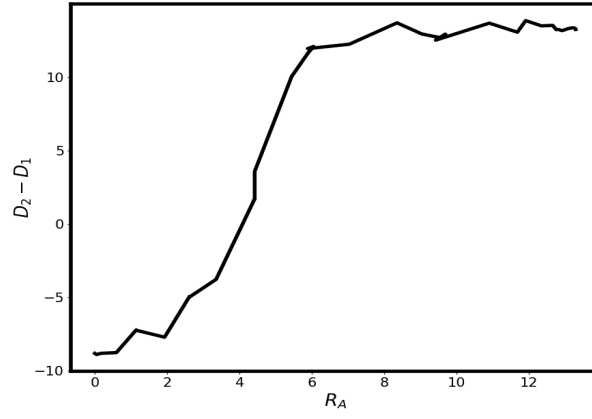
(a)
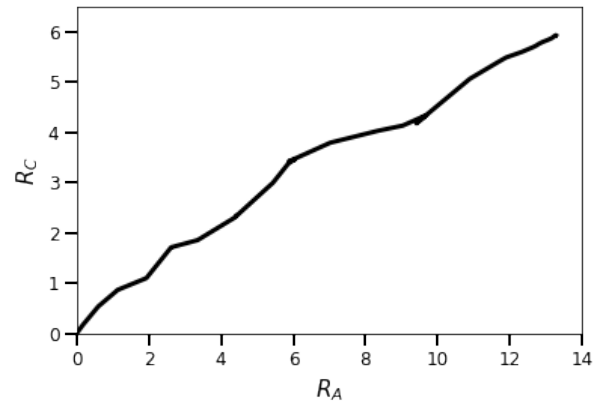
(b)
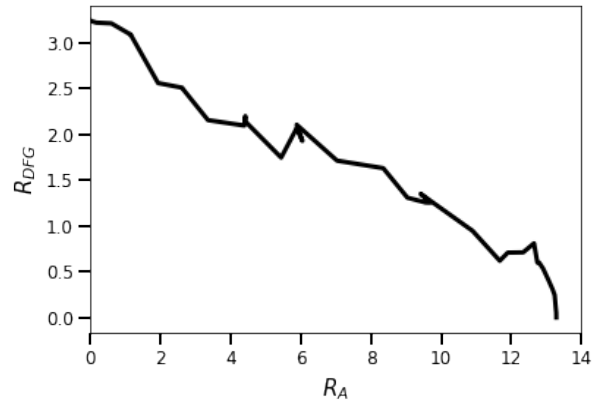
(c)

Figure 4.5: Variation of (a) D_2-D_1, (b) R_C and (c) R_{DFG} vs R_A along the MEP

Figure 4.5b shows the variation of R_C and R_A along the MEP. The variation of R_C and of R_A show more or less a linear dependence indicating that the rotation of αC-helix and activation loop go hand in hand along the MEP, as the drug dissociates from the binding pocket.

Figure Figure 4.5c shows the variation of R_{DFG} and R_A along the MEP. The functionally important DFG motif, which is part of the activation loop, stays in the flipped-in conformation when the activation loop is open for the kinase to be active state, whereas it is flipped-out when the activation loop is closed in inactive kinase conformation. Phe residue of the DFG motif in DFG-in conformation forms the base of the binding pocket and stabilises the drug in the binding pocket by forming π-π stacking interactions with the phenyl ring of the drug. The Asp and Gly residues at the binding pocket also contribute to stability of the drug in the binding pocket.

The mechanism of conformational transition and the nature and number of intermediate states visited during the active to inactive kinase conformational transition may differ based on whether it is examined in the presence [154, 171, 172] or absence of the drug [147, 148, 154, 171–178].

4.3.2 Properties governing drug-induced conformational transitions

The crystal structure of bosutinib-bound c-Src kinase (Figure 4.1a) reveals that the inhibitor is bound at the ATP binding site and makes contacts with the binding pocket residues and also makes hydrogen bonds with the water molecules trapped inside the binding pocket. These interactions contribute majorly to the binding affinity of such type-1 inhibitors and it is necessary to break these interactions for drug dissociation to occur.

4.3.2.1 Distance between drug and binding pocket

Following Ref. [154], the distance (D_d) between the centers of mass of the binding pocket residue Met82 (to be precise, Met82:O and Met82:N atoms) and the drug (N and N1 atoms of the drug) was chosen as a measure of the distance between the drug and the binding pocket of c-Src kinase. Figure 4.6 depicts the variation of D_d as the system evolves from the drug-bound active state to the drug-free inactive conformation along the MEP. A marginal drop in D_d at around frame 15 indicates that the drug is moved a little shallower inside the binding pocket such that the nitrogen atoms of the drug come closer to the O and N atoms of Met82. Therefore, we see a dip in the absolute value of D_d as the drug moves out of the binding pocket.

4.3.2.2 Number of water molecules inside binding pocket

Figure 4.6 shows the change in the number of water molecules present in the binding pocket as the system evolves along the MEP. In this analysis, all the water molecules located within 6 Å from Thr79:Og1, Glu80:O, and Asp145:Ca are considered to be inside the binding pocket of the kinase.
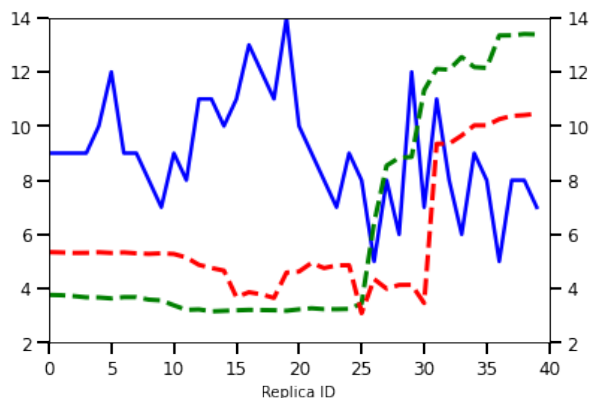
Figure 4.6: Variation of W_{bp} (blue), D_d (red) and D_l (green) along the MEP.

4.3.3 Our Proposed Mechanism

Figure 4.6 shows the variation of the number of water molecules (W_{bp}) at the binding pocket (blue), the distance (D_d) of drug from the binding pocket (red) and the distance (D_l) between Lys295-Glu310 salt-bridge residues (green) along the MEP.

As the system evolves from the drug-bound active state to the drug-free inactive state via the MEP, the distance between drug and binding pocket remains more or less constant between 4-5 Å for the first 30 frames, but abruptly increases to values greater than 9 Å in the later frames. Prior to this transition, the distance between gate residues increases abruptly from 4 Å to values greater than 12 Å between frames 25-30. Preceding the observed transitions in D_d and D_l, the number of water molecules inside the binding pocket increases significantly between the frames 10 and 20, with a negligible change in D_d and D_l. However, the transition in D_l precedes a decrease in W_{bp} during frames 17-25, but W_{bp} again increases and exhibits noticeable fluctuations beyond frame 25 on the MEP. These results offer a plausible sequence of events that are likely to occur during the drug-induced conformational transition of the kinase from the active state to the inactive state via the MEP.

For frames 10-20, we conclude the water molecules entering the binding pocket push a drug little outside the binding pocket. This is only a small push, which is reflected by a small dip in the D_d (green) in Figure 4.6. Since there was no change in D_l, these water molecules did not penetrate via Lys295-Glu310 salt-bridge but via other pathways. These water molecules were eventually unsuccessful in dissociating the drug from the binding pocket, as can be seen from value of D_d just decreasing slightly between frames 10-20. The interaction energy of the drug with the key hydrophilic and hydrophobic residues in the binding pocket and with the salt-bridge residues did not change much in this section, as can be seen in Figure 4.7, Figure 4.8.

For frames between 25 and 35, an abrupt increase in D_l is observed (green), which indicates the Lys295-Glu310 salt-bridge opens prior to the unbinding of drug from the binding pocket. Coinciding

56
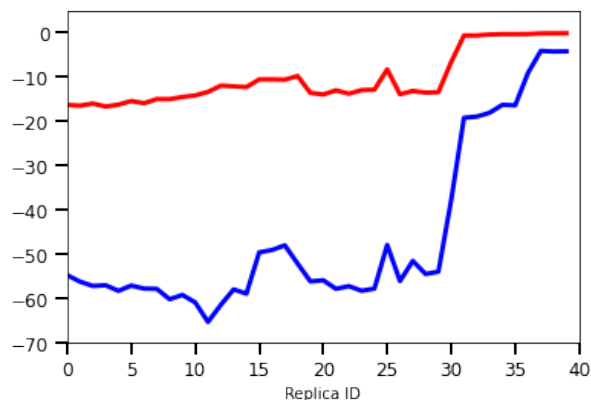
Figure 4.7: Variation of total interaction energy of drug with hydrophilic (blue) and hydrophobic (red) residues inside the binding pocket along the MEP
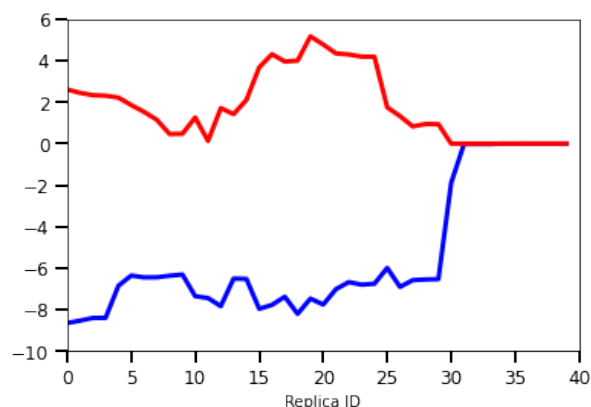
Figure 4.8: Variation of interaction energy of drug with salt-bridge residues Lys295 (blue) and Glu310 (red) along the MEP

with the opening of the gate, W_{bp} increases sharply, indicating a coupling between opening of the Lys295-Glu310 salt-bridge and dissociation of the drug, which this is triggered by water molecules penetrating the binding pocket in frames 25-35.

