University of Alberta

Computational High Throughput Screening Targeting DNA Repair Proteins To Improve Cancer Therapy

by

Khaled H. Barakat

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Physics

Spring 2012
Edmonton, Alberta

I would like to dedicate this thesis to my parents, Hassan and Twaba, my wife, Iman, my daughters, Yomna and Nour, and my son, Hassan.

Abstract

Developing a new drug is a complex, highly structured, and expensive task. The further a potential drug progresses in the development process, the more costly its failure becomes. Virtual screening (VS) is the initial stage of a drug discovery process. Its job is to screen large compound databases for bioactive molecules. Its role is critical to reduce the probability of late-stage expensive failures. A reliable VS protocol would identify a diversity of lead compounds that are suitable for further structural optimizations. Most of the current available protocols fail at integrating target flexibility or suggesting accurate ranking for the selected top hits. Here, we introduce an improved virtual screening protocol. A protocol that improves over current methodologies by employing complementary techniques comprising molecular docking, molecular dynamics simulations, iterative clustering techniques, principle component analysis and accurate scoring methods. The implemented VS protocol identified novel compounds that can bind to a number of important cancer-related targets. The targets chosen here play critical roles in tumor cell initiation and progression and their regulation promises for the improvement of current cancer therapy. Two of these important targets are DNA repair proteins that are linked to the hallmark relapse or drug resistance phenomena. These are Excision Repair Cross-Complementation Group 1 (ERCC1), and DNA polymerase beta. The former is a key player in Nucleotide Excision Repair (NER), while the latter is the error-prone polymerase of Base Excision Repair (BER). The third target is p53, a guardian of the genome that is inactivated in more than half of all human cancers. The work presented here has an outstanding significance on both the methods and their applications. On one hand the implemented protocol is generic and can be used almost

against any target. On the other hand, the compounds we identified have the promise of being successful potential drug candidates that can progress through the drug discovery process and improve cancer therapy.

Acknowledgements

My prevalent thanks go to my mentor and friend, Jack Tuszynski. He has been always a great helper, a brilliant leader, a successful model and a wise teacher. Special thanks go to Torin Huzil, Tyler Luchko and Holly Freemann. Their guidance was deeply appreciated. My immense appreciation goes to my supervisory committee members, Michael Woodside, Mariusz Klobukowski and Sharon Morsink for their help with my research and my scholarship applications. I would like also to thank my colleagues and friends Travis Craddock, Douglas Friesen, Melissa Gajewski, Philip Winter, Jonathan Mane, Ishwar Hosamani, Gerry Leenders, and Richard Tseng. They provided me with the most sociable and friendly environment to carry on my research. The rest goes to my wife, Iman, for her constant encouragement, support, and patience.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Once, the US General, Ulyses S. Grant, summarized his philosophy on warfare in just four concise statements, “The art of war is simple enough. Find out where your enemy is. Get at him as soon as you can. Strike him as hard as you can, and keep moving”. Although these overarching statements formed the basic premise of modern war strategies, the same concepts have been applied in designing new drugs aimed at combating a broad range of diseases. In this context, rational drug design has been established as an exciting research approach aimed at developing safer and more efficacious drugs. The ultimate goal of this paradigm is to design small organic, non-peptidic, compounds that bind to a specific molecular target (typically a protein), and result in the inhibition (or less frequently, activation) of a particular protein or enzyme involved in a given cellular pathway.

Without a doubt, developing a new drug is a highly structured and expensive route that begins with the identification of the target and concludes with a phase III clinical trial followed by marketing. A candidate drug may never materialize into a safe and efficacious medicine due to its failure to comply with stringent requirements at any stage of the drug discovery process. The further a potential drug progresses in the development process, the more costly its failure becomes. Accordingly, it is important to reduce the probability of late-stage expensive failures by identifying a diversity of lead compounds that are suitable for structural optimizations. Throughout the last three decades, experimental high throughput screening (HTS) and combinatorial chemistry formed the principal source for lead identification. However, as these approaches are particularly expensive and require considerable resources in terms of equipment and skills of the highly qualified personnel, it was vital to search for an alternative or a complementary low-cost computational technique that aids in the discovery of new bioactive compounds while maintaining the rapidity of HTS. More recently, a new method was developed and named computational virtual screening (VS) or *in silico* screening.

This new method formed a “change in paradigm” and declared itself as a possible replacement for the massive HTS machines. For one aspect, the basic cost of a typical HTS laboratory includes collecting and maintaining screening libraries of thousands, if not millions, of compounds. In VS, one does not need do synthesis, or even touch, these molecules. Another important aspect is the experimental difficulties that are associated with HTS such as limited solubility or aggregate formation, which are not relevant to VS and do not need to be considered. All VS predictions are usually achieved in a single computer cluster using a systematic procedure that attempts to answer a definite question: Can these compounds bind to a particular target and, hence, induce the desired biological activity? The outcome of such procedure is a set of compounds that are predicted to bind to the target with acceptable affinities. This set can be directly tested on cells or, as in many cases, can be used to guide HTS by focusing the search on a limited collection of similar structures, instead of testing millions of compounds.

The pioneering efforts of Kuntz and DesJarlais in the late 1990s defined VS as “searching for bioactive molecules within large compound databases”. Over the past decade, the method has been vastly improved and gained popularity as a result of an exponential increase in the performance of computer hardware, methods and enhanced human expertise. Currently, VS is a valuable prototype within the rational drug design tool box, helping in prioritizing compounds for experimental HTS and aiding in compound progression through lead optimization. However, as every new born, VS is still developing, and far from forming a mature field of science. This is apparent in the fact that the number of strategies followed in the field is nearly as large as the number of reported screening campaigns.

A drug usually binds to a specific location within the target (a binding site). To be biologically active, it must physically fit within the binding site. A century ago, Fischer described this event as a lock-and-key fit. However, it is not only the shape and size of the drug that is important for binding. The drug must also complement the hydrophobic and polar parts of the binding site. Although this binding reaction seems very similar to fitting the little pieces of a jigsaw puzzle, what I described so far is, in fact, only a part of the story. There are many additional factors that must be taken into account. These include the structural flexibility of the drug and its target, solvent effects, entropy contributions and the protonation states of the two molecules.

This dissertation focuses on building an improved virtual screening protocol. A protocol that takes into account most of the factors described above and intends to search for novel compounds that can bind to a number of important cancer-related molecular targets. The targets chosen here play critical roles in tumor cell initiation and progression and their regulation promises for the improvement of current cancer therapy. Two of these important targets are DNA repair proteins that are linked to the hallmark “relapse” or “drug resistance” phenomena. These are Excision Repair Cross-Complementation Group 1 (ERCC1), and DNA polymerase beta (pol β). The former is a key player in Nucleotide Excision Repair (NER), while the latter is the error-prone polymerase of Base Excision Repair (BER). The third target is p53, a guardian of the genome that is inactivated in more than half of all human cancers.

Thusly motivated, the research question here was twofold. Is it possible to build an accurate VS algorithm that overcomes the limitations of and gaps in previous research? And if this is possible, will it be able to identify novel drug candidates that can specifically recognize and bind to the aforementioned targets? The answers to these questions are presented in the next few chapters.

This thesis is based on a set of published articles and book chapters and is organized as follows. Chapters 2 and 3 present computational background material covering the state-of-the-art of virtual screening and explain the VS protocol that was implemented and used in this work. Chapter 4 describes a two-stage-filtering application of the VS protocol on the ERCC1-XPA problem. Chapters 5 and 6 switch the gears to a different target, DNA pol β. Chapter 5 reviews all currently known pol β inhibitors and chapter 6 applies our VS protocol to its lyase activity. Chapter 7 introduces a more perplexing problem, where we identified dual inhibitors for two similar p53 binding proteins. Chapter 8 is a follow up to chapter 7's question. Chapter 8 discusses the possibility of restoring the activity of one of the most frequent occurring p53 mutations, namely, the R248Q mutant. The important results and their future impacts on cancer research are summarized in chapter 9.

The order of these chapters gives chronological details on the way the VS protocol was developed and the accumulation of knowledge that I gained throughout this research. The beginning of each chapter includes an introduction of the necessary biology, the motivation behind selecting the target, and review of previous research on its targeting. Detailed methodological parameters for each problem and part of the mathematical details of the computational methods are explained in the appendices.

Chapter 2: Virtual Screening: Solving a Jigsaw Puzzle[1]

2.1 The need for Virtual Screening

Have you tried to solve a jigsaw puzzle before? Did you ever get to a situation where one little piece is missing? This piece would complement an obvious part of the puzzle. If so, I am sure you first looked at the boundaries of the absent piece to predict its shape and size. You then looked at the colors of the adjacent pieces, the environment, to speculate its colors. Next, you scanned the remaining pieces for a one that looked similar to the one you expected. You collected parts that had fitting shapes and suitable colors. You placed them one by one in the empty space, until you finally got the one.

Well, this chapter is not, really, about solving jigsaw puzzles, however, it is about a similar process that is repeated again and again inside our cells, and in this thesis, I tried to mimic. This process is called molecular recognition.

Throughout the last three decades, experimental high throughput screening (HTS) has been the main technique playing the role of a "puzzle solver". Together with combinatorial chemistry it formed the front line for discovering new drug candidates (lead compounds). While avoiding the complexity of modeling, these methods are particularly expensive and require considerable resources. These shortcomings called for the development of an alternative/complementary and economical approach that could recognize novel bioactive compounds, while maintaining the yield and rapidity of HTS. More recently, a new method was developed that holds great promise of rapid drug candidate identification using computational methods. This new methodology has been named computational virtual screening (VS) (or *in silico* screening).

VS is still in the early stages of evolution. Many strategies are currently developing and several aspects are improving. This chapter provides support for this claim. It is divided into two main parts. The first part acts as our guide in the VS "store" and explores the state-of-the-art of this field focusing on its cutting-edge "products". It guides us inside the two main VS branches and explores their different algorithms and methods. As we will see in this part, each method by itself is not enough to build a robust and perfect VS protocol. The second part of this chapter describes in details the VS algorithm that was

[1] A version of this chapter has been published in Barakat KH, Mane JY, Tuszynski JA (2011) Virtual Screening: An Overview on Methods and Applications. In: Liu LA, Wei D, Li Y, Lei H, editors. Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications: IGI.

implemented in this thesis. Most of it was, essentially, built from the little pieces described in the first part. While the first part is a general review of the algorithms, definitions and concepts behind VS, the second part is a more detailed discussion on target selection and preparation, compounds used, and how much improvement this work added to the methodology of VS.

Once the structure of a target is available, docking algorithms can be used to place each ligand (i.e. a molecule or a molecular fragment included in a typical library of compounds) and predict its most probable binding mode (optimal target-drug complex configuration) within the binding site of the target.[1,2] Most docking programs can rank the activity of each compound by analyzing the different ligand-target interactions and estimating the binding affinity of the complex. In addition to docking techniques, one can define the essential interactions between the ligand and the binding site of the receptor and translate this information into the formulation of binding-site pharmacophore models.[3] These models can be used to search the available chemical space for compounds that can complement the physico-chemical features of the receptor (target). As these two procedures require a comprehensive understanding of the structural arrangement of the target, they have been commonly termed as structure-based virtual screening (SBVS). On the other hand, and for most of the cases, the three-dimensional structure of the target, the binding site or even the target itself are not accurately known, although there may be a number of known active compounds that have been identified experimentally. In this case, data mining algorithms can be used to screen for compounds that are structurally similar to the known actives (similarity search),[4] or that comprise the chemical features of these compounds (pharmacophore search),[5] in what is called ligand-based virtual screening (LBVS). Thus, these two fundamental procedures, SBVS and LBVS, form the general layout of present-day VS protocols. Figure 2-1 illustrates the different branches and methods that are followed in current VS campaigns. Detailed descriptions of these methods are summarized below.

2.2 Structure-based VS (SBVS)

SBVS requires the knowledge of the three-dimensional structure of the target.[6] This structure can be obtained by experimental techniques such as NMR, X-ray crystallography, electron crystallography or it can be predicted computationally using homology modeling. It is also important to identify the relevant binding site(s) within the protein that is (are) deemed responsible for its biological activity. Generally, the binding site is a pocket, a groove or a protrusion having an assortment of apparent hydrogen bond donors and acceptors; hydrophobic features; and it can be associated with molecular adherence surfaces. There may be a number of metal ions or water molecules as part of the active site that are essential for the activity of the protein and they must be considered during the screening procedure. There are two basic approaches for SBVS namely docking,[2] and receptor-based pharmacophore modeling.[3]

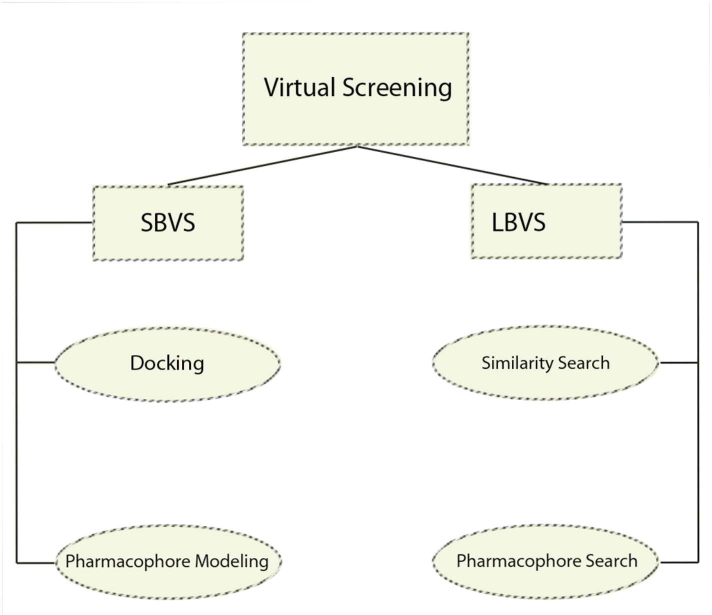


Figure 2-1: Two main virtual screening approaches.

Docking or pharmacophore modeling tools are the best alternatives when the target structure is available. On the other hand, similarity or pharmacophore search are commonly used when the target structure is inaccessible.