To further understand the mechanism, we examined the variation of the interaction energy of the drug with each of the Lys295-Glu310 salt-bridge residues along the MEP and the results are shown in Figure 4.8. It is observed that the drug interacts favourably with Lys295 for frames from 0 (active state) to frame 30 (onset of drug unbinding) and this interaction breaks at frame 30 leading to the dissociation of the drug from the binding pocket. We conclude that the water molecules that entered into the binding pocket between frames 25-35 trigger the breakage of drug-Lys295 interactions that results in the dissociation of the drug from the binding pocket. Figure 4.9 shows the number of water molecules in a 6 Å radius around drug and Lys295. We observe an abrupt increase in the value of $W_{d-Lys295}$ after the opening of the Lys295-Glu310 salt bridge. Value of $W_{d-Lys295}$ increases significantly just prior to
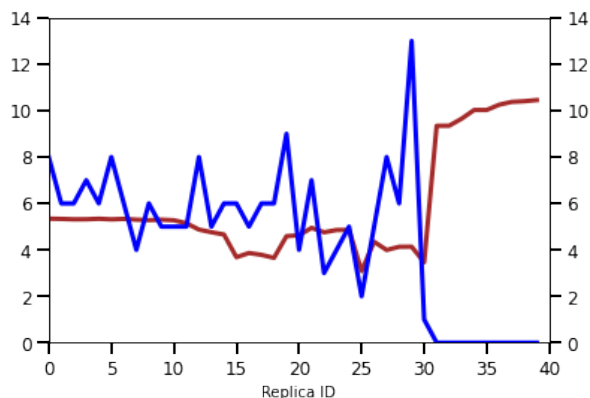
Figure 4.9: Variation of D_d (brown) and $W_{d-Lys295}$ (blue) along the MEP

the dissociation of the drug (brown), supporting the hypothesis that water entering via Lys295-Glu310 salt bridge triggers the breaking of drug-Lys295 interactions and leads to drug unbinding. W_{bp} in Figure 4.6 (blue) decreases once again after the drug dissociates completely, thus indicating the crucial role of water molecules in the drug dissociation process of c-Src kinase.

4.3.3.1 Single Entry Path of Water inside Binding Pocket

As concluded above, the opening of the Lys295-Glu310 salt-bridge creates a pathway for the entry of additional water molecules into the binding pocket and thereby triggering the unbinding of drug from the binding pocket. To examine the variation in the hydration of the salt-bridge residues during the active to inactive transition, we computed the number of water molecules ($W_{Lys295\text{-}Glu310}$) that are located within 6 Å from Lys295-Glu310 salt-bridge residues. The change of $W_{Lys295\text{-}Glu310}$ along the MEP is shown in Figure 4.10. An increase in $W_{Lys295\text{-}Glu310}$ is observed from frame 25, which suggests that the entry of additional water molecules via the distorted Lys295-Glu310 salt-bridge is necessary and a key step for drug dissociation. After the drug is dissociated from the kinase, some water molecules leave the binding pocket via different pathways giving rise to a relatively less hydrated binding pocket in the inactive conformation.

The MEP trajectory of a representative water molecule that enters into the binding pocket via the distorted Lys295-Glu310 salt-bridge is shown in Figure 4.11. The representative water molecule is shown in muave colored VMD representation.

4.3.4 MEP as a single overlapped image

Figure 4.12 shows configurations of the system along the MEP overlaid with one another. For the kinase, only the changes in the αC-helix and A-loop conformations are shown for clarity and the remaining parts correspond to the active state conformation of the kinase. The yellow colored VMD

58
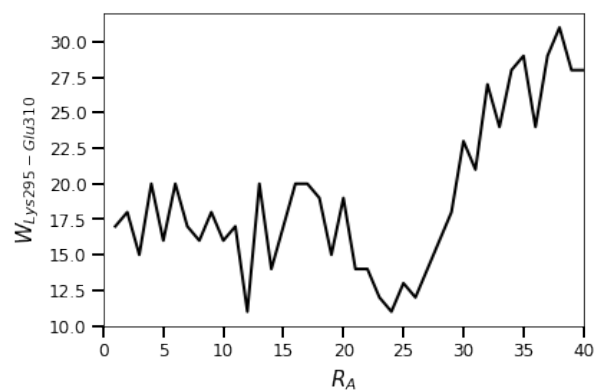
Figure 4.10: Variation of $W_{\text{Lys295-Glu310}}$ along the MEP
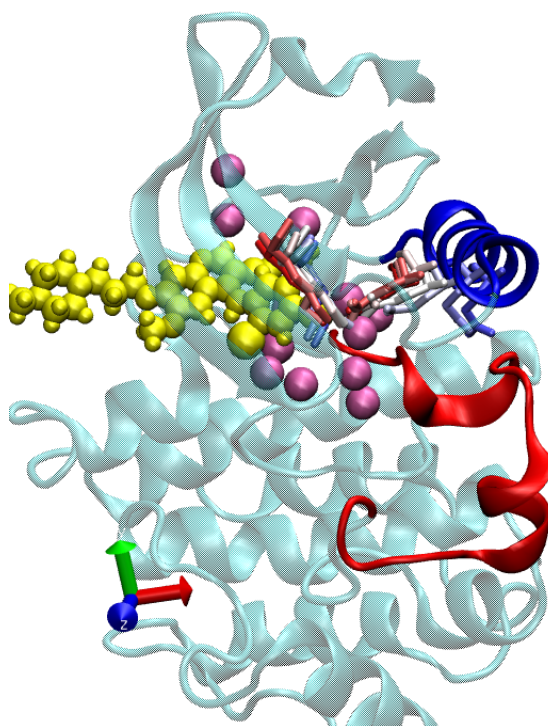
Figure 4.11: Trajectory of a representative water molecule via conserved Lys295-Glu310 salt-bridge pathway
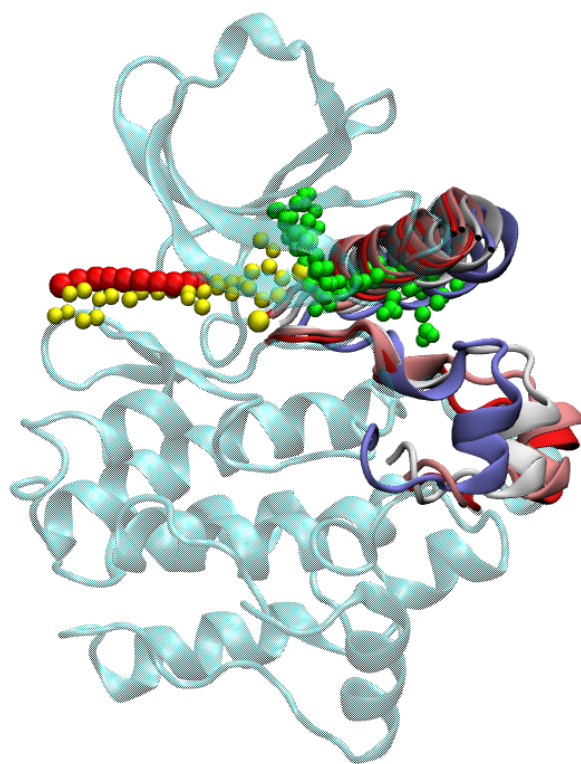
Figure 4.12: Snapshot representing the major structural feature changes (A-loop, αC-helix, Lys295-Glu310 salt bridge residues) in kinase conformation along the MEP

representation is the drug bosutinib, associating/dissociating from the binding pocket. The green colored VMD representation is for Lys295-Glu310 salt bridge residues. Red colored VMD representation represents the center-of-mass of drug binding/unbinding from the binding pocket.