2.2.1 Docking

Molecular docking is a standard element of many SBVS studies described in the literature.[7] The idea of docking and scoring as a VS tool has been proposed since the birth of docking methods.[8] The main problem which all docking algorithms try to solve can be stated as follows: given two interacting molecular structures, what is the most probable binding configuration to form a stable three-dimensional protein-ligand complex? In order to address this problem, the docking procedure can be divided into two major steps. First, explore the conformational space of the ligand within the binding site of the target. At this stage, many conformations are generated for the ligand with a limited number of configurations that can actually fit within the binding site. Second, examine all suggested ligand configurations and select the optimal target-ligand alignment by scoring their interactions and ranking the docking results (poses) according to their predicted binding affinity.

Today, there are at least 30 docking programs commercially (or freely) available with different conformational sampling algorithms and a variety of scoring functions.[9] The most commonly used programs are AUTODOCK,[10] GOLD,[11] GLIDE,[12] DOCK,[8] ICM,[13] IFREDA,[14] and FlexX.[15] These programs differ mostly in the way they deal with protein/ligand flexibility or their scoring and ranking methods.

In contrast to the poor representation of target flexibility (see below), most docking methods can handle the flexibility of ligands very efficiently.[9] In other words, for most of the cases, docking algorithms can reproduce the protein-ligand binding modes that have been observed experimentally using X-ray crystallography. As an example (as we will see in details in chapter 7), Figure 2-2 shows the successful docking of nutlin, a well-known p53-MDM2 inhibitor, to the p53-binding site within MDM2 using AutoDock 4.0.

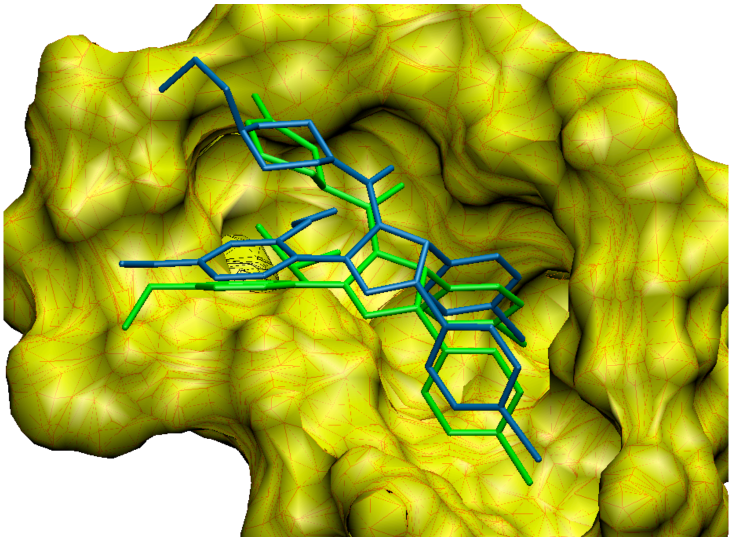


Figure 2-2: Comparison between docked and experimental structures for nutlin.

The binding site within the MDM2 protein is shown in molecular surface representation. Nutlin2 (the experimental structure) is shown in blue and the docked nutlin3 is shown in green. The docking program AutoDock 4.0 was used to carry out the docking calculation and the results are in good agreement with the experimental findings (see chapter 7 for more details).

In general, the degree of success for docking methods can be measured by comparing the predicted binding mode (pose) to the experimental conformation (the native binding mode).[10-11, 14] This assessment can be evaluated quantitatively by calculating the root-mean-square deviation (RMSD) between the two structures. However, in certain systems, where unexpected flexibility of the receptor is crucial for the binding reaction or the interaction of the ligand and protein is mediated by water molecules or metal ions, docking may fail to predict the correct binding conformation of the complex, leading to improper and unrealistic interactions. Below, we will see in more details, how ligand and target flexibility are considered within most docking algorithms. A summary of the different scoring methods will be also introduced.

2.2.1.1 Ligand flexibility

The binding reaction between a ligand and a particular target involves numerous conformational changes in the two molecules as well as water molecules and ions located in their interface. Each entity in this reaction adapts its shape and distribution in order to maximize its interactions with the other entities, forcing the whole system to reach the global minimum of their potential energy surface. This binding interaction is somehow similar to a folding problem of a protein, comprising a huge number of degrees of freedom. Consequently, the majority of docking programs avoid this conformational flare-up problem by implementing almost full-flexibility for the ligands while keeping the target completely rigid with no flexibility allowed.[2, 9]According to the nature of the searching method, one can classify the ways by which ligand-flexibility was introduced within docking methods into three main categories: (1) systematic search routines, (2) stochastic exploration, and (3) simulation techniques.

Systematic search

In a typical systematic search, all rotatable bonds in the ligand are gradually rotated in order to cover all possible combinations among the dihedral angles. Evidently, the number of generated structures using this method increases dramatically with the number of rotatable bonds involved. If not kept under control it may lead to the problem of combinatorial explosion. In this way, applying a standard systematic search to explore the entire conformational space of a ligand requires massive calculations and considerable computational time. The docking program, FLOG,[16] gets around this hindrance by limiting the created structures to a pre-generated set of conformations recorded in structural databases. Other docking algorithms adopt an incremental procedure to reconstruct the ligand within the binding site of the target. The main objective of these methods is to limit the number of degrees of freedom for the ligand, allowing for a less-expensive and rapid conformational search.

Essentially, there are two main approaches for the incremental reconstruction methods. First is the one that has been employed by LUDI,[17] FlexX,[15] DOCK,[8] ADAM [18] and Hammerhead,[19] where the ligand is split into a rigid core fragment that is docked first and a number of flexible regions that are subsequently and successfully added. This method is commonly referred to as the “incremental approach”. The other method, known as “place and join”, is to break the ligand into several fragments, dock them within the binding site of the target and finally connect them together in order to rebuild the final ligand conformation.

Stochastic exploration

Stochastic exploration samples the conformational space of a ligand by generating random variations in the orientation of all rotatable bonds and in some cases random translations for the whole ligand within the binding site. This is done mainly to enable crossing of the energy barriers and searching for local minima enclosed by the rugged energy surface of the ligand. This procedure can be applied to a single ligand or a population of conformations derived from the same molecular structure of the ligand. Each resultant conformation is then evaluated according to a probability distribution or by estimating its binding affinity with respect to the target. In fact, there are three methods that are derived from this technique, namely: Monte Carlo simulations, Genetic Algorithms and Tabu Search methods.[1]

Monte Carlo (MC) simulations are one of the most powerful techniques ever developed to allow for overcoming potential energy barriers and sampling the conformational space of a typical system. Within docking algorithms, the method usually starts with a randomly generated conformation for the ligand by arbitrarily changing one or more dihedral angles or even the whole orientation or position of the ligand with respect to the target. This new conformation is accepted or rejected according to a Metropolis algorithm that follows the Boltzmann probability distribution. Programs like ICM,[13] MCDOCK[20] and DockVision[21] employ this approach.

Genetic algorithms (GA) exploit the biological concepts introduced by Darwin in order to explore all possible conformations of the ligand and predict its native structure. In contrast to MC-based algorithms, instead of manipulating a single ligand, GA generates a random population of the same molecular structure of the ligand.[10-11] Each member of this population is unique in terms of the internal orientation and the global placement and alignment within the binding site. This random population forms the initial generation (seed) of a set of non-interacting ligand species. These poses are further subjected to a number of biological operators that add up more diversity to the generated structures. Among these operators are the mutation operator (generate new ligands from earlier ones by altering a rotatable bond or moving the whole ligand to a new position), and the crossover operator (merge two ligands in order to create a new structure comprising their common features). The fitness of each newly generated structure is evaluated by calculating its binding affinity to the target. The pose that retains the most predominant interactions with the binding site survives and becomes the parent of the new generation. This iterative procedure terminates after reaching a predefined number of generations or energy evaluations, or if no more improvement to the binding affinity has been observed (converged solution). Examples of programs that incorporate genetic algorithms in conformational sampling include AutoDock, GOLD and DARWIN.[22]

As a memory-based stochastic exploration method, Tabu Search (TS) prevents the searching machinery from revisiting the same conformation more than once. PRO_LEADS is one of the most popular programs that employ this searching technique.[23] This is generally achieved by creating a list that records all previously visited solutions, which acts as a memory for the algorithm. A decision to accept or reject a new conformation is made after comparing its RMSD to the other recorded conformations.

Simulation techniques

Simulation techniques employ a deterministic approach that either: (1) passes through both time and space giving rise to an evolving trajectory describing the biological behavior of a typical system, or (2) re-adjusts the system by rearranging its particle composition towards a more stable state. In this context, molecular dynamics (MD) simulations and energy minimization methods are the most widely used simulation techniques in a number of docking programs. Although the two approaches can handle the full-flexibility of both the ligand and target, their foremost disadvantage is that they can be readily trapped within a local energy minimum, which in turn precludes them from efficiently sampling the conformational space of the complex. Therefore, simulation techniques are usually used as a refining step subsequent to GA or MC simulations.[9]

2.2.1.2 Target flexibility

Docking a ligand against a crystal or relaxed receptor structure is a commonly used approach in structure-based drug design.[2] However, in many cases, the degree of success that may be achieved in a typical docking simulation depends on the characteristics of the target and how important is the protein flexibility in the simulation. Most of the successful cases reported in the literature were either related to nearly rigid proteins or proteins having real binding mode of their respective ligands.[24] In spite of these studies, there are cases where the binding interaction has been shown to induce significant conformational changes to the target, ranging from local reorganization of side-chains to hinge movement of domains. Sampling these conformational changes during docking is impractical, as they involve a large number of degrees of freedom. To address such problems, a number of docking packages like AutoDock, GOLD, FlexE and IFREDA, manage to include a modest amount of flexibility in the target during the docking simulations. These approaches include soft docking,[25] side chain flexibility,[1] combined protein grid and united descriptors of the target.[26]

Soft Docking

Soft docking algorithms allow the ligand to penetrate through the surface of the protein in order to approximate and predict the dynamical changes that may take place within the active site as a result of ligand binding. This is generally achieved by attenuating Lennard-Jones repulsive parameters in the potential energy function that describes the system.[25]

Side Chain Flexibility

Another commonly used technique to introduce active site dynamics in the context of docking is to allow key side chains that have been shown to mediate the interactions with the ligand to rotate freely and search for their preferred conformation. These side chain rotations are usually restricted to a number of pre-defined experimental conformations stored in rotamer libraries or predicted from a prior MD simulation. While this method reduces the risks associated with the lack of flexibility to some extent, it neglects backbone dynamics, which may affect the ultimate docking results.[24]

Combined protein grid

In order to account for a larger degree of receptor flexibility at a reasonable computational cost, a number of dominant protein conformations can be combined simultaneously to generate a comprehensive model that describes the essential dynamics of the binding site.[26] This approach is generally termed as "combined protein grid" and is usually implemented in two steps. First, for each conformation, all possible protein-ligand atomic interactions are calculated and recorded in what is called a docking grid. Applying a weighted average for all the resulting grids representing the various conformations then creates a combined grid. Alternatively, the averaging procedure may be applied to the atomic coordinates to generate an average structure for the protein.

2.2.1.3 Scoring Methods

As ranking of the binding modes is crucial in prioritizing and ranking of the compounds, it is important to use sensitive and accurate scoring functions that can replicate and predict experimental data. This is normally achieved using an objective scoring function that directs the conformational search algorithm in predicting the native conformation and ultimately estimates the binding affinity. Nevertheless, it has been broadly demonstrated that docking scoring functions are less successful at predicting the actual binding affinities and at discriminating true binders from inactive (decoy) compounds.[1-2, 27]These puzzling results are direct outcomes of many factors that have been mistreated while analyzing the binding interactions of the resulting poses as a compromise to speed up the docking process. These factors mostly include the lack of proper salvation, the neglect of protein flexibility and the bias toward the training set of structures that have been used in optimizing the scoring process.[24] In fact, developing new scoring functions and innovative ranking schemes is a wide-open area of research in the field of docking. Although a more precise scoring method can be practically implemented within docking, the large computational cost that is associated with such a function will be the actual barrier from using it. In this way, many assumptions have been proposed in the currently used docking scoring functions in order to reduce the complexity and computational time required to evaluate a particular pose. Overall, a typical scoring function includes at least among its ingredients, a descriptor for the hydrophobic effects, van de Waals dispersion interactions, hydrogen bonding, electrostatic interactions, and solvation effects. Based on their scoring functions, all docking programs that are in use today can be divided into four major categories: force field-based, empirical, knowledge-based and consensus methods.[9, 24, 28]

Force Field-Based Scoring Methods

According to the energy landscape theory, the native conformation of a ligand within the binding site of its target is correlated with a profound deep minimum on the energy surface. Therefore, potential energy (force-field) functions have been used to describe protein-ligand interactions and assess their binding affinity by exploring the energy surface and locating these minima. Over the past 30 years, rigorous efforts have been devoted to build new force field models and make them available for a substantial number of applications ranging from molecular docking to molecular dynamics simulations.[29] One of the main problems of such models is the selection of a potential energy functional form and adjusting its various parameters to better represent experimental data or quantum mechanical predictions. These energy functions are commonly restricted to a number of assumptions and approximations for the sake of minimizing their computational time, reducing the efforts of refitting the parameters to more complex representations and, aligning them with many applications that are currently running with force fields of standard functional forms. An obvious example of such restrictions is the use of atom-centered charges in electrostatic calculations. Rationally, a more accurate representation of atomic charges should explicitly represent lone pairs on electronegative acceptors such as oxygen and take electronic polarization into account. As docking algorithms usually deal with a single target conformation, the internal protein interactions are typically neglected. Accordingly, force field methods approximate the ligand-protein binding interactions by adding the interaction energy between the protein and the ligand to the ligand internal energy. These internal interactions are approximated by harmonic springs that describe the vibrations and rotations of the different bonds forming the ligands. The non-bonded interactions between the ligand and its target are estimated by van der Waals, hydrogen bonding, and electrostatic terms. For example, the potential energy function of the general AMBER force field (known as GAFF) is shown below:[30]

$$\begin{aligned} E_{pair} = \sum_{bonds} k_r (r - r_{equ})^2 + \sum_{angles} k_\theta (\theta - \theta_{equ})^2 + \\ \sum_{dihedrals} \frac{v_n}{2} \times \left[1 + \cos(n\phi - \gamma)\right] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \end{aligned} \qquad \text{EQ. 2-1}$$

where r_{equ} and θ_{equ} are equilibrium structural parameters; k_r and k_θ, v_n are force constants; *n* is the multiplicity and γ is the phase angle for the torsional angular parameters. The *A*, *B* and *q* parameters represent the nonbonded potentials (charge-charge and van der Waals terms). Nonbonded interactions can be obtained from liquid state calculations and available experimental data. Other parameters such as stretching, bending, and torsional terms are generally fit to quantum chemical calculations. Noticeably, the major drawback of standard force field scoring functions is the lack of solvation and entropy contributions to the binding energy. Examples of force-field-based scoring functions include D-Score[31] and GoldScore.[32]

Empirical Scoring Methods

Another widely used scoring approach is to hypothesize an empirical scoring function that has been optimized to reproduce a collection of experimental data [17]. These data may include binding affinities or native conformations for known active compounds. Notable examples include F-Score [15], ChemScore [33], SCORE [34] and Fresno [35]. The basic concept behind this type of scoring functions is that the binding energies can be approximated by a summation of unrelated contributions. Each element of this summation describes a certain binding interaction such as hydrophobic, hydrogen bonding, electrostatic or solvation effects. Some functions may comprise an approximation for the loss of entropy due to binding, which is proportional to the number of rotatable bonds included in the ligand. Overall, the terms that build up a typical empirical scoring scheme are simple enough to be rapidly evaluated in order to speed up the docking process. A fairly accurate estimate for the coefficients pre-multiplying these terms can be obtained by performing regression analysis and fit the whole function against the set of experimental data. An example of such functions is the AutoDock scoring function [10] (see below), which has an accuracy of ~ ±2 kcal/mol:

$$\Delta G = \Delta G_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + \Delta G_{tor} N_{tor} + \Delta G_{sol} \sum_{i,j} \left(S_i V_j \right) e^{(-r_{ij}{}^2 / 2\sigma^2)} \qquad \text{EQ. 2-2}$$

here, the five ΔG terms on the right-hand side are coefficients empirically determined using linear regression analysis from the set of protein-ligand complexes with known binding constants. The function includes three in vacuo interaction terms, namely, a Lennard-Jones 12-6 dispersion/repulsion term; a directional 12-10 hydrogen bonding term, where E(*t*) is a directional weight based on the angle, *t*, between the probe and the target atom; and screened Columbic electrostatic potential. In addition, the unfavorable entropy contributions are estimated by a term that is proportional to the number of rotatable bonds in the ligand and solvation effects are represented by a pairwise volume-based term that is calculated by summing up, for all ligand atoms, the fragmental volumes of their surrounding protein atoms weighted by an exponential function and then multiplied by the atomic solvation parameter of the ligand atom (S_i). It should be noted that, although several empirical functions like the above-mentioned AutoDock scoring function have been successfully used for many cases, they are generally biased toward the experimental data that was used in their optimization and are not efficient at eliminating false binders from a set of tested compounds.[9, 24]

Knowledge-Based Scoring Methods

Knowledge-based scoring methods are Similar to the empirical scoring functions. They attempt to reproduce experimentally determined structures using simpler atomic interaction-pair potentials. These potentials are based on the frequency of occurrence of all possible interactions between the ligand and its target. Using statistical analysis, knowledge-based models implicitly describe binding effects that are hard to represent explicitly during docking. Therefore, the accuracy of these methods depends on the extent of the used protein-ligand data set and the diversity of atomic interactions included in these complexes. Example of such functions is DrugScore.[36]

Consensus Scoring Methods

A more recent scoring technique is called consensus scoring. It collects assessments from several scoring functions in order to evaluate a particular docking result.[37] The method is expected to reduce the errors that result from the individual scoring functions and improve the probability of selecting true binders. Nevertheless, it has been recommended to use different and uncorrelated scoring functions in constructing a successful consensus scheme. This is because correlated functions tend to produce similar results leading to error amplification and misleading results. Despite these constraints, several studies have pointed to the success of a number of consensus scoring functions when compared to using a single scoring method.[38] These methods include X-CSORE [39] and FlexX [15] scoring functions.

2.2.2 Structure-Based Pharmacophore Modeling (SBPM)

Pharmacophores are straightforward models that describe the essential interactions behind the binding of a ligand to its target.[40]The concepts behind pharmacophore modeling dates back to 1909, when Ehrlich defined a pharmacophore model as "a molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon's*) biological activity". These features are generally classified into two major categories, namely, chemical-based and shape-based features. The former include hydrogen bond acceptors or donors, charge centers, metal binding regions, aromatic rings and hydrophobic regions, while the latter mainly include volume-excluded regions and geometrical constraints like distances, angles and dihedral angles. By allocating the different features and including their three-dimensional distributions within the binding site, one can understand the essential properties required for bioactivity of known true binders.[41]

Currently, there are two approaches to generate a pharmacophore model depending on the accessibility of the target structure. If the structure of the target is available, one can build a pharmacophore hypothesis that would complement the chemical features within the binding site of the target.[5] These pharmacophore models can be further improved if there is any active compound that has been co-crystallized with the target. This approach is called "binding sites-based pharmacophore models" or "structure-based pharmacophore models". The second approach, that is more widely used, relies only on the known active compounds and no information about the target is required. This technique is commonly referred to as "ligand-based pharmacophore modeling" and will be explained in detail later in this chapter. In this section, we will focus on the "structure-based pharmacophore models".

The most straightforward approach in designing a SBPM is to analyze the experimentally determined crystal structures of protein-ligand complexes. For example, the program LigandScout[42] introduced by Wolber and co-workers is very effective at manipulating these structures and automatically interprets the various interactions between a particular macromolecule and its co-crystallized ligands into functional pharmacophore models. The program starts by cleaning up the structures of the ligands by assigning hybridization states and bond characteristics that are missing in the crystal structure. This is accomplished by using an extended heuristic approach combined with template-based numeric analysis. Following this step, pharmacophore models are created by analyzing the atomic interactions between the ligand and all residues located within the binding site of the target. These interactions are classified into complementing groups in terms of hydrogen bonding, electrostatic charges and hydrophobic contacts. Moreover, by aligning several bound confirmations of the ligand, one can, partly, incorporate the flexibility of the complex to generate what is called a common-feature pharmacophore model.

When only the structure of the active site is available, programs like structure-based focusing (SPF) can suggest its complementing pharmacophore hypotheses.[43] The process starts by mapping favorable regions or "hot spots" for protein-ligand interactions within the binding site of the target. These regions are then clustered into hydrogen bond donating and hydrogen bond accepting vectors and hydrophobic interaction sites. The clustered groups are then used to build the pharmacophore model. Other algorithms that can construct SBPMs in addition to LBPMs (see below) are Unity (Tripos. http://www.tripos.com/) and MOE (Molecular Operating Environment; http://www.chemcomp.com/). Once a pharmacophore hypothesis is created, the model can be converted to a query that is used to screen chemical databases for molecules that satisfy these proposed hypotheses.

2.3 Ligand Based Virtual Screening (LBVS)

Despite the advances in macromolecular structure prediction methods, the number of protein structures that have been determined experimentally is still lagging compared to that of their sequenced counterparts. In this case, homology modeling plays a key role in understanding and predicting the three-dimensional structure of the target. However, homology modeling has its own limitations and the degree of success of incorporating the method within the context of VS depends mainly on the quality of the predicted structure[44]. Therefore, it is important to seek alternative routes that depend merely on known active compounds and in which no information about the target is required. These ligand-based filtering techniques have played a significant role in discovering potent inhibitors for many targets. In fact, ligand-based screening methods use known active and inactive compounds as templates and employ comparative algorithms to identify new compounds that are similar to these templates. Overall, one can classify the different LBVS methods into two main approaches, namely, similarity search,[4] and pharmacophore search.[41]

2.3.1 Similarity search

The fundamental theory behind this approach is Maggiora's "similar property principle", which states that similar molecules are more than likely to have similar properties. While not universally correct, there are many cases where this simple idea showed great success and helped in the discovery of novel active molecules.[45] According to this concept, one can use known active compounds as reference structures and filter a given chemical library for ligands that are structurally similar to the active molecules. The filtered compounds are expected to display some activity that in some cases could be greater than the original reference structures. In fact, there are mainly two ways to assess the similarity between two molecular structures. These methods include molecular alignment and molecular descriptors algorithms [46].

Molecular alignment algorithms

Molecular alignment algorithms such as FlexS [47] or GASP [48] typically align the filtered compounds with the reference structure and rank them according to their degree of similarity. During the superimposition process, the two aligned molecules can be treated either as rigid or flexible. Similar to docking methods, flexibility can be introduced by employing an incremental construction approach (FlexS) or a genetic algorithm procedure (GASP). Other algorithms like Fflash [49] apply fragment-based techniques to incorporate ligand flexibility during the filtering process. Other algorithms incorporate Gaussian functions as in the program MIMIC,[50] or constructing interaction potential grids around molecules.[51]

A major drawback of molecular alignment techniques is that the time required for a single molecule comparison is long enough to discourage a user from employing the method in screening large databases.[4] As a result, more efficient and accurate techniques have been developed to describe the information inherited in the molecular structure of a given ligand along with its physiochemical and topological properties. These are molecular descriptors methods.

Molecular descriptors algorithms

Molecular descriptors are generated on-the-fly and are compared to the reference structure very rapidly. Based on the dimension of the information that is used, molecular descriptors can be classified into 1D-, or 2D-descriptors. Evidently, the higher the dimension of the descriptor approach, the longer its computational time will be and the higher the accuracy one can expect from the searching protocol. Generally speaking, bulk properties like molecular weight, molar refractivity or log P values are adequate to construct a 1D-molecular descriptor.[52] However, since there is no information about the structural properties or chemical features of the ligand, it is impossible to only rely on such descriptors in filtering a typical chemical library for active molecules. Consequently, one should draw on a higher level of information and include structural properties as an additional descriptor in order to increase the accuracy of the method. This introduced molecular fingerprints as the most successful and widely used similarity search approach in LBVS.

Molecular fingerprints are bit-string representations that reflect structural features and other properties of a molecule given its chemical structure.[4] Key advantages of this approach over direct comparisons of molecules are that it is very simple to implement, remarkably fast to calculate and the final outcome is expressed as a single number that quantifies the degree of similarity. According to the complexity level and design scheme, one can recognize two basic approaches in generating a molecular fingerprint for a specified structure. The first approach is what is known as "keyed" representation. In this case, an individual bit within the string can be set as "on" or "off" reflecting the presence or absence of a pre-defined functional group (pattern) in the sub-structural space of the ligand. While the order of the bit-string map is the same for each molecule, the individual bits are turned on or off if their representative substructure exists or not. A widely used VS algorithm that employs this procedure is MACCS whose bit-strings may include up to 166 bits representing commonly known fragments. The second approach is known as the "hashed" representation. This method resembles human fingerprints by not restricting the definition of bits to describe a pre-specified set of patterns. That is, like human fingerprints, which are very characteristic of individuals, a pattern's fingerprint characterizes the pattern, but the meaning of any particular bit is not well defined. To do so, a typical hashed representation algorithm starts with generating a pattern for each atom. Then it creates a pattern representing each atom and its nearest neighbors in addition to the bonds that join them. This hierarchal construction evolves to include higher order nearest neighbors until the complete structure is recovered.

In the heart of these similarity-based VS techniques lays a similarity measure, usually termed a similarity coefficient that is used to quantify the degree of resemblance between two molecules. In fact, the most commonly used parameter is Tanimoto (Jaccard) coefficient (described in equation 2).[4] To understand the concept behind this parameter, let us consider the case of 2D fingerprints representing two molecules A and B that have *a* and *b* bits that are set as true, respectively. Now, if there are *c* common bits that are mutually set as true in the two molecules, where *c* is the intersection subset of *a* and *b*, one defines their Tanimoto coefficient as:

$$T = \frac{c}{a + b - c} \qquad \text{EQ. 2-3}$$

The Tanimoto coefficient gives values between zero (no similarity) and one (maximum similarity). This coefficient is the most popular choice for both in-house and commercial screening packages.

2.3.2 Pharmacophore search

While the basic concepts behind the pharmacophore search approach have been introduced in previous sections, in this part of the chapter we will focus on pharmacophore modeling techniques that have been broadly followed in the literature if no target structure is available [40-41, 53]. In this case, the only information that can be exploited is a set of known active compounds that are recognized experimentally and the general procedure can be summarized in two fundamental steps. First, this set of molecules is analyzed in order to identify all chemical features within their structures. Then, for each molecule an ensemble of different conformations is generated and used to produce the best alignment between the different compounds to overlay their corresponding features. Although the main approach seems feasible and simple to implement, searching the conformational space is the most important and most difficult part of the method. This is because it is hard to predict the active conformer of a given ligand without understanding how it interacts with the target, with the solvent molecules and other elements of the binding environment. Nevertheless, there are several programs that have been successfully used in building ligand-based pharmacophore models for many targets. These programs differ mostly in the way they handle ligand flexibility and the method of searching a typical chemical database for promising hits. The most popular programs are Catalyst [53], DiscoTech,[45b] and GASP.[48]

Catalyst introduces ligand flexibility very efficiently and, in the mean time, is extremely fast in searching 3D chemical databases [53]. In brief, the program extensively explores the conformational space of a ligand by using a random search algorithm along with a poling function that creates a large number of low-energy conformations. Catalyst follows two alternative algorithms in building up pharmacophore models. The first algorithm, HypoGen, is a quantitative approach in which each chemical feature allocated to the ligand structure is associated with a particular weighting factor that is related to its relative importance in describing the bioactivity of the molecule. Following this procedure, the algorithm builds up a number of pharmacophore hypotheses and ranks them based on their ability to explain available experimental data. In the other approach, Catalyst follows a qualitative procedure that is termed the HipHop algorithm. In this process, for each ligand, the algorithm checks for the surface accessibility for receptor interactions. Then, chemical features are defined based on their absolute coordinates in the different conformations of the molecule rather than by their inter-feature distances. This procedure usually starts with the most active compounds in the training set followed by highlighting their matching features from other less-important molecules. This results in a considerable number of proposed pharmacophore hypotheses that is significantly reduced by rejecting models that cannot explain the bioactivity of these molecules. Regardless of which approach is used, Catalyst can merge different models in order to generate a more comprehensive pharmacophore hypotheses.

Disco not only suggests pharmacophore models that demonstrate the important features in a ligand, but it also predicts their potential complementary regions that should be located within the binding site.[45b] This is accomplished by breaking up a pharmacophore to groups of ligand points and binding pocket interaction sites. Ligand points include atoms with hydrogen bonding properties, charge centers and hydrophobic characteristics. Binding pocket interaction sites are predicted to be complementary regions within the target and are calculated using the coordinates of the heavy atoms of the ligand. Similarly to Catalyst, the conformational flexibility of the ligands is explored using a set of pre-calculated conformations for each ligand in the training set. However, one pitfall of using Disco is that all chemical features that make up the final pharmacophore model must be identified in every molecule, which may result in the exclusion of talented models.