4.4   Conclusion

In this work, we have investigated the dynamics of drug-induced conformational transitions between active and inactive states of human c-Src kinase via minimum energy path using nudged elastic band algorithm. The minimum energy path analysis gives us key insights into the mechanism of drug binding/unbinding, as the kinase transitions between active and inactive states. We highlight the crucial role of water in drug association/dissociation and its coupling with key kinase residues and the distance between drug and the binding pocket. We observe the Lys295-Glu310 salt bridge distortion as a key event, that triggers the drug dissociation process. This distortion creates a conserved entry path for water molecules to penetrate the binding pocket and break its key interactions with the kinase residues. Water penetration via other pathways is not effective in the drug binding/unbinding process and we also observe water leaving the binding pocket once the dissociation has taken place. Our study elucidates the crucial role of water and the coupling between these processes during drug-induced conformational transitions. The actual mechanism of transition might differ from ours, which can be studied by taking the kinetic energy component into consideration as well and calculating free energy profiles using our MEP as order parameter. That been said, the molecular mechanisms elucidated in our study will be useful for rational design of novel therapeutics against cancer and other kinase-related diseases.

Chapter 5

Conclusion

Proteins and DNA are complex dynamical systems that form essential components of cell biology. The conformational dynamics of these macromolecules is crucial for the cellular biological functions and characterization of binding-induced changes in their conformational dynamics becomes crucial to understand the fundamental processes of life. Molecular dynamics simulations serve as useful tools to understand the mechanics, energetics and kinetics of such complex systems computationally on an atomic level. Proteins engage with DNA to recognise and repair DNA damages and interact with inhibitors to de-regulate downstream processes in case of diseases like cancer. The rugged nature of potential energy surfaces of DNA and proteins makes it extremely difficult to study the conformational dynamics of such binding-induced molecular events using traditional molecular dynamics techniques. Such conformational changes involve a network of metastable states, barrier crossing events and the timescales to capture these conformational changes is well out of the scope of normal MD simulations. Enhanced sampling methods and path finding algorithms are therefore used extensively to simulate conformational transitions in complex biological systems to study conformational dynamics and energetics surrounding their cellular processes. In this thesis, we have employed umbrella sampling technique (an enhanced sampling method) and nudged elastic band method (a path finding algorithm on the potential energy surface) to study the conformational dynamics surrounding the following two molecular processes: (a) The association of DNA damage sensing protein Rad4/XPC with the DNA damaged site, and (b) Drug-induced conformational transitions and the activation/deactivation mechanisms of human c-Src kinase as a type-1 inhibitor binds/unbinds from the ATP binding pocket.

Initially, we cover the basic scientific knowledge one should have about DNA and protein kinases. Their functions, behaviour and the essential roles they play for healthy functioning of our cells have been discussed. We have discussed about DNA lesions, types of DNA lesions and some mechanisms used by damage sensing and repairing proteins to correct these lesions and protect cells against diseases such as cancer. We then go on to discuss more about types of kinases, and specifically in detail about Src family of non-receptor tyrosine kinases that regulate numerous signal transduction pathways for cellular processes. We discuss about some important structural motifs present in the Src family of

kinases and how their relative orientation and arrangement deterministically defines the kinase's ability to upregulate or downregulate various cellular pathways. We also discuss about kinase inhibitors, types of kinase inhibitors and their stable binding geometries with different kinase conformations.

Next, we introduce scientific concepts like potential energy surfaces, force fields, energy minimisation techniques on potential energy surfaces and minimum energy paths. We then get into the world of molecular dynamics by introducing MD equations and how they are solved and integrated at each step to evolve the system. We then introduce computational sampling methods, umbrella sampling and nudged elastic band algorithm, and the partial NEB implementation of nudged elastic band algorithm in Amber MD to find minimum energy paths. We present the algorithm's results for our test system, conformational transitions in alanine-dipeptide, and explain how we are going to apply this method to find minimum energy path for conformational transitions in large complex systems, like c-Src kinase.

Further, we investigate the molecular process of association of the DNA-damage sensing protein Rad4 with different sequence mismatched DNA lesions. We study the mechanism, energetics, dynamics, and molecular basis of the association of Rad4/XPC enzyme with three different base-pair sequence mismatched DNA lesions, namely, TTT/TTT, TAT/TAT and CCC/CCC. We define appropriate collective variable to study the mechanism of association and use umbrella sampling simulations to generate free energy surfaces for the three Rad4-DNA complexes. We identify three regimes in each of the free energy profiles, where different interactions between Rad4-DNA, DNA-water and Rad4-water dominate, depending upon the value of the association collective variable ξ. Comparing the local minimum from the free energy profiles for the sequences, we observe Rad4 recognising and docking at the CCC/CCC mismatched site earlier than the other two sequences. This hints towards higher sequence specificity Rad4 exhibits towards CCC/CCC mismatch in comparison to the other two mismatches. We also characterise the number of hydrogen bonds and quantify interaction energy between Rad4-DNA complexes for the three sequences. Free energy surfaces obtained from the unbiased MD runs of the three sequences were also calculated to further validate our results.

Lastly, we investigate the dynamics of drug-induced conformational transitions and the activation/deactivation mechanisms of human c-Src kinase as a type-1 inhibitor bosutinib binds/unbinds from the ATP binding pocket. We employ a popular path finding technique, nudged elastic band method, to find the minimum energy path connecting the active and inactive conformational states of human c-Src kinase, as the drug binds/unbinds from the binding pocket. The dynamics of these drug-induced conformational transitions gives us key insights into the mechanism of binding/unbinding. We highlight the crucial role of hydration/dehydration of the binding pocket, and its coupling with key kinase residues in the binding pocket and distance of drug from the binding pocket. We also identify a conserved entry/exit path water needs to take, created via distortion of the evolutionary Lys295-Glu310 salt bridge, to effectively contribute towards the dissociation of the drug. Interaction energies between

drug and various kinase residues in or near the binding pocket were calculated and analysed to further strengthen our claims.

Future Work

For studying Rad4-damaged DNA association process more efficiently, future works would employ better and more suitable collective variables to capture the association of the Rad4 enzyme to damaged DNA site more appropriately. This would involve considering the angle of approach with which the enzyme approaches the damaged site. We also understand the association of Rad4 with the DNA damaged site and insertion of BDH3 β-hairpin inside the damaged site are interdependent processes and future studies should also consider coupling between these events to make more realistic observations. Finally, other base modifications DNA damages such as CPD lesion, Methylation, Deamination, Oxidative damage, etc should also be studied in the future with appropriate force field parameters.

The current study of finding minimum energy paths of drug-induced conformational transitions in kinases does not take kinetic energy into account. The actual mechanics of these conformational transitions therefore might differ than what we propose here. To study the actual mechanism of conformational transition, future works would look to obtain free energy profiles along the MEP, as order parameter. This can be done by simulating each of the images at the desired temperature, and collecting the order parameter data over the simulations performed. Future works would also employ similar techniques to find inhibitor induced conformational transitions for type-2 and type-3 inhibitors, whose unbinding dynamics would be more tricky, since they occupy deep cavities and allosteric sites in the kinase domain. Deeper understanding of these allosteric sites on the protein surface and their impact on drug binding/unbinding dynamics would also be intriguing to theoretical researchers and experimentalists in the future.

Related Publications

1. Abhinandan Panigrahi, Hemanth Vemuri, Madhur Aggarwal, Kartheek Pitta, Marimuthu Krishnan, "Sequence specificity, energetics and mechanism of mismatch recognition by DNA damage sensing protein Rad4/XPC", Nucleic Acids Research, Volume 48, Issue 5, 18 March 2020, Pages 2246–2257, https://doi.org/10.1093/nar/gkaa078

2. Madhur Aggarwal, Aadarsh Raghunathan, Shaunak Badani, Marimuthu Krishnan, "Minimum Energy Path Analysis of Drug-induced Inhibition of c-Src Kinase", Dec 2022 (Manuscript Under Preparation)

Bibliography

[1] Paul J Sapienza and Andrew L Lee. Using nmr to study fast dynamics in proteins: methods and applications. *Current Opinion in Pharmacology*, 10(6):723–730, 2010. ISSN 1471-4892. doi: https://doi.org/10.1016/j.coph.2010.09.006. URL `https://www.sciencedirect.com/science/article/pii/S147148921000144X`. Endocrine and metabolic diseases/New technologies - the importance of protein dynamics.

[2] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, Jun 1977. ISSN 1476-4687. doi: 10.1038/267585a0. URL `https://doi.org/10.1038/267585a0`.

[3] Katherine A. Henzler-Wildman, Ming Lei, Vu Thai, S. Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, Dec 2007. ISSN 1476-4687. doi: 10.1038/nature06407. URL `https://doi.org/10.1038/nature06407`.

[4] Martin Karplus and J. Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652, Sep 2002. ISSN 1545-9985. doi: 10.1038/nsb0902-646. URL `https://doi.org/10.1038/nsb0902-646`.

[5] Christopher A Hunter. Sequence-dependent dna structure: the role of base stacking interactions. *J. Mol. Biol.*, 230(3):1025–1054, 1993.

[6] Myron F Goodman. On the wagon—dna polymerase joins" h-bonds anonymous". *Nat. Biotechnol.*, 17(7):640, 1999.

[7] Leslie Pray. Discovery of dna structure and function: Watson and crick. *Nat. Education*, 1(1):100, 2008.

[8] Joshua Lederberg. The transformation of genetics by dna: an anniversary celebration of avery, macleod and mccarty (1944). *Genetics*, 136(2):423, 1994.