In contrast to both Catalyst and Disco, the program GASP handles the ligand conformational flexibility in a very sophisticated manner.[48] Instead of using a pre-calculated set of ligand conformations, the program uses a genetic algorithm to explore the conformational space of the ligand during the pharmacophore generation process. GASP algorithm starts by detecting all possible chemical features in the structure of each ligand. The molecule with the least number of features is selected as a reference structure. Every structure in the training set is then fitted to the reference structure using a genetic algorithm that is similar to what is used in docking programs (see above). However, in this case, the fitness of a particular model is measured based on a combination of similarity, the number of overlaid features and the volume integral of the overlay. One more advantage for GASP over DISCO and Catalyst is that, models generated by GASP account for the steric clashes between the ligands in generating the final pharmacophore model. On the other hand, the other two programs propose their models by only matching the chemical features of the ligands without taking their overall shape into account.

No matter which approach is used to generate a pharmacophore model, which includes SBPM, there are two main ways to screen chemical libraries for compounds that satisfy the constraints of the pharmacophore hypotheses.[40-41] First, one can use a database file format that includes a set of pre-defined conformers for each compound in the database. Although this approach remarkably speeds up the search process, it requires massive storage of the different conformations. Alternatively, a single conformation can be used as a precursor for generating an ensemble of conformations followed by fitting these structures to the pharmacophore query during the screening process. While this procedure eliminates the need for substantial storage, it is much slower than the former method.

2.4 Conclusion

The present chapter introduced us to the world of computational virtual screening (VS). It showed that VS tries to solve the same problem as its experimental high throughput-screening counterpart. They both attempt to complement a given binding site of a particular target with a small ligand that would block or provoke the target biological activity. VS is cheaper, faster and sometimes more yielding than the experimental approach. Nevertheless, there are also so many cases where the computational VS showed very low success due to mistreatment of various factors during the simulations. Although VS can employ two independent computational alternatives, namely, SBVS and LBVS, it is important to think about these two methods as two arms in the same body. They hold the same problem together and complement each other. For example, as was implemented in this work (see next chapter), LBVS can enrich the used ligand VCC with similar compounds to the known actives. The enriched VCC is then used for SB screening in order to identify more active compounds and suggest them for experimental testing. The protein flexibility and scoring are still persisting as the two main problems facing ranking of hit compounds in VS.

2.5 Bibliography

1. Schneider, G.; Bohm, H. J., Virtual screening and fast automated docking methods. *Drug Discov Today* **2002,** *7* (1), 64-70.
2. Abagyan, R.; Totrov, M., High-throughput docking for lead generation. *Current opinion in chemical biology* **2001,** *5* (4), 375-82.
3. Good, A. C.; Krystek, S. R.; Mason, J. S., High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discov Today* **2000,** *5* (12 Suppl 1), 61-69.
4. Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* **2006,** *11* (23-24), 1046-53.
5. Krovat, E. M.; Fruhwirth, K. H.; Langer, T., Pharmacophore identification, in silico screening, and virtual library design for inhibitors of the human factor Xa. *Journal of chemical information and modeling* **2005,** *45* (1), 146-59.
6. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002,** *7* (20), 1047-55.
7. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002,** *47* (4), 409-43.
8. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology* **1982,** *161* (2), 269-88.
9. Bissantz, C.; Folkers, G.; Rognan, D., Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of medicinal chemistry* **2000,** *43* (25), 4759-67.
10. Goodsell, D. S.; Olson, A. J., Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990,** *8* (3), 195-202.
11. Jones, G.; Willett, P.; Glen, R. C., Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of molecular biology* **1995,** *245* (1), 43-53.
12. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* **2004,** *47* (7), 1739-49.
13. Totrov, M.; Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* **1997,** *Suppl 1*, 215-20.
14. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K., Gaussian docking functions. *Biopolymers* **2003,** *68* (1), 76-90.
15. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology* **1996,** *261* (3), 470-89.
16. Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P., FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J Comput Aided Mol Des* **1994,** *8* (2), 153-74.
17. Bohm, H. J., The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *Journal of computer-aided molecular design* **1992,** *6* (1), 61-78.
18. Mizutani, M. Y.; Tomioka, N.; Itai, A., Rational automatic search method for stable docking models of protein and ligand. *Journal of molecular biology* **1994,** *243* (2), 310-26.
19. Welch, W.; Ruppert, J.; Jain, A. N., Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry & biology* **1996,** *3* (6), 449-62.
20. Liu, M.; Wang, S., MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *Journal of computer-aided molecular design* **1999,** *13* (5), 435-51.
21. Hart, T. N.; Read, R. J., A multiple-start Monte Carlo docking method. *Proteins* **1992,** *13* (3), 206-22.

22. Taylor, J. S.; Burnett, R. M., DARWIN: a program for docking flexible molecules. *Proteins* **2000,** *41* (2), 173-91.

23. Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D., Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998,** *33* (3), 367-82.

24. Reddy, A. S.; Pati, S. P.; Kumar, P. P.; Pradeep, H. N.; Sastry, G. N., Virtual screening in drug discovery -- a computational perspective. *Current protein & peptide science* **2007,** *8* (4), 329-51.

25. Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S., Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002,** *46* (1), 34-40.

26. Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M., Molecular docking to ensembles of protein structures. *Journal of molecular biology* **1997,** *266* (2), 424-40.

27. Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M., Structure-based discovery of inhibitors of thymidylate synthase. *Science (New York, N.Y* **1993,** *259* (5100), 1445-50.

28. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **2004,** *3* (11), 935-49.

29. Guvench, O.; Greene, S. N.; Kamath, G.; Brady, J. W.; Venable, R. M.; Pastor, R. W.; Mackerell, A. D., Jr., Additive empirical force field for hexopyranose monosaccharides. *J Comput Chem* **2008,** *29* (15), 2543-64.

30. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-74.

31. Kramer, B.; Rarey, M.; Lengauer, T., Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999,** *37* (2), 228-41.

32. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins* **2003,** *52* (4), 609-23.

33. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design* **1997,** *11* (5), 425-45.

34. Tao, P.; Lai, L., Protein ligand docking based on empirical method for binding affinity estimation. *Journal of computer-aided molecular design* **2001,** *15* (5), 429-46.

35. Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V., Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *Journal of medicinal chemistry* **1999,** *42* (22), 4650-8.

36. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *Journal of molecular biology* **2000,** *295* (2), 337-56.

37. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P., Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of medicinal chemistry* **1999,** *42* (25), 5100-9.

38. Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S., A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein-ligand binding affinities. *Journal of medicinal chemistry* **2001,** *44* (14), 2333-43.

39. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design* **2002,** *16* (1), 11-26.

40. Chen, Z.; Li, H. L.; Zhang, Q. J.; Bao, X. G.; Yu, K. Q.; Luo, X. M.; Zhu, W. L.; Jiang, H. L., Pharmacophore-based virtual screening versus docking-based virtual screening: a benchmark comparison against eight targets. *Acta pharmacologica Sinica* **2009,** *30* (12), 1694-708.

41. Sun, H., Pharmacophore-based virtual screening. *Current medicinal chemistry* **2008,** *15* (10), 1018-24.

42. Wolber, G.; Langer, T., LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model* **2005,** *45* (1), 160-9.

43. Hoffren, A. M.; Murray, C. M.; Hoffmann, R. D., Structure-based focusing using pharmacophores derived from the active site of 17beta-hydroxysteroid dehydrogenase. *Current pharmaceutical design* **2001,** *7* (7), 547-66.

44. Cavasotto, C. N.; Phatak, S. S., Homology modeling in drug discovery: current trends and applications. *Drug discovery today* **2009,** *14* (13-14), 676-83.

45. (a) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E., Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *Journal of medicinal chemistry* **1996,** *39* (16), 3049-59; (b) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A., A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *Journal of computer-aided molecular design* **1993,** *7* (1), 83-102.

46. Lemmen, C.; Lengauer, T., Computational methods for the structural alignment of molecules. *Journal of computer-aided molecular design* **2000,** *14* (3), 215-32.

47. Lemmen, C.; Lengauer, T.; Klebe, G., FLEXS: a method for fast flexible ligand superposition. *Journal of medicinal chemistry* **1998,** *41* (23), 4502-20.

48. Jones, G.; Willett, P.; Glen, R. C., A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *Journal of computer-aided molecular design* **1995,** *9* (6), 532-49.

49. (a) Kramer, A.; Horn, H. W.; Rice, J. E., Fast 3D molecular superposition and similarity search in databases of flexible molecules. *Journal of computer-aided molecular design* **2003,** *17* (1), 13-38; (b) Pitman, M. C.; Huber, W. K.; Horn, H.; Kramer, A.; Rice, J. E.; Swope, W. C., FLASHFLOOD: a 3D field-based similarity search and alignment method for flexible molecules. *Journal of computer-aided molecular design* **2001,** *15* (7), 587-612.

50. Mestres, J.; Rohrer, D. C.; Maggiora, G. M., A molecular field-based similarity approach to pharmacophoric pattern recognition. *Journal of molecular graphics & modelling* **1997,** *15* (2), 114-21, 103-6.

51. Goodford, P. J., A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry* **1985,** *28* (7), 849-57.

52. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **2001,** *46* (1-3), 3-26.

53. Barnum, D.; Greene, J.; Smellie, A.; Sprague, P., Identification of common functional configurations among molecules. *Journal of chemical information and computer sciences* **1996,** *36* (3), 563-71.

Chapter 3: The Implemented Virtual Screening Workflow[(1)]

After reviewing all VS methods and concepts that have been used and developed during the last two decades, it is time to describe the VS protocol that was implemented to carry out this work. Figure 3-1 illustrates the essential steps that construct the overall workflow of the VS procedure. In a nutshell, the developed protocol employs molecular docking, molecular dynamics simulations and clustering techniques to filter any given library of compounds for inhibitors of a particular target. The detailed description and rationale behind each step are summarized below. Except for a few steps that need careful preparation, the whole process has been automated. It starts with a collection of 3D structures of ligands and a well-prepared target structure. It finally yields a set of top hit structures in their preferred binding modes to the target. Although the following steps were applied to the few problems described in the following chapters, the procedure is general and the same method is applicable for almost any target.

3.1 Target preparation

All the targets that were studied in this work are proteins. Their structures can be downloaded easily from the protein data bank (PDB) database.[1] However, it is not a good practice to use the downloaded structures directly in simulations. The target structures should be cleaned up and prepared carefully prior to any computational work. Although, the following target preparation steps can be automated within the screening procedure, it is not recommended to automate them, as they need careful visual examination, which takes into account all possibilities and individual differences among them.

3.1.1 Primary assessment of target structure

In general, the downloaded "crude" crystal structure of a target contains many details that must be taken into account. This includes nonstandard amino acids; cofactors; other small molecules that are there

[(1)] A version of this chapter has been published in Barakat K, H, Tuszynski J (2011) Virtual Screening for DNA Repair Inhibitors. In: Storici F, editor. DNA Repair - On the Pathways to Fixing DNA Damage and Errors. 1 ed. Rijeka: InTech.

due to the crystallization process; ions and co-crystallized water molecules. At this stage, an important decision must be made on whether to keep these extra molecules or delete them before carrying out the simulations. For most of small molecules like polyethylene glycol, it is better to remove them from the structure, as they are not included in the native form of the target, however, they were required for the crystallization process. Nonstandard amino acids must be assessed and modeled. In many protein structures, these unusual amino acids lack several atoms because most structure handling packages don't check on them automatically. Their parameters must be appended to the used Force Field (FF) before starting any type of simulation. Cofactors, ions and co-crystallized water molecules should be included within the structure. This conclusion has been strongly recommended in the literature and was necessary to study the DNA-p53 binding characteristics shown in figure and introduced latter in chapter 8.

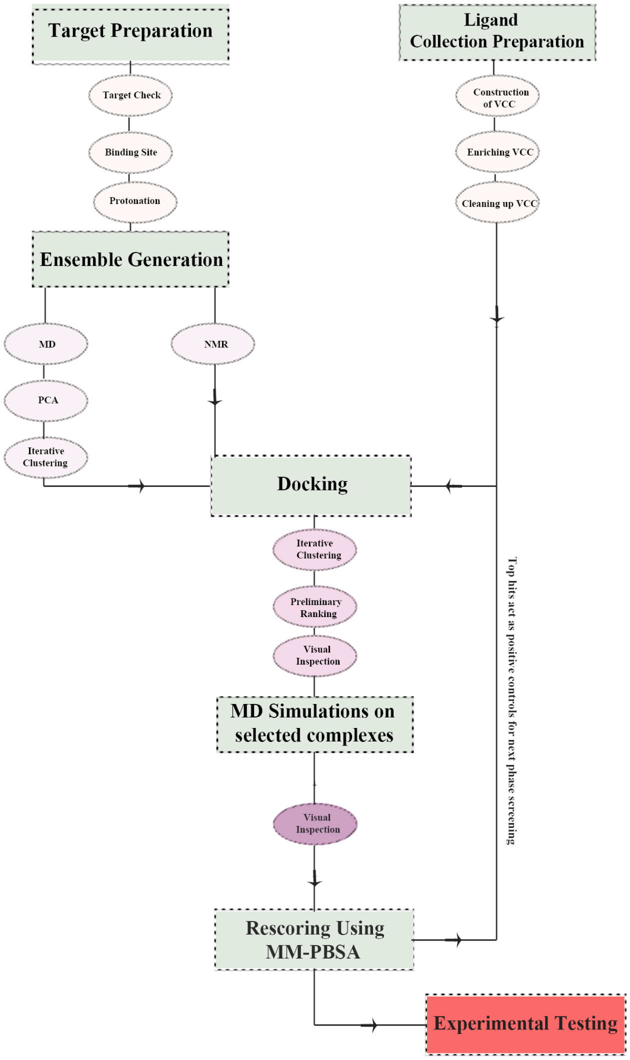


Figure 3-1: The implemented VS protocol. See text for details.

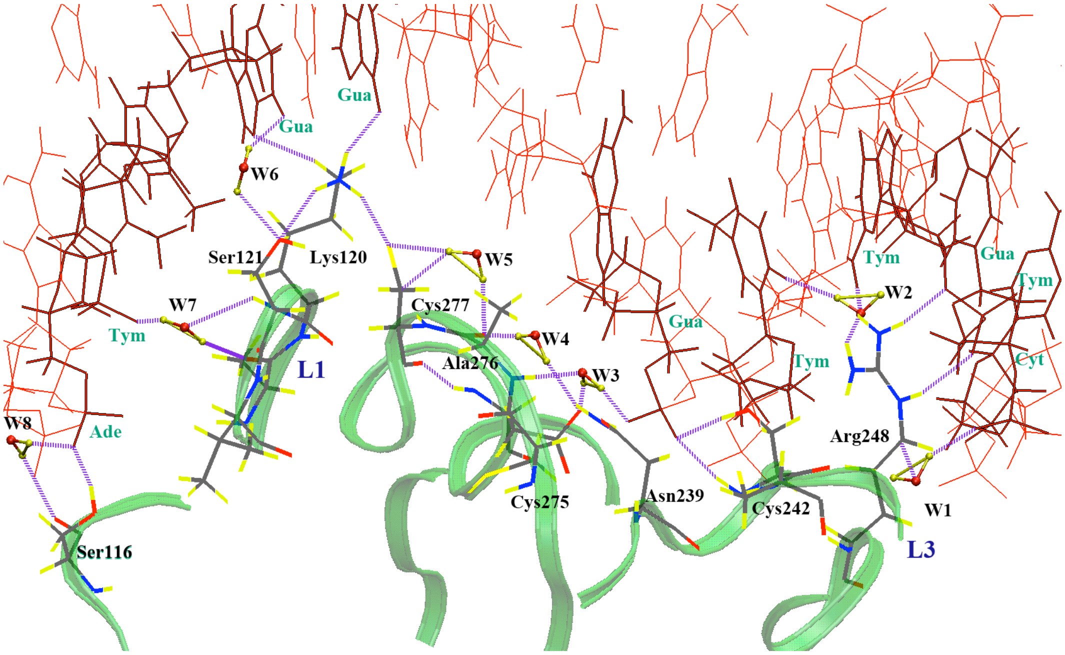


Figure 3-2: Hydrogen bond network at the DNA-p53 binding interface at 300K. Water molecules played an important role in maintaining the binding stability between the protein and DNA (see chapter 8 for more details).

Water molecules that are located close to or within the binding site can mediate several interactions with the ligands. However, it is important to find out which water molecules are conserved within these regions. Any unpreserved (misplaced) water molecule can obstruct the docking simulation and lead to wrong results. One way to identify important water molecules is to compare several crystal structures of the same target (if applicable) and choose which water molecules to keep during the docking procedure. Another “tedious” way if only a limited number of target structures are available is to run different docking exercise by removing/keeping these water molecules and select the ones that lead to realistic and favorable binding modes. A final way to decide on these water molecules is to use prediction software packages (e. g. ConSolv 1.0)[2] that check whether a bound water molecule is likely to be conserved or displaced in other, independently-solved crystallographic structures of the same target.

One other aspect in target preparation is that sulfur atoms of two neighboring cysteine residues can covalently bind and form disulfide bonds. This information should be included within the crystal structure text file (pdb-file) and must be visually inspected before simulation. Usually, a disulfide bond will form between two adjacent cysteine residues if they have the proper geometry and the distance between the two sulfur atoms is around 2 Å. Finally, it is necessary to verify that no parts of the protein structure are missing. These missing residues are usually mentioned at the header of the pdb-file and must be added and relaxed within the target structure. Regarding the targets that were studied in this work, all missing amino acids were distant away from the binding site. Nevertheless, we added and relaxed them using molecular dynamics simulations before running the docking experiments (see below).

3.1.2 Identifying the binding site

The starting point of any VS protocol is the identification of the binding site within the target protein. This portion of the protein is directly related to the biological activity that needs to be regulated. At this stage, it is important to consult previously published work and determine if there are any known active compounds and to ascertain their location of binding. If the binding site is not exactly known, however, there is a set of active molecular structures that exist, one should run a series of blind docking experiments until a suitable and experimentally verified binding site is recognized. Regarding the other three targets in this work, only two, namely MDM2-p53 and pol β, had previously identified inhibitors. Moreover, for the three proteins, the binding sites were accurately known, mainly because they are protein-protein interaction sites (e.g. ERCC1-XPA and p53-MDM2/MDMX), or protein-DNA binding sites (e.g. DNA-pol β) where crystal structures of the interacting subunits are available.

Typically, the binding site is defined as “all target atoms that are located within 6-10 Å from any atoms of the bound ligand/peptide”. The selection of the cutoff is crucial and is essentially done on a try and error basis during the initial docking optimization step. If the cutoff is too small (i.e. less than 6Å of ligand atoms) it may bias the docking results toward the binding mode of the known inhibitor and generate a few number of ligand conformations that does not represent all possible binding modes for the other ligands. On the other hand, if it is too large (i.e. greater than 10Å) the docking simulation can produce enormous solutions that are hard to be clustered into definitive binding modes. A more accurate way to choose the residues constructing the binding site is to perform binding energy analysis between the co-crystallized ligand/peptide and the target using molecular dynamics simulations and break the predicted energy into residue contributions. This strategy was used to define the binding site within ERCC1 that interacts with the XPA protein (see chapter 4).

3.1.3 Protonation states of charged residues

The proper adjustment of the protonation states “the assignment of Hydrogen atoms” of the ionizable groups contained by the target structure is important for any successful VS simulation. These residues play key roles in interaprotein, protein-solvent and protein-ligand interactions. The protonation states can be determined by predicting the pK_a values of charged residues and compare it to the pH value in which the simulation takes place. In this work, all protonation states of ionizable residues were calculated using the software PROBKA and adjusted at physiological pH of 7.0.[3] PROBKA is a very fast and accurate method that relates the structure and environment of the charged residues to the change of the pK_a values from their intrinsic ones. Again, visual inspection of the location of these residues is necessary to validate the predictions of PROBKA, especially, for histidine residues, which possess complicated protonation characteristics. An example for such predictions is shown in Figure 3-3 for the Zinc binding domain of p53. Once the protonation states have been decided, all hydrogen atoms are then added to the system according to a given force field (FF). For this work the AMBER99SB FF was used. By the end of this phase, the protein structure is ready for docking or molecular dynamics simulations.

3.2 Ligand Collection Preparation

Parallel to target preparation is the organization and cleaning up of the set of compounds with which to screen. Currently, there are many suitable, easy to access compound databases that contain millions of molecules spanning various structural families. These databases resulted from the combined efforts of the pharmaceutical industry and many research groups all over the world. Prior to any screening campaign, one should decide on which set of compounds that will be filtered and build up his virtual compound collection (VCC) of compounds. This collection will be repeatedly used against many targets. A typical VCC should include marketed drugs, lead-like compounds, fragment structures, commercially available chemicals and other high-activity molecules. It is also important to represent these molecules in different protonation, stereo and conformational states. An effective VCC should be constructed from molecules that are suitable for further lead optimizations, after they show biological activity.

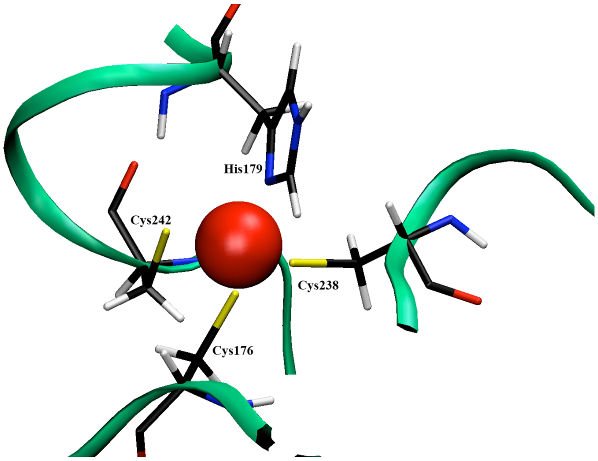


Figure 3-3: The protonation states for the residues constituting the Zinc binding site in p53. All charged atoms that are facing Zinc (shown in red) are deprotonated (has no hydrogen atoms).

3.2.1 Construction Of The VCC

In this work, 4 different databases comprise the core of our VCC. They are the National Cancer Institute diversity set (NCIDS),[4] the DrugBank database,[5] subsets of the ZINC database[6] and finally, the French national chemical library "la Chimiothèque Nationale" (CN).[7] Some of them are used in the first iteration of VS and others are retained for the higher order searching rounds.

The NCIDS is a collection of approximately 2,000 compounds that are structurally representative of a wide range of molecules, representing almost 140,000 compounds that are available for testing at the NCI. A number of its ligands contain rare earth elements and cannot be properly parameterized for docking experiments, leaving us with 1,883 compounds that can be actually used. This work exploits a cleaned 3D version of the NCIDS formatted for use in AutoDock (the main docking program to produce this thesis) and was prepared by the AutoDock Scripps team. What makes the NCIDS so valuable and extensively screened by many groups (even in HTS) is that its individual molecules have distinctive structures and are the cluster representatives of their parent families. Once screened and a number of its molecules rank high in the hit list, one can return back and screen the whole family of the representative structure, instead of screening the actual NCI set of compounds.

The DrugBank database is not only a set of molecules representing FDA-approved dugs, but also it is a unique bioinformatics and cheminformatics resource. It relates each drug to its target(s). It includes details about the different pathways, structural information and chemical characteristics of these targets and the way they take part in inducing a particular disease. This information is stored in a free available website that is linked to other databases (KEGG, PubChem, ChEBI, PDB, Swiss-Prot and GenBank) and a range of structure displaying applets. The DrugBank collection includes ~4,800 drug structures including >1,350 FDA-approved small molecule drugs, 123 FDA-approved biotech (protein/peptide) drugs, 71 nutraceuticals and >3,243 experimental drugs. Once a hit is identified from this library, it is simply a drug. This means it overcomes many barriers of preclinical and clinical experiments and can be tested directly for its novel biological activity. Moreover, a hit from this collection may explain a mysterious side effect that would not be discovered before its identification as a regulator for the examined target.

ZINC is a free database dedicated to VS. It includes more than 13 million purchasable compounds most of them are "drug-like" or "lead-like". The compounds are available in several 3D formats and compatible with several docking programs. The ZINC database has many other interesting features. For example, one can easily create a subset of the whole database with any given set of properties such as functional groups, molecular weight, and calculated logP. Most of the compounds also exist in multiple protonation states suitable for different pH values, several tautomeric forms, all possible stereochemistries, and different 3D conformations. The database is also organized so that the origin for each molecule is known. That is, one can determine the vendor and original catalog number for each commercial source of that compound. Similar to the DrugBank database, a molecule can be annotated for its function or activity. It also has a powerful web server that helps in searching, browsing, creating subsets, and downloading some or all of the molecules of the database.

The CN chemical library (~50,000 compounds) is a repository of all synthetic, natural compounds and natural extracts in the existing French public laboratories. The whole database is divided into two main categories. The first part includes information about all synthetic products, while the second contains the natural compounds and extracts. In this work, we used the whole CN database in our screening. Contrary to the previously mentioned databases, compounds in this library are represented by 2D SDF structures with no hydrogen atoms attached. This required a number of cleaning and preparation steps before using them in VS simulations (see below).

3.2.2 Enriching the VCC Core

This is where ligand-based methods come to play a significant role in the pre-screening process. Any molecule that is known to bind to the target-binding site can serve as a positive control. Such molecules can be identified through published articles or previous patents. Besides their function in directing and verifying the simulation parameters, they can be used as seeds for searching for similar chemical structures to enrich the VCC. This step is crucial and should be done even if the identified similar structures have been previously removed from the VCC in its early construction steps. As previously mentioned in section 2.3.1: Similar molecules are more than likely to have similar properties, these compounds can bind with comparable affinity to their parent seeds.

Following this strategy, we used known inhibitors for the p53-MDM2 interaction (see chapter 7) and DNA pol β (see chapter 6) to enrich their representative VCCs. For the ERCC1-XPA interaction (chapter 4), initially, there was no active compound that was confirmed to bind to the ERCC1 pocket. Hence, for the first round of VS, we started from scratch and didn’t apply this enrichment method. However, it was used in the second round of screening, after the first iteration identified a list of novel binders to the ERCC1 target.

3.2.3 Cleaning Up the VCC

After deciding on which collection of compounds to be used in the screening process, one should spend time and effort to assure the quality of the used ligand structures. As mentioned before, it is important to have proper protonation and conformational states for the ligands. For example, the original CN library of compounds is a collection of 2D structures with no hydrogen atoms. Ligands in this state are not suitable for docking using many of the popular docking programs. These software packages require 3D structures with proper placement of hydrogen atoms. One solution to this problem, which was followed in this work, is to use conversion software that can translate the 2D information into its 3D representative structure. Many of such programs are available (e.g. Open Babel[8] and LigPrep from Schrödinger[9]). We prefer LigPrep for this task because it produces structures with fewer errors compared to Open Babel, especially in bond connection and hydrogen atoms assignment. LigPrep has strong and consistent capabilities in handling of chemical structures. Besides the conversion from 2D to 3D configurations, it can generate variants of the same ligand with different tautomeric, stereochemical, and ionization properties. It also includes energy minimization protocols that can relax the generated structures into their preferred configurations. Moreover, it includes flexible filters that can be used to clean the processed structures from any ligand with no desirable properties.

3.3 Generation of an Ensemble of Target Structures

Proteins are dynamical entities in nature. Their dynamical behavior is essential to recognize and bind to other molecules inside the cell. As we have seen in section 2.2.1.2, although many attempts have been done to partly include the flexibility of the target within docking algorithms, there are many barriers and challenges that preclude the progress of this field. One major challenge is the enormous number of conformations that are accessible to the target under equilibrium conditions. The range of these conformations is wide and includes many local and global movements within the structure of the protein. These dynamics can be as small as little rotations of the side-chains or as large as the complete dislocation of domains within the same target. There are many crystal structures in the Protein Data Bank (PDB) that give evidence to this bizarre dynamic behavior. These conformational changes can be illustrated by comparing different crystal structures of the same target, especially, between its bound and unbound forms. Depending on the time-scale of such movements, one should employ the right method to detect it. For small movements such as side-chain rotations or little loop movements, X-ray crystallography and standard MD simulations can be used. However, NMR crystallography and multiple trajectories MD (e.g. Replica Exchange MD) or Coarse Grained computations would be appropriate for understanding larger motions.