[9] Boris Magasanik, Ernst Vischer, Ruth Doniger, David Elson, and Erwin Chargaff. The separation and estimation of ribonucleotides in minute quantities. *J. Biol. Chem.*, 186(1):37–50, 1950.

[10] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171 (4356):737–738, 1953.

[11] Soraya de Chadarevian. Portrait of a discovery: Watson, crick, and the double helix. *Isis*, 94(1): 90–105, 2003.

[12] Martin Chaplin. Do we underestimate the importance of water in cell biology? *Nat. Rev. Mol. Cell Biol.*, 7(11):861, 2006.

[13] John J. DiGiovanna and Kenneth H. Kraemer. Shining a light on xeroderma pigmentosum. *J. Invest. Dermatol.*, 132(3):785–796, 2012. ISSN 0022-202X.

[14] K. H. Kraemer. Sunlight and skin cancer: Another link revealed. *Proc. Natl. Acad. Sci. U.S.A.*, 94(1):11–14, 1997. ISSN 0027-8424, 1091-6490.

[15] Kum Kum Khanna and Stephen P Jackson. Dna double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet.*, 27(3):247, 2001.

[16] Stephen P Jackson. Sensing and repairing dna double-strand breaks. *Carcinogenesis*, 23(5): 687–696, 2002.

[17] MG Marinus. Dna mismatch repair. *EcoSal Plus*, 5(1), 2012.

[18] Guo-Min Li. Mechanisms and functions of dna mismatch repair. *Cell Res.*, 18(1):85, 2008.

[19] Joanne J Dobbins. Prescott's microbiology. *J. Microbiol. Biol. Educ.*, 11(1):64, 2010.

[20] Joyce T Reardon and Aziz Sancar. Purification and characterization of escherichia coli and human nucleotide excision repair enzyme systems. *Methods Enzymol.*, 408:189–213, 2006.

[21] Michael R Lieber. The mechanism of double-strand dna break repair by the nonhomologous dna end-joining pathway. *Annu. Rev. Biochem.*, 79:181–211, 2010.

[22] Peter Blume-Jensen and Tony Hunter. Oncogenic kinase signalling. *Nature*, 411(6835):355–365, May 2001. ISSN 1476-4687. doi: 10.1038/35077225. URL https://doi.org/10.1038/35077225.

[23] Martin E M Noble, Jane A Endicott, and Louise N Johnson. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303(5665):1800–1805, March 2004.

[24] Jianming Zhang, Priscilla L. Yang, and Nathanael S. Gray. Targeting cancer with small molecule kinase inhibitors. *Nature Reviews Cancer*, 9(1):28–39, Jan 2009. ISSN 1474-1768. doi: 10.1038/nrc2559. URL https://doi.org/10.1038/nrc2559.

[25] Robert Roskoski. Targeting erk1/2 protein-serine/threonine kinases in human cancers. *Pharmacological Research*, 142:151–168, 2019. ISSN 1043-6618. doi: https://doi.org/10.1016/j.phrs.2019.01.039. URL `https://www.sciencedirect.com/science/article/pii/S1043661819300866`.

[26] Robert Roskoski. Cyclin-dependent protein serine/threonine kinase inhibitors as anticancer drugs. *Pharmacological Research*, 139:471–488, 2019. ISSN 1043-6618. doi: https://doi.org/10.1016/j.phrs.2018.11.035. URL `https://www.sciencedirect.com/science/article/pii/S1043661818318784`.

[27] Eugene S. Kandel and Nissim Hay. The regulation and activities of the multifunctional serine/threonine kinase akt/pkb. *Experimental Cell Research*, 253(1):210–229, 1999. ISSN 0014-4827. doi: https://doi.org/10.1006/excr.1999.4690. URL `https://www.sciencedirect.com/science/article/pii/S0014482799946906`.

[28] Robert Roskoski. Small molecule inhibitors targeting the egfr/erbb family of protein-tyrosine kinases in human cancers. *Pharmacological Research*, 139:395–411, 2019. ISSN 1043-6618. doi: https://doi.org/10.1016/j.phrs.2018.11.014. URL `https://www.sciencedirect.com/science/article/pii/S104366181831747X`.

[29] Toshimitsu Yamaoka, Sojiro Kusumoto, Koichi Ando, Motoi Ohba, and Tohru Ohmori. Receptor tyrosine kinase-targeted cancer therapy. *International Journal of Molecular Sciences*, 19 (11), 2018. ISSN 1422-0067. doi: 10.3390/ijms19113491. URL `https://www.mdpi.com/1422-0067/19/11/3491`.

[30] Charles Pottier, Margaux Fresnais, Marie Gilon, Guy Jérusalem, Rémi Longuespée, and Nor Eddine Sounni. Tyrosine kinase inhibitors in cancer: Breakthrough and challenges of targeted therapy. *Cancers*, 12(3), 2020. ISSN 2072-6694. doi: 10.3390/cancers12030731. URL `https://www.mdpi.com/2072-6694/12/3/731`.

[31] J E Smart, H Oppermann, A P Czernilofsky, A F Purchio, R L Erikson, and J M Bishop. Characterization of sites for tyrosine phosphorylation in the transforming protein of rous sarcoma virus (pp60v-src) and its normal cellular homologue (pp60c-src). *Proceedings of the National Academy of Sciences*, 78(10):6013–6017, 1981. ISSN 0027-8424. doi: 10.1073/pnas.78.10.6013. URL `https://www.pnas.org/content/78/10/6013`.

[32] T E Kmiecik, P J Johnson, and D Shalloway. Regulation by the autophosphorylation site in overexpressed pp60c-src. *Molecular and Cellular Biology*, 8(10):4541–4546, 1988. ISSN 0270-7306. doi: 10.1128/MCB.8.10.4541. URL `https://mcb.asm.org/content/8/10/4541`.

[33] Renee J. Boerner, Daniel B. Kassel, Sean C. Barker, Byron Ellis, Pam DeLacy, and Wilson B. Knight. Correlation of the phosphorylation states of pp60c-src with tyrosine kinase activity: The intramolecular py530-sh2 complex retains significant activity if y419 is phosphorylated. *Biochemistry*, 35(29):9519–9525, Jan 1996. ISSN 0006-2960. doi: 10.1021/bi960248u. URL `https://doi.org/10.1021/bi960248u`.

[34] JA Cooper, KL Gould, CA Cartwright, and T Hunter. Tyr527 is phosphorylated in pp60c-src: implications for regulation. *Science*, 231(4744):1431–1434, 1986. ISSN 0036-8075. doi: 10.1126/science.2420005. URL `https://science.sciencemag.org/content/231/4744/1431`.

[35] M Okada and H Nakagawa. A protein tyrosine kinase involved in regulation of pp60c-src function. *Journal of Biological Chemistry*, 264(35):20886–93, 1989. URL `http://www.jbc.org/content/264/35/20886.abstract`.

[36] J. R. Engen, T. E. Wales, J. M. Hochrein, M. A. Meyn, S. Banu Ozkan, I. Bahar, and T. E. Smithgall. Structure and dynamic regulation of src-family kinases. *Cellular and Molecular Life Sciences*, 65(19):3058–3073, Oct 2008. ISSN 1420-9071. doi: 10.1007/s00018-008-8122-2. URL `https://doi.org/10.1007/s00018-008-8122-2`.

[37] Justin M. Summy and Gary E. Gallick. Src family kinases in tumor progression and metastasis. *Cancer and Metastasis Reviews*, 22(4):337–358, Dec 2003. ISSN 1573-7233. doi: 10.1023/A:1023772912750. URL `https://doi.org/10.1023/A:1023772912750`.

[38] Paul A. Bromann, Hasan Korkaya, and Sara A. Courtneidge. The interplay between src family kinases and receptor tyrosine kinases. *Oncogene*, 23(48):7957–7968, Oct 2004. ISSN 1476-5594. doi: 10.1038/sj.onc.1208079. URL `https://doi.org/10.1038/sj.onc.1208079`.

[39] Titus J. Boggon and Michael J. Eck. Structure and regulation of src family kinases. *Oncogene*, 23(48):7918–7927, Oct 2004. ISSN 1476-5594. doi: 10.1038/sj.onc.1208081. URL `https://doi.org/10.1038/sj.onc.1208081`.

[40] Sarah J. Parsons and J. Thomas Parsons. Src family kinases, key regulators of signal transduction. *Oncogene*, 23(48):7906–7909, Oct 2004. ISSN 1476-5594. doi: 10.1038/sj.onc.1208160. URL `https://doi.org/10.1038/sj.onc.1208160`.

[41] Clifford A Lowell. Src-family kinases: rheostats of immune cell signaling. *Molecular Immunology*, 41(6):631–643, 2004. ISSN 0161-5890. doi: https://doi.org/10.1016/j.molimm.2004.04.010. URL `https://www.sciencedirect.com/science/article/pii/S0161589004001269`. Cell Signaling in the Immune System.

[42] Bruce J. Mayer, Michinari Hamaguchi, and Hidesaburo Hanafusa. A novel viral oncogene with structural similarity to phospholipase c. *Nature*, 332(6161):272–275, 1988. ISSN 1476-4687. doi: 10.1038/332272a0. URL `https://doi.org/10.1038/332272a0`.