3.3.1 "Induced Fit" vs. "Conformational Sampling"

The existence of different structures and the diversity of their conformations for the same target, pose an important question. Is it the interaction between the ligand and its target that induces these conformational changes for the two interacting molecules? If so, why then some bound proteins acquire large movements that may extend to distant locations from the binding site, instead of changing the local environment of the ligand? Certainly, the "lock and key" principle proposed by Fischer, cannot answer either of the two questions. Hence, two models were proposed to explain this inconsistency.

The first was Koshland's theory of "induced fit".[10] This theory answered the first question. Yes, it is the ligand-binding that changes the structure of the binding site in order to maximize its interactions and, hence, its binding affinity to the bound ligand. But the “induced fit” principle can explain the local movements around the ligand, including side-chain rotations or even tiny loop movements. However, it is hard to apply the same principle to account for larger dynamics that can rearrange considerable regions of the target. To explain this behavior, Monod, Wyman and Chaneux suggested their “MWC” model, or the “conformational selection model”.[11] Any particular unbound target conformation is in equilibrium with many other conformations. A ligand can only “select” and bind to one of these structures, hence the name, “conformational selection”. Figure 3-4 describes the difference between the three proposed concepts of target-ligand binding.

3.3.2 Single Structure Vs. Multiple Structures During Docking

Introducing protein dominant dynamics during docking experiments can indentify new scaffolds that exploit the newly opened (or closed) regions in the binding site. The importance of protein flexibility during docking was demonstrated in many studies since the late 1990 and shown to have a remarkable influence on the final results. For example, Bouzida et al.(1999) investigated the docking of sb203386 and skf107457 inhibitors to HIV-1 protease using MC simulations.[12] Their work compared docking against a single protein conformation to using an ensemble of protein structures. It was not a surprise to conclude that using multiple conformations of the same target was better than using a single structure. In another study by Murray et al. (1999) only 49% of the ligands were cross-docked correctly to another target that was co-crystallized with a different ligand. This showed that inducing small movements of the side chains (not only the backbone) in the binding site resulted in large variations in the predicted binding affinities. These early studies have drawn the attention of docking research groups to the importance of target flexibility and motivated them to implement various techniques to include this factor in the context of docking.

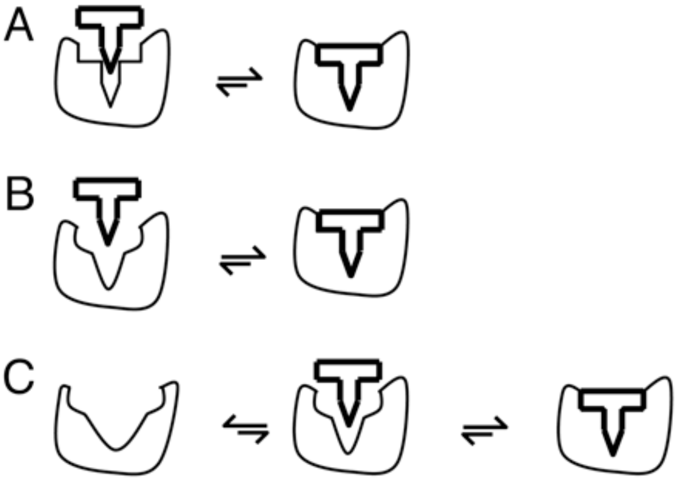


Figure 3-4: Three concepts of target-ligand binding.

A) Lock and key. B) Induced fit. C) Conformational selection followed by induced fit. (Adopted from Tobi et al.[13])

3.3.3 Hybrid MD-Docking Methods

One way to accommodate receptor flexibility and allow for using of more accurate scoring techniques is to implement a hybrid between docking and MD simulations. Originally, the use of MD simulations in VS studies was intended to create a set of receptor conformations.[14] However, it was always debatable whether to use structures derived from MD simulations or NMR data. For example, Philippopoulos et al. suggested NMR structures as the most effective source for protein conformations.[15] A set of 15 NMR conformations for ribonuclease HI was compared to a trajectory obtained from a 1.7ns MD simulation. The NMR data explored the conformational space of the protein more efficiently than the conventional MD simulation. In spite of their findings, it should be noted that Philippopoulos used a standard single trajectory MD simulations for a relatively short simulation time. As was noted before, in generating such ensembles, one should employ multiple trajectory MD simulations (REMD), or run the simulation over longer times. In this thesis, we think that if a practical ensemble of NMR structures exist, one should consider using it all, instead of running long MD simulation. However, if the VS exercise is departing from a single X-ray crystal structure, it is important to generate such ensemble using MD simulation.

In this context, a successful approach, reported by McCammon and his team, is the relaxed complex scheme (RCS).[16],[17] The method, illustrated in Figure 3-5, forms the foundation of the VS protocol presented in this thesis. In the RCS approach, all-atom MD simulations (e.g., 2-5 ns simulation) are applied to explore the conformational space of the target, while docking is subsequently used for the fast screening of drug libraries against an ensemble of receptor conformations. This ensemble is extracted at predetermined time intervals (e.g., 10 ps) from the simulation, resulting in hundreds of thousands of protein conformations. Each conformation is then used as a target for an independent docking experiment.

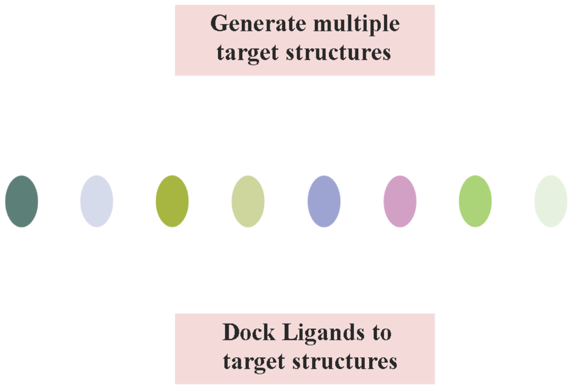

Generate multiple target structures
Dock Ligands to target structures

Figure 3-5. Basic idea behind the relaxed complex scheme developed by McCammon et al.

The RCS methodology has been successfully applied to a number of cases. An excellent example is an HIV inhibitor, raltegravir,[18,19] which became the first FDA approved drug targeting HIV integrase. MD simulations played a significant role in discovering a novel binding site, and compounds that can exchange between the two binding sites have formed a new generation of HIV integrase inhibitors. Other successful examples include the identification of novel inhibitors for acetylcholine binding protein,[20] RNA-editing ligase 1,[21] the influenza protein neuraminidase,[22] Trypanosoma brucei uridine diphosphate galactose 4'-epimerase,[23]and many others described in the literature.

These applications employed alternative ways to solve two main problems with the method, namely, reducing the number of extracted target conformations and deciding on how to select the final set of hits after carrying out the screening process. For the first problem, a number of studies suggested extracting the structures at larger intervals of the MD simulation (e.g. every 5ns or so),[20] condensing the structural ensemble generated from MD simulations using QR factorization,[21] or clustering the MD trajectory using root-mean-square-deviation (RMSD) conformational clustering,[22,23] On the other hand, to rank the screened compounds and suggest a final set of top hits, some studies used only docking predictions,[20,21,22]while others suggested (as in this thesis) using a more accurate scoring method (e.g. MM/PBSA (Molecular Mechanics/Poisson Boltzmann Surface Area)) to refine the final selected hits.[24] All of these approaches, similar to the work presented here, were aiming at keeping the balance between significantly reducing the number of target structures and, in the meantime, retaining their capacity to describe the conformational space of the target.

The following steps represent the approach that was used to put together and improve the RCS method in this thesis. Our implementation follows the same guidelines of the method. We first use MD simulations and generate large enough trajectories that can progress through the phase space of the binding site. The length of the MD simulations (usually in the order of 100ns) is determined through applying metrics that employ principal component analysis (PCA). Once the trajectory reaches an adequate sampling of target conformations, clustering analysis extracts representative structures that describe the dominant dynamics of the binding site. The extracted structures are then used as rigid targets to screen the whole library of compounds and suggest models for the most preferred ligand-protein complexes, hence, utilizing the "conformational sampling" model. These bound structures are then solvated and used to run all-atom MD simulations to relax the two molecules and generate new trajectories that represent their "induced fit" models. MM-PBSA method finally ranks the newly generated structures and suggests a set of top hits for experimental testing.

3.3.4 Principal Component Analysis And Convergence Of Sampling

A typical MD trajectory displays how the atomistic Cartesian coordinates are traveling in time. Although the duration of the whole trajectory is very short (at best, in the order of 100s of ns) compared to real life biological dynamics, it involves a huge number of snapshots that contain a mixture of fast and slow modes of motion. It is impossible to segregate or understand these mixed dynamics through simple analysis (e.g. visual inspection). However, covariance, or principle component, analysis (PCA) can break up these two types of motions and extract the essential dynamics (ED) spanned by the protein structure. These essential dynamics are the collective movements that are directly linked to the function of the protein and are essential for its role. In fact, PCA transforms the original space of correlated variables from a large MD simulation into a reduced space of independent variables.[25,26] For a typical protein, the system's dimensionality is thereby reduced from tens of thousands to fewer than fifty degrees of freedom.

To perform PCA for a subset of N atoms, the entire MD trajectory is RMSD fitted to a reference structure, in order to remove all rotations and translations. The covariance matrix can then be calculated from their Cartesian atomic coordinates as (see Appendix A for more details):

$$\sigma_{ij} = \left\langle \left(r_i - \left\langle r_i \right\rangle\right)\left(r_j - \left\langle r_j \right\rangle\right) \right\rangle \qquad \text{EQ. 3-1}$$

where r_i represents the three Cartesian coordinates (x,y,z) and the eigenvectors of the covariance matrix constitute the essential vectors of the motion. It is generally accepted that the larger an eigenvalue, the more important its corresponding eigenvector in the collective motion. PCA can also be employed to predict the completeness of sampling during the MD simulation. This step is critical and was used in this study to answer a very important question: when to stop the simulation and start extracting the dominant conformations of the protein? In this, we follow a method proposed by Hess[27] that divides an MD trajectory into separate parts, and their normalized overlap (see appendix A for details) is calculated using the covariant matrices for each pair of parts:

$$\text{Normalized overlap } (C_1, C_2) = 1 - \frac{\sqrt{tr\left(\left(\sqrt{C_1} - \sqrt{C_2}\right)^2\right)}}{\sqrt{tr(C_1) + tr(C_2)}} \qquad \text{EQ. 3-2}$$

where C_1 and C_2 are the covariant matrices, and the symbol *tr* denotes the trace operation. If the overlap is 0, then the two sets are considered to be orthogonal, whereas an overlap of 1 indicates that the matrices are identical. In this context, for the three targets studied in this thesis, the individual whole trajectories were divided into three parts and the normalized overlap between each pair was calculated to determine the completeness of sampling.

3.3.5 Iterative Clustering To Extract Dominant Conformations

Once a sufficient sampling is confirmed through the aforementioned PC calculations, clustering analysis are then used to extract a set of target structures that represent its dominant conformations. Unfortunately, there is no universally accepted clustering algorithm or parameters that can be used to extract all of the information contained within the MD simulation. However, recent studies suggest that a number of clustering algorithms, such as average-linkage, means and self-organizing maps (SOM) can be used accurately to cluster MD data.[28] In this work, the clustering quality was anticipated by calculating a number of clustering metrics. These metrics can reveal the optimal number of clusters to be extracted and their population size. These are the Davies-Bouldin index (DBI)[29] and the "elbow criterion".[28] A high-quality clustering scheme is correlated with high DBI values. On the other hand, using the elbow criterion, the percentage of variance explained by the data is expected to plateau for cluster counts exceeding the optimal number of clusters. Using these metrics, by varying the number of clusters, one should expect for adequate clustering, a local minimum for DBI and a horizontal line for the percentage of variance explained by the data. Figure 3-6 describes an example of such calculations.

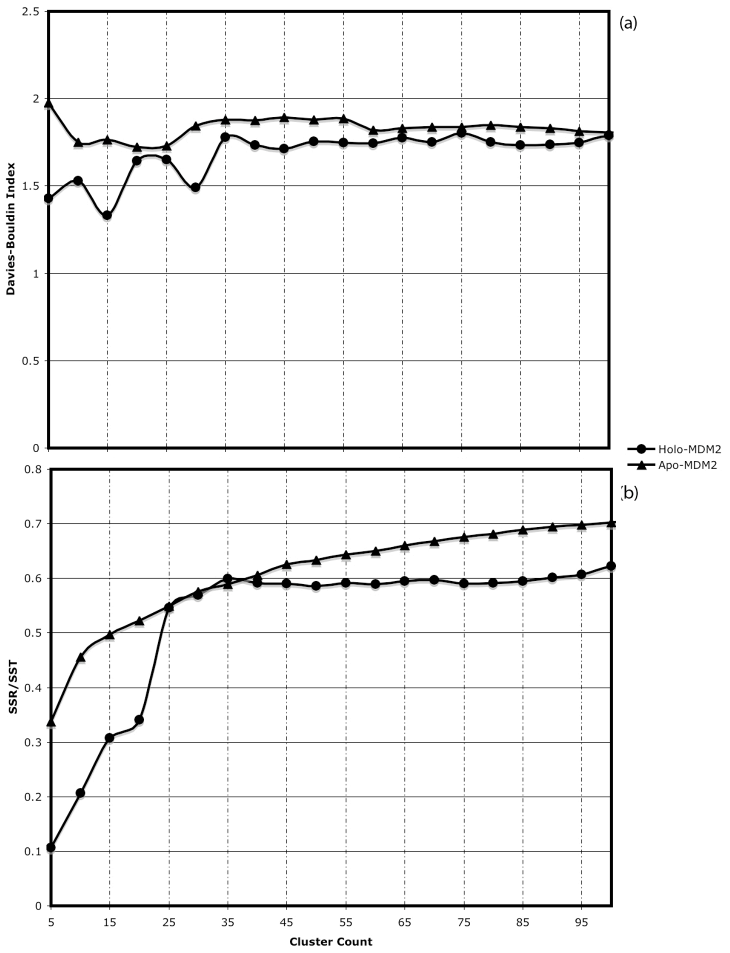


Figure 3-6: Clustering Analysis.

The DBI and SSR metrics against the number of clusters for the MDM2 target (chapter 7). A high-quality clustering is obtained when a local minimum in DBI correlates with saturation in the SSR/SST ratio. This is clear at cluster count of 60 for the apo-structure and 30 clusters for the holo-structure.

This work employs an iterative clustering algorithm using the abovementioned hypothesis. The procedure is implemented as an in-house code using the PTRAJ utility of AMBER10. A modified version of the code is also used to cluster the docking results (see section 3.4.2). MD trajectories' clustering runs the average-linkage algorithm for a number of clusters ranging from 5 to 150 clusters. Structures are extracted at 2 ps intervals over the entire simulation times. In order to remove the overall rotation and translation, all C_{α} atoms are fitted to the minimized initial structure. RMSD-clustering is performed on the residues contained in the investigated binding sites. These residues are clustered into groups of similar conformations using the atom-positional RMSD of the entire amino acid, including side chains and hydrogen atoms, as the similarity criterion. The centroid of each cluster, the structure having the smallest RMSD to all members of the cluster, is chosen as the cluster representative structure and the most dominant structures are used as rigid templates for the ensemble-based docking experiments (see below).

3.4 Docking Ligands To The Ensemble Of Target Structures

As stated above, the outcome of the iterative clustering step is an ensemble of protein structures that are used as targets for docking. The main docking program that was used throughout this thesis is AutoDock version 4.[30] AutoDock is one of the most popular docking packages that utilize different conformational search methods, including Simulated Annealing (SA), traditional Genetic Algorithm (GA), and Lamarckian Genetic Algorithm (LGA). Here, we used the LGA approach. The LGA method is a hybrid between classical GA (described in section 2.3.1.1.2) and local minimization search. The approach is well-described in the original paper by Morris et al.[30] and is compared to the classical GA below.

3.4.1 Classical Vs. Lamarckian GA

AutoDock represents the state of the ligand within the binding site (i.e. position and orientation) with a set of parameters and defines them as the ligand's "state variables". These state variables are the genome of the ligand. Genomic state variables are subject to change by mutations or crossover during the evolution "docking" process. The atomic coordinates of the ligand represent its phenotypic features. LGA is based on Jean Baptiste de Lamarck's ideas. That is, phenotypic features that are gained during an individual's lifetime can be inherited in its genome and passed to the newly generated offspring. This is the main difference between classical and Lamarckian GAs. While the former explores only the phenotype side of the problem and has a one-way translation (mapping) from genotype codons into phenotype features (see Figure 3-7), the latter can search and exchange information between the two realms. LGA follows the same concepts of the classical algorithm except for two essential parts. First, when evaluating the fitness (scoring) of the individual conformations, classical GA uses only the scoring function (**Error! Reference source not found.**), while LGA, do that by first evaluating the scoring function, minimizing the coordinates of the ligand and search for the closest minimum. Second, this local search takes place in the phenotypic features of the ligand "coordinates" and then translates it into a new set of genomic information "state variables" which can be inherited in the new generation of offspring, if the resulting conformation is strongly fitted to the environment.

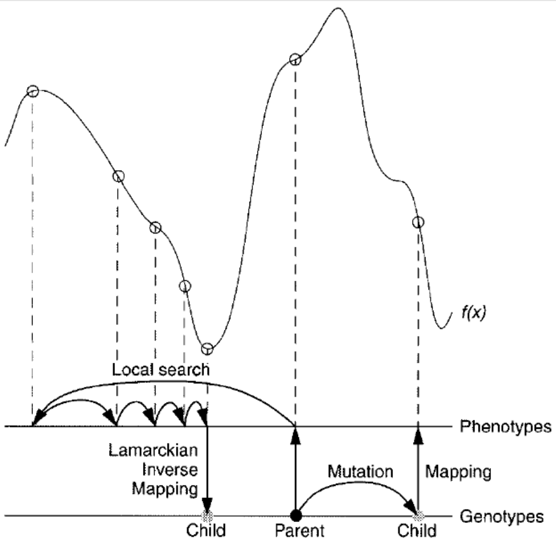


Figure 3-7: Difference between classical and Lamarckian genetic algorithms. (adopted from Garrett et al.[30])

3.4.2 Automated Clustering of Docked Poses

The previously described virtual screening experiments involve millions of conformations of each ligand bound to its target. AutoDock can cluster these output poses into subgroups depending on their RMSD values referred to a reference structure. Although this approach is widely used, however, similar to clustering of MD trajectories (section 3.3.5), the number of clusters and the population size for each cluster depends heavily on the RMSD cut-off used. Consequently, it is impossible to expect the optimal cut-off for the RMSD in order to produce a clustering pattern with the highest confidence. This motivated us to use an alternative approach when clustering the docked ligand structures. Here, we extended and automated the clustering methodology that was used in section 3.3.5 to couple the elbow criterion,[31] with the clustering module of PTRAJ (a well-known utility in AMBER10). This method exploits the fact that the percentage of variance exhibited by the data (λ), is expected to plateau for cluster counts exceeding the optimal number.

The percentage of variance is defined by:

$$\lambda = \frac{SSR}{SST}$$

EQ. 3-3

where (SSR) is the sum-of-squares regression from each cluster summed over all clusters and (SST) is the total sum of squares. Here, we used the SOM algorithm to cluster the docking results. This modified clustering program increases the number of clusters required until the percentage of variance exhibited by the data (λ) plateaus. The convergence of clustering can be determined by calculating the first and second derivatives of the percentage of variance with respect to the clusters number ($\frac{d\lambda}{dN}$ and $\frac{d^2\lambda}{dN^2}$) after each attempt to increase the cluster counts. The clustering process then stops at an acceptable value for these derivatives that is close to zero. In this way, the clustering procedure depends only on the system itself and adjusts itself to arrive at the optimal clustering pattern for that specific system.

3.4.3 Preliminary Ranking of Docking Results

The VS protocol then sorts the docking results by the lowest binding energy of the most populated cluster. The compounds can also be ranked using their weighted average binding energies according to the following formula:

$$\text{Weighted Average Binding Energy (WABE)} \;=\; \sum_{i}^{M} \text{percent distribution}(i) \times \text{binding energy}(i) \qquad \text{EQ. 3-4}$$

where *i* is the index number of each ensemble cluster whose percent distribution sums up to 100% and *M* is the number of different structures included in the target ensemble. The VS protocol only considers a compound among the top hits if the most populated cluster from any of the VS experiments includes at least 25% of all docked conformations. The top N hits of the combined docking runs construct an irredundant set of promising compounds that are used for further analysis. In this work, a typical preliminary set (N) includes from 200 to 500 compounds.

3.5 Visual Inspection And Selection Of A Focused Set Of Hits

Visual inspection to the preliminary set of hits is necessary before proceeding to the later computationally rigorous steps. Although this step involves more human intervention, it assures the quality of the docking results, which are not precise in terms of ranking or the final suggestion of binding geometries. Moreover, even with the exhaustive preparation of the VCC in the initial steps of the screening process, some compounds may include mistakes that were not noticed in the earlier steps. It is also important to perform this visual inspection step as a post-filtering approach to select compounds that have specific interactions with the target. For example, if a hydrogen bond(s) with a particular residue(s) or a hydrophobic interaction with a portion of the binding site is required. In this work and during this stage, it was important to return back to the original clustering results for a number of ligands and select additional binding modes that were ignored during the filtering step but showed promising binding energies. These post-filtering investigations act as a refining step for the previous high throughput docking steps in order not to waste the computational resources on improbable successful hits during the next steps.

3.6 Molecular Dynamics Simulations on Selected Hit-target Complexes

As was mentioned in the previous sections, there are many factors (e.g. water, protein flexibility, etc.) that are not well characterized within the docking context. During this step, the VS protocol aims at accounting for these factors through performing MD simulations. Each simulation starts from the final docked structure. The important aspect at this stage is the solvation of the docked models. Based on our knowledge from this work, for a number of cases, water molecules were not only involved in solvation/desolvation of the protein-ligand complexes, but also they mediated their interactions and helped in generating more suitable binding modes. MD simulations relax the structures, rearrange water and ion molecules and generate trajectories that are used during the next step for binding energy calculations. The output from this step is a set of snapshots representing the trajectory of the MD simulations for each complex. Although this procedure requires extensive computational resources, it tends to improve the protein–ligand interactions and enhance their molecular complementarity. In fact, during this stage it simulates their induced fit model and allows the binding-site environment to inflict the requirements to be met by the selected ligand. Moreover, the stability of the complexes over the simulation time is a direct measure of the consistency of binding, since improperly docked structures are expected to produce unstable trajectories. The MD simulation protocol depends on the system studied and the details of these calculations and their set up will be presented in the next chapters.

3.7 Rescoring of Hits Using the MM-PBSA Method

Besides using MD simulations to refine the docked structures, another essential constraint for a successful VS experiment is to accurately predict their binding energies. To correctly fulfill this task, we need to move far from the simple docking scoring methods. However, we are also restricted by the need to have a fairly fast method that can be applied to many systems at a reasonable time. At this stage, it is also the time to consider the factors that were ignored or mistreated during the initial docking scoring, such as solvation and entropic terms. In this context, the VS protocol utilizes a fast and efficient scoring method to suggest the final ranked set of top hits. It is the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) technique. The method was initially proposed by Kollman et al.[32] and it combines molecular mechanics with continuum solvation models. The method has been extensively tested on many systems and shown to reproduce, with an acceptable range of accuracy, experimental binding data. It was also validated as a VS refining tool and revealed excellent results in predicting the actual binding affinities and in discriminating true binders from inactive (decoy) compounds [33]. Its main advantages are the lack of adjustable parameters and the option of using a single MD simulation for the complete system to determine all energy values.

In this work we used the MM-PBSA method as implemented in AMBER. The total free energy is the sum of average molecular mechanical gas-phase energies (E_{MM}), solvation free energies (G_{solv}), and entropy contributions ($-TS_{solute}$) of the binding reaction:

$$G = E_{MM} + G_{solv} - TS_{solute} \qquad \text{EQ. 3-5}$$

The total molecular mechanical energies can be further decomposed into contributions from electrostatic (E_{ele}), van der Walls (E_{vdw}) and internal energies (E_{int}):

$$E_{MM} = E_{ele} + E_{vdw} + E_{\text{int}} \qquad \text{EQ. 3-6}$$

Furthermore, the solvation free energy can be expressed as a sum of non-electrostatic and electrostatic contributions:

$$\Delta G_{solv} \approx \Delta G_{solv}^{nonele} + \Delta G_{solv}^{ele} \qquad \text{EQ. 3-7}$$

The non-electrostatic part was approximated by a linear function of the (SASA). That is:

$$\Delta G_{solv}^{nonele} = \gamma \times SASA, \ \text{where } \gamma = 7.2 \text{ cal/mol/A}^2 \qquad \text{EQ. 3-8}$$

where $\gamma = 7.2$ cal/mol/Å^2.

In this work, the molecular mechanical (E_{MM}) energy of each snapshot is calculated using the SANDER module of AMBER with all pair-wise interactions included using a dielectric constant (ε) of 1. The solvation free energy (G_{solv}) is estimated as the sum of electrostatic solvation free energy, calculated by the finite-difference solution of the Poisson–Boltzmann equation (see appendix B) in the Adaptive Poisson-Boltzmann Solver (APBS) program. The non-polar solvation free energy is directly proportional to the solvent-accessible surface area (SASA) of the target. The solute entropy is approximated using normal mode analysis. As has been reported earlier by other groups, the most computationally demanding step is the calculation of the solute entropy. Although this component can be neglected if only relative binding (relative ranking) of compounds is required,[32]for all of the three studied systems we calculated the entropy contribution to the binding energies. The binding free energy can be approximated by:

$$\Delta G^{o} = \Delta G_{gas}^{protein-ligand} + \Delta G_{solv}^{protein-ligand} - \{\Delta G_{solv}^{ligand} + \Delta G_{solv}^{protein}\} \qquad \text{EQ. 3-9}$$

Here, ($\Delta G_{gas}^{protein-ligand}$) represents the free energy per mole for the non-covalent association of the ligand-protein complex in vacuum (gas phase) at 310 K, while ($-\Delta G_{solv}$) stands for the work required to transfer a molecule from its solution conformation to the same conformation in vacuum at 310 K (assuming that the binding conformation of the ligand-protein complex is the same in solution and in vacuum).

3.8 Conclusion

This chapter introduced the virtual screening workflow that was implemented in this thesis. The VS protocol prepares ligand collections for docking and extracts dominant confirmations of the target through MD simulations combined with clustering analysis and PCA. The VS algorithm introduced in this chapter tried to improve over the well-known RCS technique. The main improvements are the reduction of the number of target structures used and the employment of a more accurate scoring method than that of AutoDock. In the next chapters, the VS protocol introduced here will be applied to different problems.

3.9 Bibliography

1. http://www.rcsb.org/pdb/home/home.do.
2. http://www.bch.msu.edu/~kuhn/software/consolv/.
3. Li, H.; Robertson, A. D.; Jensen, J. H., Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **2005,** *61* (4), 704-21.
4. http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html.
5. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J., DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **2006,** *34* (Database issue), D668-72.
6. Irwin, J. J.; Shoichet, B. K., ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **2005,** *45* (1), 177-82.
7. http://chimiotheque-nationale.enscm.fr/?lang=fr.
8. http://openbabel.org.
9. http://www.schrodinger.com/.
10. Koshland, D. E., Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci U S A* **1958,** *44* (2), 98-104.
11. Monod, J.; Wyman, J.; Changeux, J. P., On the Nature of Allosteric Transitions: A Plausible Model. *J Mol Biol* **1965,** *12*, 88-118.
12. Gennady, M. V.; Paul, A. R.; Djamal, B.; Sandra, A.; Anthony, B. C.; Stephan, T. F.; Daniel, K. G.; Veda, L.; Brock, A. L.; Tami, M.; Peter, W. R., Parallel simulated tempering dynamics of ligand±protein binding with ensembles of protein conformations. *Chemical Physics Letters* **2001,** *337*.
13. Tobi, D.; Bahar, I., Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc Natl Acad Sci U S A* **2005,** *102* (52), 18908-13.
14. (a) Broughton, H. B., A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening. *Journal of molecular graphics & modelling* **2000,** *18* (3), 247-57, 302-4; (b) Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A., Developing a dynamic pharmacophore model for HIV-1 integrase. *Journal of medicinal chemistry* **2000,** *43* (11), 2100-14.
15. Philippopoulos, M.; Lim, C., Exploring the dynamic information content of a protein NMR structure: comparison of a molecular dynamics simulation with the NMR and X-ray structures of Escherichia coli ribonuclease HI. *Proteins* **1999,** *36* (1), 87-110.
16. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A., Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J Am Chem Soc* **2002,** *124* (20), 5632-3.
17. Amaro, R. E.; Baron, R.; McCammon, J. A., An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J Comput Aided Mol Des* **2008,** *22* (9), 693-705.
18. Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A., Discovery of a novel binding trench in HIV integrase. *J Med Chem* **2004,** *47* (8), 1879-81.
19. Markowitz, M. N., B.Y. Gotuzzo, F.; Mendo, F.; Ratanasuwan, W.; Kovacs, C.; Zhao, J.; Gilde, L.; Isaacs, R.; Teppler, H., Potent antiviral effect of MK-0518, novel HIV-1 integrase inhibitor, as part of combination ART in treatment-naive HIV-1 infected patients. *16th International AIDS Conference, Toronto, Canada.* **2006**.
20. Babakhani, A.; Talley, T. T.; Taylor, P.; McCammon, J. A., A virtual screening study of the acetylcholine binding protein using a relaxed-complex approach. *Comput Biol Chem* **2009,** *33* (2), 160-70.
21. Amaro, R. E.; Schnaufer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A., Discovery of drug-like inhibitors of an essential RNA-editing ligase in Trypanosoma brucei. *Proc Natl Acad Sci U S A* **2008,** *105* (45), 17278-83.
22. Durrant, J. D.; McCammon, J. A., Potential drug-like inhibitors of Group 1 influenza neuraminidase identified through computer-aided drug design. *Comput Biol Chem* **2010,** *34* (2), 97-105.