[43] I Sadowski, J C Stone, and T Pawson. A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of fujinami sarcoma virus p130gag-fps. *Molecular and Cellular Biology*, 6(12):4396–4408, 1986. ISSN 0270-7306. doi: 10.1128/MCB.6.12.4396. URL `https://mcb.asm.org/content/6/12/4396`.

[44] Megan T. Brown and Jonathan A. Cooper. Regulation, substrates and functions of src. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1287(2):121 – 149, 1996. ISSN 0304-419X. doi: https://doi.org/10.1016/0304-419X(96)00003-0. URL `http://www.sciencedirect.com/science/article/pii/0304419X96000030`.

[45] Heng Wu. *Protein Conformational Transitions Using Computational Methods*. PhD thesis, The Purdue University Graduate School, Department of Medicinal Chemistry and Molecular Pharmacology, West Lafayette, Indiana, December 2018.

[46] Sonya M. Hanson, George Georghiou, Manish K. Thakur, W. Todd Miller, Joshua S. Rest, John D. Chodera, and Markus A. Seeliger. What makes a kinase promiscuous for inhibitors? *Cell Chemical Biology*, 26(3):390–399.e5, 2019. ISSN 2451-9456. doi: https://doi.org/10.1016/j.chembiol.2018.11.005. URL `https://www.sciencedirect.com/science/article/pii/S2451945618304124`.

[47] Peng Wu, Thomas E. Nielsen, and Mads H. Clausen. Small-molecule kinase inhibitors: an analysis of fda-approved drugs. *Drug Discovery Today*, 21(1):5–10, 2016. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2015.07.008. URL `https://www.sciencedirect.com/science/article/pii/S1359644615002792`.

[48] Doriano Fabbro. 25 years of small molecular weight kinase inhibitors: Potentials and limitations. *Molecular Pharmacology*, 87(5):766–775, 2015. ISSN 0026-895X. doi: 10.1124/mol.114.095489. URL `https://molpharm.aspetjournals.org/content/87/5/766`.

[49] Basis for resistance to imatinib in 16 bcr-abl mutants as determined using molecular dynamics. *Recent Patents on Anti-Cancer Drug Discovery*, 4(2), 2009.

[50] Louis J. Lombardo, Francis Y. Lee, Ping Chen, Derek Norris, Joel C. Barrish, Kamelia Behnia, Stephen Castaneda, Lyndon A. M. Cornelius, Jagabandhu Das, Arthur M. Doweyko, Craig Fairchild, John T. Hunt, Ivan Inigo, Kathy Johnston, Amrita Kamath, David Kan, Herbert Klei, Punit Marathe, Suhong Pang, Russell Peterson, Sidney Pitt, Gary L. Schieven, Robert J.

Schmidt, John Tokarski, Mei-Li Wen, John Wityak, and Robert M. Borzilleri. Discovery of n-(2-chloro-6-methyl- phenyl)-2-(6-(4-(2-hydroxyethyl)- piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide (bms-354825), a dual src/abl kinase inhibitor with potent antitumor activity in preclinical assays. *Journal of Medicinal Chemistry*, 47(27):6658–6661, Dec 2004. ISSN 0022-2623. doi: 10.1021/jm049486a. URL `https://doi.org/10.1021/jm049486a`.

[51] Richard A. Norman, Dorin Toader, and Andrew D. Ferguson. Structural approaches to obtain kinase selectivity. *Trends in Pharmacological Sciences*, 33(5):273–278, 2012. ISSN 0165-6147. doi: https://doi.org/10.1016/j.tips.2012.03.005. URL `https://www.sciencedirect.com/science/article/pii/S0165614712000375`.

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

[53] Philip M Morse. Diatomic molecules according to the wave mechanics. ii. vibrational levels. *Phys. Rev.*, 34(1):57, 1929.

[54] Katarina Roos, Chuanjie Wu, Wolfgang Damm, Mark Reboul, James M Stevenson, Chao Lu, Markus K Dahlgren, Sayan Mondal, Wei Chen, Lingle Wang, et al. Opls3e: Extending force field coverage for drug-like small molecules. *J. Chem. Theory Comput.*, 15(3):1863–1874, 2019.

[55] 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.

[56] 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.

[57] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell Jr. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1):71, 2017.

[58] Paul Robustelli, Stefano Piana, and David E Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.*, 115(21):E4758–E4766, 2018.

[59] Geoffrey C Maitland, Maurice Rigby, E Brian Smith, and William A Wakeham. *Intermolecular forces: their origin and determination*, volume 3. Clarendon Press Oxford, 1981.

[60] Anthony J Stone. *The theory of intermolecular forces*. Oxford, 1996.

[61] M Sprik. Effective pair potentials and beyond. In *Computer simulation in chemical physics*, pages 211–259. Springer, 1993.

[62] Rodrigo Galindo-Murillo, James C Robertson, Marie Zgarbova, Jiri Sponer, Michal Otyepka, Petr Jurečka, and Thomas E Cheatham III. Assessing the current state of amber force field modifications for dna. *J. Chem. Theory Comput.*, 12(8):4114–4127, 2016.

[63] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P Eastwood, Ron O Dror, and David E Shaw. Systematic validation of protein force fields against experimental data. *PLoS One*, 7(2):e32131, 2012.

[64] J. E. Jones. On the determination of molecular fields. II. from the equation of state of a gas. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 106(738): 463–477, October 1924.

[65] Michael P Allen et al. Introduction to molecular dynamics simulation. *Comput. Soft Matter*, 23: 1–28, 2004.

[66] MJ Abraham, D van der Spoel, E Lindahl, B Hess, and The GROMACS development team. GROMACS User Manual, version 2019, 2019. URL http://www.gromacs.org.

[67] Magnus Rudolph Hestenes and Eduard Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

[68] Andrew R. Leach. *Molecular modelling: principles and applications*. Prentice Hall, 2001.

[69] Martin Field. *A Practical Introduction to the Simulation of Molecular Systems*. Cambridge University Press, 1999.

[70] Smit B. Frenkel D. *Understanding molecular simulation: from algorithms to applications*, volume 1. Academic press, 2001.

[71] Tildesley D. J. Allen M. P. *Computer simulation of liquids*. Oxford university press, 1989.

[72] 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.

[73] Johannes Kästner. Umbrella sampling. *WIREs Comput. Mol. Sci.*, 1(6):932–942, 2011.

[74] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.*, 99(20):12562–12566, 2002.

[75] 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.

[76] 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.

[77] Johannes Kästner and Walter Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method:"umbrella integration". *J. Chem. Phys.*, 123 (14):144104, 2005.

[78] 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.

[79] 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.

[80] Daniel Sheppard, Rye Terrell, and Graeme Henkelman. Optimization methods for finding minimum energy paths. *The Journal of Chemical Physics*, 128(13):134106, April 2008. doi: 10.1063/1.2841941. URL https://doi.org/10.1063/1.2841941.

[81] Graeme Henkelman, Gísli Jóhannesson, and Hannes Jónsson. *Methods for Finding Saddle Points and Minimum Energy Paths*, pages 269–302. Springer Netherlands, Dordrecht, 2002. ISBN 978-0-306-46949-7. doi: 10.1007/0-306-46949-9_10. URL https://doi.org/10.1007/0-306-46949-9_10.

[82] Peng Tao, Milan Hodošček, Joseph D. Larkin, Yihan Shao, and Bernard R. Brooks. Comparison of three chain-of-states methods: Nudged elastic band and replica path with restraints or constraints. *Journal of Chemical Theory and Computation*, 8(12):5035–5051, 2012. doi: 10.1021/ct3006248. URL https://doi.org/10.1021/ct3006248. PMID: 23526888.

[83] Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113(22):9901–9904, 2000. doi: 10.1063/1.1329672. URL https://doi.org/10.1063/1.1329672.

[84] Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of Chemical Physics*, 113(22):9978–9985, 2000. doi: 10.1063/1.1323224. URL https://doi.org/10.1063/1.1323224.

[85] Daniel Sheppard, Rye Terrell, and Graeme Henkelman. Optimization methods for finding minimum energy paths. *The Journal of Chemical Physics*, 128(13):134106, 2008. doi: 10.1063/1.2841941. URL `https://doi.org/10.1063/1.2841941`.

[86] Søren Smidstrup, Andreas Pedersen, Kurt Stokbro, and Hannes Jónsson. Improved initial guess for minimum energy path calculations. *The Journal of Chemical Physics*, 140(21):214106, 2014. doi: 10.1063/1.4878664. URL `https://doi.org/10.1063/1.4878664`.

[87] HANNES JÓNSSON, GREG MILLS, and KARSTEN W. JACOBSEN. *Nudged elastic band method for finding minimum energy paths of transitions*, pages 385–404. doi: 10.1142/9789812839664_0016. URL `https://www.worldscientific.com/doi/abs/10.1142/9789812839664_0016`.

[88] Sudipta Kundu, Satadeep Bhattacharjee, Seung-Cheol Lee, and Manish Jain. Pasta: Python algorithms for searching transition states. *COMPUTER PHYSICS COMMUNICATIONS*, 233: 261–268, DEC 2018. ISSN 0010-4655. doi: 10.1016/j.cpc.2018.06.026.

[89] Gregory Mills, Hannes Jónsson, and Gregory K. Schenter. Reversible work transition state theory: application to dissociative adsorption of hydrogen. *Surface Science*, 324(2):305–337, 1995. ISSN 0039-6028. doi: https://doi.org/10.1016/0039-6028(94)00731-4. URL `https://www.sciencedirect.com/science/article/pii/0039602894007314`.

[90] Aiichiro Nakano. A space–time-ensemble parallel nudged elastic band algorithm for molecular kinetics simulation. *Computer Physics Communications*, 178(4):280–289, 2008. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2007.09.011. URL `https://www.sciencedirect.com/science/article/pii/S0010465507004225`.

[91] Esben L. Kolsbjerg, Michael N. Groves, and Bjørk Hammer. An automated nudged elastic band method. *The Journal of Chemical Physics*, 145(9):094107, 2016. doi: 10.1063/1.4961868. URL `https://aip.scitation.org/doi/abs/10.1063/1.4961868`.

[92] Sudipta Kundu, Satadeep Bhattacharjee, Seung-Cheol Lee, and Manish Jain. PASTA: Python algorithms for searching transition stAtes. *Computer Physics Communications*, 233:261–268, December 2018. doi: 10.1016/j.cpc.2018.06.026. URL `https://doi.org/10.1016/j.cpc.2018.06.026`.

[93] Delaram Ghoreishi, David S. Cerutti, Zachary Fallon, Carlos Simmerling, and Adrian E. Roitberg. Fast implementation of the nudged elastic band method in amber. *Journal of Chemical Theory and Computation*, 15(8):4699–4707, Aug 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.9b00329. URL `https://doi.org/10.1021/acs.jctc.9b00329`.

[94] David H. Mathews and David A. Case. Nudged elastic band calculation of minimal energy paths for the conformational change of a gg non-canonical pair. *Journal of Molecular Biology*, 357 (5):1683 – 1693, 2006. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2006.01.054. URL `http://www.sciencedirect.com/science/article/pii/S0022283606000842`.

[95] Christina Bergonzo, Arthur J. Campbell, Ross C. Walker, and Carlos Simmerling. A partial nudged elastic band implementation for use with large or explicitly solvated systems. *International Journal of Quantum Chemistry*, 109(15):3781–3790, 2009. doi: 10.1002/qua.22405. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.22405`.

[96] Michael SY Huen, Shirley MH Sy, and Junjie Chen. BRCA1 and its toolbox for the maintenance of genome integrity. *Nat. Rev. Mol. Cell Biol.*, 11(2):138–148, 2009. ISSN 1471-0072, 1471-0080.

[97] Yosef Shiloh. ATM and related protein kinases: Safeguarding genome integrity. *Nat. Rev. Cancer*, 3(3):155–168, 2003. ISSN 1474-175X, 1474-1768.

[98] Karlene A. Cimprich and David Cortez. ATR: An essential regulator of genome integrity. *Nat. Rev. Mol. Cell Biol.*, 9(8):616–627, 2008. ISSN 1471-0072, 1471-0080.

[99] James E. Cleaver, Larry H. Thompson, Audrey S. Richardson, and J. Christopher States. A summary of mutations in the UV-sensitive disorders: Xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy. *Hum. Mutat.*, 14(1):9, 1999. ISSN 1059-7794, 1098-1004. URL `https://doi.org/10.1002/(sici)1098-1004(1999)14:1⟨9::aid-humu2⟩3.3.co;2-y`.

[100] Jiri Lukas, Claudia Lukas, and Jiri Bartek. More than just a focus: The chromatin response to DNA damage and its role in genome integrity maintenance. *Nat. Cell Biol.*, 13(10):1161–1169, 2011. ISSN 1465-7392, 1476-4679.

[101] Nicholas E Geacintov and Suse Broyde. Introduction and perspectives on the chemistry and biology of DNA damage. *Chem. Biol. DNA Damage*, pages 1–20, 2010.

[102] Gregory L. Verdine and Steven D. Bruner. How do DNA repair proteins locate damaged bases in the genome? *Chem. Biol.*, 4(5):329–334, 1997. ISSN 1074-5521.

[103] Anjum Ansari and Serguei V. Kuznetsov. Dynamics and mechanism of DNA-bending proteins in binding site recognition. In *Biological and Medical Physics, Biomedical Engineering*, pages 107–142. Springer New York, 2010. ISBN 9780387928074, 9780387928081. URL `https://doi.org/10.1007/978-0-387-92808-1_6`.

[104] Manas Kumar Sarangi, Viktoriya Zvoda, Molly Nelson Holte, Nicole A Becker, Justin P Peters, L James Maher, and Anjum Ansari. Evidence for a bind-then-bend mechanism for architectural

DNA binding protein yNhp6A. *Nucleic Acids Res.*, 47(6):2871–2883, 2019. ISSN 0305-1048, 1362-4962.

[105] Yogambigai Velmurugu, Paula Vivas, Mitchell Connolly, Serguei V Kuznetsov, Phoebe A Rice, and Anjum Ansari. Two-step interrogation then recognition of DNA binding site by integration host factor: An architectural DNA-bending protein. *Nucleic Acids Res.*, 46(4):1741–1755, 2017. ISSN 0305-1048, 1362-4962.

[106] Alberto Ciccia and Stephen J. Elledge. The DNA damage response: Making it safe to play with knives. *Mol. Cell*, 40(2):179–204, 2010. ISSN 1097-2765.

[107] S. E. Polo and S. P. Jackson. Dynamics of DNA damage response proteins at DNA breaks: A focus on protein modifications. *Genes Dev.*, 25(5):409–433, 2011. ISSN 0890-9369.

[108] Lei Jia, Konstantin Kropachev, Shuang Ding, Bennett Van Houten, Nicholas E. Geacintov, and Suse Broyde. Exploring damage recognition models in prokaryotic nucleotide excision repair with a benzo[a]pyrene-derived lesion in UvrB. *Biochemistry (Mosc.)*, 48(38):8948–8957, 2009. ISSN 0006-2960, 1520-4995.

[109] Ronjon K. Chakraverty and Ian D. Hickson. Defending genome integrity during DNA replication: A proposed role for RecQ family helicases. *Bioessays*, 21(4):286–294, 1999. ISSN 0265-9247. URL `https://doi.org/10.1002/(sici)1521-1878(199904)21:4⟨286::aid-bies4⟩3.0.co;2-z`.

[110] Yuan-Cho Lee, Yuqin Cai, Hong Mu, Suse Broyde, Shantu Amin, Xuejing Chen, Jung-Hyun Min, and Nicholas E. Geacintov. The relationships between XPC binding to conformationally diverse DNA adducts and their excision by the human NER system: Is there a correlation? *DNA Repair*, 19:55–63, 2014. ISSN 1568-7864.

[111] Hong Mu, Nicholas E Geacintov, Suse Broyde, Jung-Eun Yeo, and Orlando D Schärer. Molecular basis for damage recognition and verification by XPC-RAD23B and TFIIH in nucleotide excision repair. *DNA Repair*, 71:33–42, 2018. ISSN 1568-7864.

[112] Kazantsev A Sancar A Huang JC, Hsu DS. Substrate spectrum of human excinuclease: repair of abasic sites, methylated bases, mismatches, and bulky adducts. *Proceedings of the National Academy of Sciences of the United States of America*, 91(25):12213–12217, 1994. doi: 10.1073/pnas.91.25.12213. URL `https://doi.org/10.1073/pnas.91.25.12213`.

[113] Elisa T. Zhang, Yuan He, Patricia Grob, Yick W. Fong, Eva Nogales, and Robert Tjian. Architecture of the human XPC DNA repair and stem cell coactivator complex. *Proc. Natl. Acad. Sci. U.S.A.*, 112(48):14817–14822, 2015. ISSN 0027-8424, 1091-6490.

[114] Jung-Hyun Min and Nikola P. Pavletich. Recognition of DNA damage by the rad4 nucleotide excision repair protein. *Nature*, 449(7162):570–575, 2007. ISSN 0028-0836, 1476-4687.

[115] Xuejing Chen, Yogambigai Velmurugu, Guanqun Zheng, Beomseok Park, Yoonjung Shim, Youngchang Kim, Lili Liu, Bennett Van Houten, Chuan He, Anjum Ansari, and Jung-Hyun Min. Kinetic gating mechanism of DNA damage recognition by Rad4/XPC. *Nat. Commun.*, 6(1):5849, 2015. ISSN 2041-1723.

[116] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996. ISSN 0263-7855.

[117] Yang Liu, Dara Reeves, Konstantin Kropachev, Yuqin Cai, Shuang Ding, Marina Kolbanovskiy, Alexander Kolbanovskiy, Judith L. Bolton, Suse Broyde, Bennett Van Houten, and Nicholas E. Geacintov. Probing for DNA damage with β-hairpins: Similarities in incision efficiencies of bulky DNA adducts by prokaryotic and human nucleotide excision repair systems in vitro. *DNA Repair*, 10(7):684–696, 2011. ISSN 1568-7864.

[118] Hong Mu, Nicholas E. Geacintov, Yingkai Zhang, and Suse Broyde. Recognition of damaged DNA for nucleotide excision repair: A correlated motion mechanism with a mismatched cis-syn thymine dimer lesion. *Biochemistry (Mosc.)*, 54(34):5263–5267, 2015. ISSN 0006-2960, 1520-4995.

[119] Hong Mu, Nicholas E Geacintov, Jung-Hyun Min, Yingkai Zhang, and Suse Broyde. Nucleotide excision repair lesion-recognition protein rad4 captures a pre-flipped partner base in a benzo [a] pyrene-derived DNA lesion: How structure impacts the binding pathway. *Chem. Res. Toxicol.*, 30(6):1344–1354, 2017.

[120] David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5(11):789–796, 2009. ISSN 1552-4450, 1552-4469.

[121] Chunyang Cao, Yu Lin Jiang, James T Stivers, and Fenhong Song. Dynamic opening of DNA during the enzymatic search for a damaged base. *Nat. Struct. Mol. Biol*, 11(12):1230–1236, 2004. ISSN 1545-9993, 1545-9985.

[122] 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. *Proc. Natl. Acad. Sci. U.S.A.*, 113(16):E2296–E2305, 2016. ISSN 0027-8424, 1091-6490.

[123] Debamita Paul, Hong Mu, Hong Zhao, Ouathek Ouerfelli, Philip D Jeffrey, Suse Broyde, and Jung-Hyun Min. Structure and mechanism of pyrimidine–pyrimidone (6-4) photoproduct recognition by the Rad4/XPC nucleotide excision repair complex. *Nucleic Acids Res.*, 47(12):6015–6028, 2019. ISSN 0305-1048.

[124] Hong Mu, Yingkai Zhang, Nicholas E. Geacintov, and Suse Broyde. Lesion sensing during initial binding by yeast XPC/Rad4: Toward predicting resistance to nucleotide excision repair. *Chem. Res. Toxicol.*, 31(11):1260–1268, 2018. ISSN 0893-228X, 1520-5010.

[125] Sagnik Chakraborty, Peter J Steinbach, Debamita Paul, Hong Mu, Suse Broyde, Jung-Hyun Min, and Anjum Ansari. Enhanced spontaneous DNA twisting/bending fluctuations unveiled by fluorescence lifetime distributions promote mismatch recognition by the rad4 nucleotide excision repair complex. *Nucleic Acids Res.*, 46(3):1240–1255, 2017. ISSN 0305-1048, 1362-4962.

[126] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977. ISSN 0021-9991. doi: https://doi.org/10.1016/0021-9991(77)90121-8. URL https://www.sciencedirect.com/science/article/pii/0021999177901218.

[127] 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. *J. Comput. Chem.*, 25(13):1605–1612, 2004.

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

[129] 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.

[130] Daniel J. Price and Charles L. Brooks. A modified TIP3P water potential for simulation with ewald summation. *J. Chem. Phys.*, 121(20):10096–10103, 2004. ISSN 0021-9606, 1089-7690.

[131] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005. ISSN 0192-8651, 1096-987X.

[132] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91(1-3):43–56, 1995. ISSN 0010-4655.

[133] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, 2014. ISSN 0010-4655.

[134] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A. Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.*, 180(10):1961–1972, 2009. ISSN 0010-4655.

[135] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, September 1997. doi: 10.1002/(sici)1096-987x(199709)18:12⟨1463::aid-jcc4⟩3.0.co;2-h. URL `https://doi.org/10.1002/(sici)1096-987x(199709)18:12⟨1463::aid-jcc4⟩3.0.co;2-h`.

[136] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993. doi: 10.1063/1.464397. URL `https://doi.org/10.1063/1.464397`.

[137] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.

[138] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[139] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[140] 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. *J. Comput. Chem.*, 13(8):1011–1021, 1992. ISSN 0192-8651, 1096-987X.

[141] Michael J Eck and Paul W Manley. The interplay of structural information and functional studies in kinase drug design: insights from bcr-abl. *Current Opinion in Cell Biology*, 21(2):288–295, 2009. ISSN 0955-0674. doi: https://doi.org/10.1016/j.ceb.2009.01.014. URL `https://www.sciencedirect.com/science/article/pii/S0955067409000295`. Cell regulation.

[142] Martin E. M. Noble, Jane A. Endicott, and Louise N. Johnson. Protein kinase inhibitors: Insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004. doi: 10.1126/science.1095920. URL `https://www.science.org/doi/abs/10.1126/science.1095920`.

[143] Matthias Rabiller, Matthäus Getlik, Sabine Klüter, André Richters, Sandra Tückmantel, Jeffrey R. Simard, and Daniel Rauh. Proteus in the world of proteins: Conformational changes in protein kinases. *Archiv der Pharmazie*, 343(4):193–206, 2010. doi: https://doi.org/10.1002/ardp.201000028. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ardp.201000028`.

[144] Nicholas M. Levinson and Steven G. Boxer. A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity. *Nature Chemical Biology*, 10(2):127–132, Feb 2014. ISSN 1552-4469. doi: 10.1038/nchembio.1404. URL `https://doi.org/10.1038/nchembio.1404`.

[145] Wenqing Xu, Amish Doshi, Ming Lei, Michael J. Eck, and Stephen C. Harrison. Crystal structures of c-src reveal features of its autoinhibitory mechanism. *Molecular Cell*, 3(5):629–638, May 1999. ISSN 1097-2765. doi: 10.1016/S1097-2765(00)80356-1. URL `https://doi.org/10.1016/S1097-2765(00)80356-1`.

[146] Matthäus Getlik, Christian Grütter, Jeffrey R. Simard, Sabine Klüter, Matthias Rabiller, Haridas B. Rode, Armin Robubi, and Daniel Rauh. Hybrid compound design to overcome the gatekeeper t338m mutation in csrc. *Journal of Medicinal Chemistry*, 52(13):3915–3926, Jul 2009. ISSN 0022-2623. doi: 10.1021/jm9002928. URL `https://doi.org/10.1021/jm9002928`.

[147] Diwakar Shukla, Yilin Meng, Benoît Roux, and Vijay S. Pande. Activation pathway of src kinase reveals intermediate states as targets for drug design. *Nature Communications*, 5(1):3397, Mar 2014. ISSN 2041-1723. doi: 10.1038/ncomms4397. URL `https://doi.org/10.1038/ncomms4397`.

[148] Wenxun Gan, Sichun Yang, and Benoît Roux. Atomistic view of the conformational activation of src kinase using the string method with swarms-of-trajectories. *Biophysical Journal*, 97(4):L8–L10, 2009. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2009.06.016. URL `https://www.sciencedirect.com/science/article/pii/S0006349509011175`.

[149] Cheng-Chieh Tsai, Zhi Yue, and Jana Shen. How electrostatic coupling enables conformational plasticity in a tyrosine kinase. *Journal of the American Chemical Society*, 141(38):15092–15101, Sep 2019. ISSN 0002-7863. doi: 10.1021/jacs.9b06064. URL `https://doi.org/10.1021/jacs.9b06064`.

[150] Hao Wu, Fabian Paul, Christoph Wehmeyer, and Frank Noé. Multiensemble markov models of molecular thermodynamics and kinetics. *Proceedings of the National Academy of Sciences*, 113(23):E3221–E3230, 2016. doi: 10.1073/pnas.1525092113. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1525092113`.

[151] Pratyush Tiwary, Vittorio Limongelli, Matteo Salvalaglio, and Michele Parrinello. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the*

National Academy of Sciences, 112(5):E386–E391, 2015. doi: 10.1073/pnas.1424461112. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1424461112`.

[152] Ivan Teo, Christopher G. Mayne, Klaus Schulten, and Tony Lelièvre. Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time. *Journal of Chemical Theory and Computation*, 12(6):2983–2989, Jun 2016. ISSN 1549-9618. doi: 10.1021/acs.jctc.6b00277. URL `https://doi.org/10.1021/acs.jctc.6b00277`.

[153] Huei-Jiun Li, Cheng-Tsung Lai, Pan Pan, Weixuan Yu, Nina Liu, Gopal R. Bommineni, Miguel Garcia-Diaz, Carlos Simmerling, and Peter J. Tonge. A structural and energetic model for the slow-onset inhibition of the mycobacterium tuberculosis enoyl-acp reductase inha. *ACS Chemical Biology*, 9(4):986–993, Apr 2014. ISSN 1554-8929. doi: 10.1021/cb400896g. URL `https://doi.org/10.1021/cb400896g`.

[154] Pratyush Tiwary, Jagannath Mondal, and B. J. Berne. How and when does an anticancer drug leave its binding site? *Science Advances*, 3(5):e1700014, 2017. doi: 10.1126/sciadv.1700014. URL `https://www.science.org/doi/abs/10.1126/sciadv.1700014`.

[155] Satyabrata Bandyopadhyay, Bibhab Bandhu Majumdar, and Jagannath Mondal. Solvent's role in cavity–ligand recognition would depend on the mode of ligand diffusion. *The Journal of Physical Chemistry B*, 126(16):2952–2958, 2022. doi: 10.1021/acs.jpcb.1c09645. URL `https://doi.org/10.1021/acs.jpcb.1c09645`. PMID: 35436126.

[156] Peter Schmidtke, F. Javier Luque, James B. Murray, and Xavier Barril. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *Journal of the American Chemical Society*, 133(46):18903–18910, 2011. doi: 10.1021/ja207494u. URL `https://doi.org/10.1021/ja207494u`. PMID: 21981450.

[157] Jagannath Mondal, Richard A. Friesner, and B. J. Berne. Role of desolvation in thermodynamics and kinetics of ligand binding to a kinase. *Journal of Chemical Theory and Computation*, 10 (12):5696–5705, 2014. doi: 10.1021/ct500584n. URL `https://doi.org/10.1021/ct500584n`. PMID: 25516727.

[158] Navjeet Ahalawat, Satyabrata Bandyopadhyay, and Jagannath Mondal. On the role of solvent in hydrophobic cavity–ligand recognition kinetics. *The Journal of Chemical Physics*, 152(7): 074104, 2020. doi: 10.1063/1.5139584. URL `https://doi.org/10.1063/1.5139584`.

[159] Tom Young, Robert Abel, Byungchan Kim, Bruce J. Berne, and Richard A. Friesner. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proceedings of the National Academy of Sciences*, 104(3):808–813, 2007. doi: 10.1073/pnas.0610202104. URL `https://www.pnas.org/doi/abs/10.1073/pnas.0610202104`.

[160] Jagannath Mondal, Joseph A. Morrone, and B. J. Berne. How hydrophobic drying forces impact the kinetics of molecular recognition. *Proceedings of the National Academy of Sciences*, 110 (33):13277–13282, 2013. doi: 10.1073/pnas.1312529110. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1312529110`.

[161] Pratyush Tiwary, Jagannath Mondal, Joseph A. Morrone, and B. J. Berne. Role of water and steric constraints in the kinetics of cavity–ligand unbinding. *Proceedings of the National Academy of Sciences*, 112(39):12015–12019, 2015. doi: 10.1073/pnas.1516652112. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1516652112`.

[162] John E. Ladbury. Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry Biology*, 3(12):973–980, 1996. ISSN 1074-5521. doi: https://doi.org/10.1016/S1074-5521(96)90164-7. URL `https://www.sciencedirect.com/science/article/pii/S1074552196901647`.

[163] 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, 2008. doi: https://doi.org/10.1002/jcc.20945. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20945`.

[164] Juan C Meza. Steepest descent. *Wiley Interdiscip. Rev. Comput. Stat.*, 2(6):719–722, 2010.

[165] Cheng-Tsung Lai, Huei-Jiun Li, Weixuan Yu, Sonam Shah, Gopal R. Bommineni, Victoria Perrone, Miguel Garcia-Diaz, Peter J. Tonge, and Carlos Simmerling. Rational modulation of the induced-fit conformational change for slow-onset inhibition in mycobacterium tuberculosis inha. *Biochemistry*, 54(30):4683–4691, Aug 2015. ISSN 0006-2960. doi: 10.1021/acs.biochem.5b00284. URL `https://doi.org/10.1021/acs.biochem.5b00284`.

[166] Nikita A. Kuznetsov, Christina Bergonzo, Arthur J. Campbell, Haoquan Li, Grigory V. Mechetin, Carlos de los Santos, Arthur P. Grollman, Olga S. Fedorova, Dmitry O. Zharkov, and Carlos Simmerling. Active destabilization of base pairs by a DNA glycosylase wedge initiates damage recognition. *Nucleic Acids Research*, 43(1):272–281, 12 2014. ISSN 0305-1048. doi: 10.1093/nar/gku1300. URL `https://doi.org/10.1093/nar/gku1300`.

[167] Haoquan Li, Anton V. Endutkin, Christina Bergonzo, Arthur J. Campbell, Carlos de los Santos, Arthur Grollman, Dmitry O. Zharkov, and Carlos Simmerling. A dynamic checkpoint in oxidative lesion discrimination by formamidopyrimidine–DNA glycosylase. *Nucleic Acids Research*, 44 (2):683–694, 11 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1092. URL `https://doi.org/10.1093/nar/gkv1092`.

[168] Christina Bergonzo, Arthur J. Campbell, Carlos de los Santos, Arthur P. Grollman, and Carlos Simmerling. Energetic preference of 8-oxog eversion pathways in a dna glycosylase. *Journal of the American Chemical Society*, 133(37):14504–14506, Sep 2011. ISSN 0002-7863. doi: 10.1021/ja205142d. URL https://doi.org/10.1021/ja205142d.

[169] Lucy Fallon, Kellon A. A. Belfon, Lauren Raguette, Yuzhang Wang, Darya Stepanenko, Abbigayle Cuomo, Jose Guerra, Stephanie Budhan, Sarah Varghese, Christopher P. Corbo, Robert C. Rizzo, and Carlos Simmerling. Free energy landscapes from sars-cov-2 spike glycoprotein simulations suggest that rbd opening can be modulated via interactions in an allosteric pocket. *Journal of the American Chemical Society*, 143(30):11349–11360, Aug 2021. ISSN 0002-7863. doi: 10.1021/jacs.1c00556. URL https://doi.org/10.1021/jacs.1c00556.

[170] Haoquan Li, Anton V. Endutkin, Christina Bergonzo, Lin Fu, Arthur Grollman, Dmitry O. Zharkov, and Carlos Simmerling. Dna deformation-coupled recognition of 8-oxoguanine: Conformational kinetic gating in human dna glycosylase. *Journal of the American Chemical Society*, 139(7):2682–2692, Feb 2017. ISSN 0002-7863. doi: 10.1021/jacs.6b11433. URL https://doi.org/10.1021/jacs.6b11433.

[171] Haohao Fu, Haochuan Chen, Marharyta Blazhynska, Emma Goulard Coderc de Lacam, Florence Szczepaniak, Anna Pavlova, Xueguang Shao, James C. Gumbart, François Dehez, Benoît Roux, Wensheng Cai, and Christophe Chipot. Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations. *Nature Protocols*, 17(4): 1114–1141, Apr 2022. ISSN 1750-2799. doi: 10.1038/s41596-021-00676-1. URL https://doi.org/10.1038/s41596-021-00676-1.

[172] Angela M. Barragan and Benoit Roux. Unveiling the reaction mechanism of bruton's tyrosine kinase inhibition by ibrutinib. *Biophysical Journal*, 121(3):8a, Feb 2022. ISSN 0006-3495. doi: 10.1016/j.bpj.2021.11.2670. URL https://doi.org/10.1016/j.bpj.2021.11.2670.

[173] Trayder Thomas and Benoît Roux. Tyrosine kinases: complex molecular systems challenging computational methodologies. *The European Physical Journal B*, 94(10):203, Oct 2021. ISSN 1434-6036. doi: 10.1140/epjb/s10051-021-00207-7. URL https://doi.org/10.1140/epjb/s10051-021-00207-7.

[174] Yilin Meng, Lalima G. Ahuja, Alexandr P. Kornev, Susan S. Taylor, and Benoît Roux. A catalytically disabled double mutant of src tyrosine kinase can be stabilized into an active-like conformation. *Journal of Molecular Biology*, 430(6):881–889, 2018. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2018.01.019. URL https://www.sciencedirect.com/science/article/pii/S0022283618300457.

[175] Yilin Meng, Matthew P. Pond, and Benoît Roux. Tyrosine kinase activation and conformational flexibility: Lessons from src-family tyrosine kinases. *Accounts of Chemical Research*, 50(5): 1193–1201, 2017. doi: 10.1021/acs.accounts.7b00012. URL `https://doi.org/10.1021/acs.accounts.7b00012`. PMID: 28426203.

[176] Yilin Meng, Yen-lin Lin, and Benoît Roux. Computational study of the "dfg-flip" conformational transition in c-abl and c-src tyrosine kinases. *The Journal of Physical Chemistry B*, 119(4):1443–1456, 2015. doi: 10.1021/jp511792a. URL `https://doi.org/10.1021/jp511792a`. PMID: 25548962.

[177] Mikolai Fajer, Yilin Meng, and Benoît Roux. The activation of c-src tyrosine kinase: Conformational transition pathway and free energy landscape. *The Journal of Physical Chemistry B*, 121 (15):3352–3363, 2017. doi: 10.1021/acs.jpcb.6b08409. URL `https://doi.org/10.1021/acs.jpcb.6b08409`. PMID: 27715044.

[178] Yilin Meng, Diwakar Shukla, Vijay S. Pande, and Benoît Roux. Transition path theory analysis of c-src kinase activation. *Proceedings of the National Academy of Sciences*, 113(33):9193–9198, 2016. doi: 10.1073/pnas.1602790113. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1602790113`.