23. Durrant, J. D.; Urbaniak, M. D.; Ferguson, M. A.; McCammon, J. A., Computer-aided identification of Trypanosoma brucei uridine diphosphate galactose 4'-epimerase inhibitors: toward the development of novel therapies for African sleeping sickness. *J Med Chem* **2010,** *53* (13), 5025-32.

24. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A., The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers* **2003,** *68* (1), 47-62.

25. Garcia, A. E., Large-amplitude nonlinear motions in proteins. *Phys Rev Lett* **1992,** *68* (17), 2696-2699.

26. Amadei, A.; Linssen, A. B.; Berendsen, H. J., Essential dynamics of proteins. *Proteins* **1993,** *17* (4), 412-25.

27. Hess, B., Convergence of sampling in protein simulations. *Phys Rev E Stat Nonlin Soft Matter Phys* **2002,** *65* (3 Pt 1), 031910.

28. Shao, J.; Tanner, S.; Thompson, N.; Cheatham, T., Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of Chemical Theory and Computation* **2007,** (3), 2312.

29. Davies DL; Bouldin DW, A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intelligence* **1979,** *1*, 224.

30. Garrett MM; David SG; Robert SH; Ruth H; William EH; Richard KB; Arthur JS, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem.* **1999,** *19*, 1639.

31. Shao J; Tanner SW; Thompson N; Cheatham TE, Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. Journal of Chemical Theory and Computation. **2007,** *3*, 2312.

32. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd, Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* **2000,** *33* (12), 889-97.

33. (a) Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M., Structure-based discovery of inhibitors of thymidylate synthase. *Science (New York, N.Y* **1993,** *259* (5100), 1445-50; (b) Schneider, G.; Bohm, H. J., Virtual screening and fast automated docking methods. *Drug discovery today* **2002,** *7* (1), 64-70; (c) Abagyan, R.; Totrov, M., High-throughput docking for lead generation. *Current opinion in chemical biology* **2001,** *5* (4), 375-82.

Chapter 4: Discovering Inhibitors for the ERCC1-XPA interaction: The First Puzzle[(1)]

4.1 Introduction

Nucleotide excision repair (NER) can be considered as an old friend, but is in fact a new enemy in the context of cancer. In normal cells, NER removes many types of DNA lesions, protecting cell integrity.[1] However, in cancer cells exposed to DNA damaging agents that distort the DNA helix or form bulky injuries to the genome, NER comes into play and removes the damage, thus protecting cancer cells from death.[1-2] A striking example of this mechanism is represented by the use of platinum compounds such as cisplatin, the backbone for many treatments of solid tumors including testicular, bladder, ovarian, head and neck, cervical, lung and colorectal cancer [3]. It has been demonstrated that NER is the major DNA repair mechanism that removes cisplatin-induced DNA damage, and that resistance to platinum-based therapy correlates with high expression of specific elements of the NER machinery.[4] Accordingly, a novel strategy to reverse resistance and potentiate the efficacy of cisplatin is to regulate the NER pathway. However, as we will see below, the NER pathway is very complex and many proteins are involved and necessary to fulfill its function. Nevertheless, there are many clues described in the literature that suggest a few of these proteins as targets for chemotherapy. Therefore, before describing the targets that we chose in this thesis to selectively inhibit the NER pathway and, hence, sensitize tumor cells to platinum chemotherapy, a brief listing of the essential steps comprising the NER pathway will be given below.

4.1.1 The NER pathway

The Nucleotide excision repair, shown in Figure 4-1, occurs in a stepwise mechanism and involves more than 30 different proteins. It is a “cut-and-paste” mechanism that replaces a ~30 nucleotide DNA strand that contains the lesion with a correct one. The pathway has been

[(1)] A version of this chapter has been published in Barakat KH, Torin Huzil J, Luchko T, Jordheim L, Dumontet C, Tuszynski J. J Mol Graph Model. 2009 Sep; 28(2): 113-30.

extensively studied so that all the genes that are involved in it have been cloned and expressed as recombinant proteins.

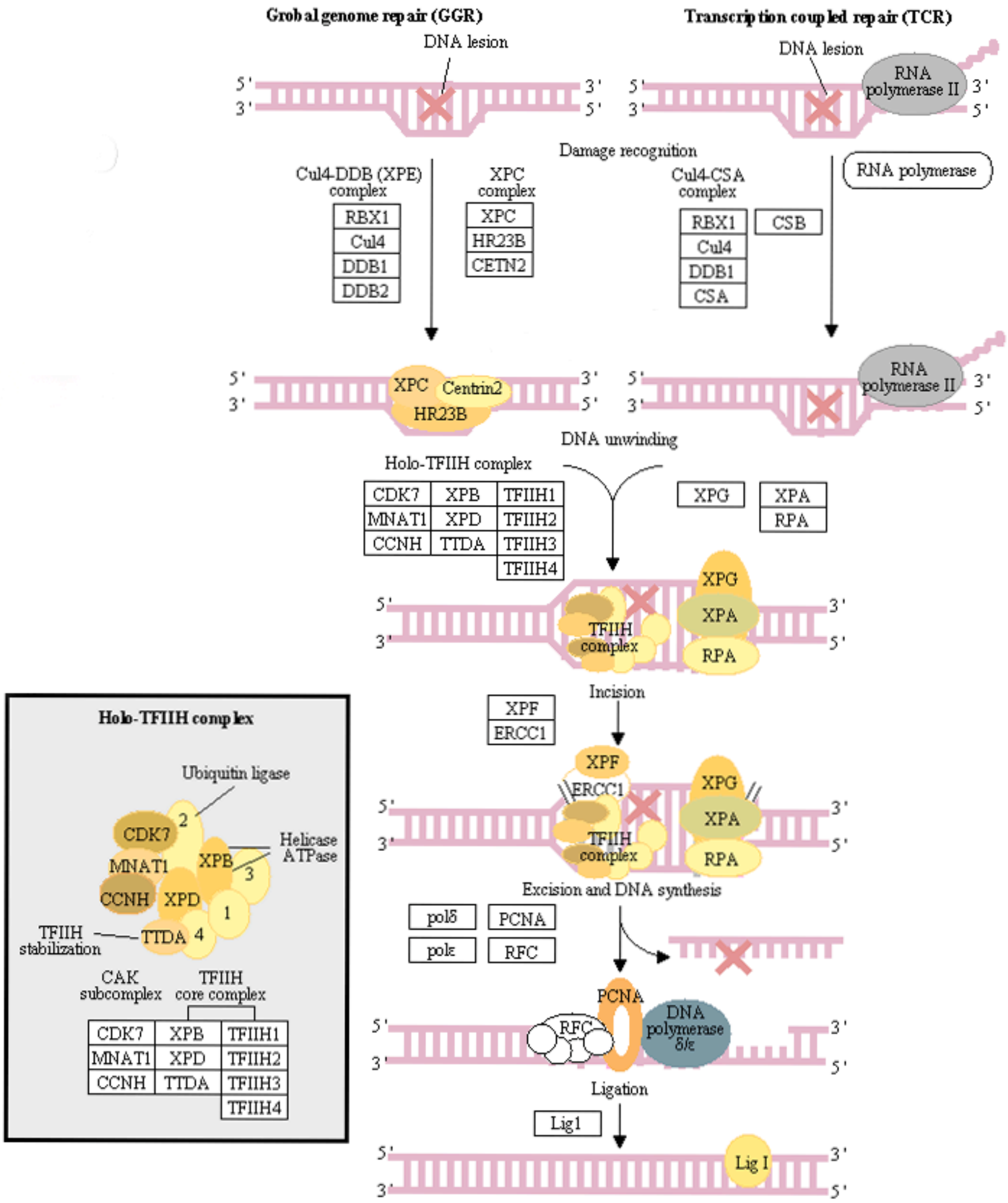


Figure 4-1: Steps of the nucleotide excision repair pathway.

See text for details (adopted from the KEGG database.[5])

The main players within NER includes the seven Xeroderma Pigmentosum (XP) complementation groups, XPA to XPG proteins; the Excision Repair Cross Complementing group 1 protein (ERCC1); the human Homolog of yeast RAD23 (hHR23B), the Replication Protein A (RPA), the subunits of Transcription Factor that possess Helicase activity (TFIIH), and the Cockayne Syndrome proteins A and B (CSA and CSB).[6] Depending on the location of the DNA damage within the genome, one can recognize two NER sub-pathways. First is the transcription-coupled repair (TCR-NER), if the DNA damage is located within the actively transcribed genes of the genome. The second is the global genome repair (GGR-NER), if the damage is located within the whole genome. The two types are thought to be identical except for the initial damage recognition step. The two mechanisms involves 5 sequential steps.[7]

The foremost step is the detection of the damage. As mentioned above, the recognition step is the only difference between TCR and GGR. In the GGR subpathway the XPC-hHR23B-XPE complex continuously scans the genome for bulky DNA damages until it recognizes a lesion and, consequently, initiates the rest of the NER sequence. On the other hand, a stalled RNAPII and Cockayne syndrome proteins, CSA and CSB, recognize the damage and activate the TCR-NER pathway. Once the damage is recognized the second step starts by recruiting the TFIIH complex in order to unwind the DNA helix surrounding the lesion. TFIIH is composed of two major sub-complexes. The core is formed from the association of a large number of proteins including XPB, XPD, p62, p52, p44, p34 and p8. The rest of TFIIH is the cdk-activating kinase sub-complex, which contains cdk7, cyclin H and MAT1. Interestingly, TFIIH possesses both 3'-5' and 5'-3' helicase activities through the two ATP-dependent helicases XPB and XPD respectively.[8] It opens the DNA structure forming a ~30 base pair bubble around the lesion. The two proteins RPA and XPA stabilize the opened DNA structure and recruit the two endonucleases that are necessary for the subsequent incision step. The interaction of XPA with the 34-kDa subunit of RPA (RPA34) activates XPA to recruit the other components of NER.

The Damaged strand-incision is the rate-limiting step for the whole pathway. The two endonucleases XBG and XPF-ERCC1 cut the two ends of the strand that contains the damage. The correct location of XPA is crucial for the recruitment of the XPF-ERCC1 heterodimer endonuclease. XPG cuts the 3' end of the damage, while XPF-ERCC1 cuts the 5' end.[9] The damaged strand is then released. DNA polymerases fill the single strand gap using the complementary intact strand as a template and DNA ligase I closes the 3' nick as a final step.[7]

4.1.2 ERCC1 Over expression Correlates With Cisplatin Resistance

ERCC1 is a 33-kDa protein that forms a tight heterodimer endonuclease complex with XPF. As described above, this endonuclease cleaves the DNA strand at the phosphodiester bonds on the 5′ side of the damage. It is important to mention that the ERCC1-XPF complex has additional functions in other DNA repair pathways including interstrand cross-link repair, double-strand break repair, and homologous recombination. Many studies has shown a considerable correlation between resistance to cisplatin and the over expression of ERCC1.[10] This spectacular conclusion has been reached from several independent clinical trial investigations on ovarian,[11] colorectal,[12] and non–small cell lung cancer.[13] For example, a study on ~750 patients who suffer from late stages of lung cancer revealed that patients with low levels of ERCC1 and who received platinum therapy had better survival rates than those with the same levels of the protein but did not receive the platinum treatment.[14] A more recent study on 444 patients who experienced non-small lung cancer showed that non-platinum-containing chemotherapy is more effective than platinum-based therapy on patients with high ERCC1 levels.[15] Very recently, Stefanie and coworkers performed a retrospective study investigating the correlation of ERCC1 expression with patients' survival in ovarian cancer after platinum-based treatment.[16] Their work revealed that patients with ERCC1-negative ovarian cancer had significant better survival rates than those with ERCC1-positive. They concluded that ERCC1 protein over expression was a marker for poor survival of high-grade ovarian cancer even in patients operated with residual disease. All of these investigations lead to the conclusion that ERCC1 is not only a gene that is usually activated when utilizing platinum-based therapy but also it may act as a predictive criterion for who could benefit from platinum treatments.[17,18] This latter role of ERCC1 as a biomarker is important because it can guide clinicians in their therapeutic decision-making and select the best treatment approach for a particular group of patients.

4.1.3 The ERCC1-XPA Interaction Is Essential For a Functional NER Pathway

Regardless of the type of NER that is initiated, the XPA protein is equally essential to fulfill both pathways.[19] It plays a vital role in DNA lesion recognition and attraction of many other NER repair proteins. For example, prior to the incision step in NER, the ERCC1-XPF endonuclease is recruited to the damaged DNA site through a secondary interaction between ERCC1 and XPA.[20,21,22] Therefore, this protein-protein interaction is necessary for a functional NER mechanism. The NMR crystal structure was resolved by Tsodikov's group[23] and the critical residue-residue interactions were determined through our binding energy predictions (Figure 4-2) (see below). A 14-residue peptide from XPA that includes three essential consecutive glycines (residues 72–74) is buried within a hydrophobic cleft within the central domain of ERCC1. This peptide has two critical characteristics. First, it is necessary and sufficient for binding to ERCC1. Second, and more importantly, it can compete with the full-length XPA protein in binding to ERCC1 and disrupting NER in vitro.

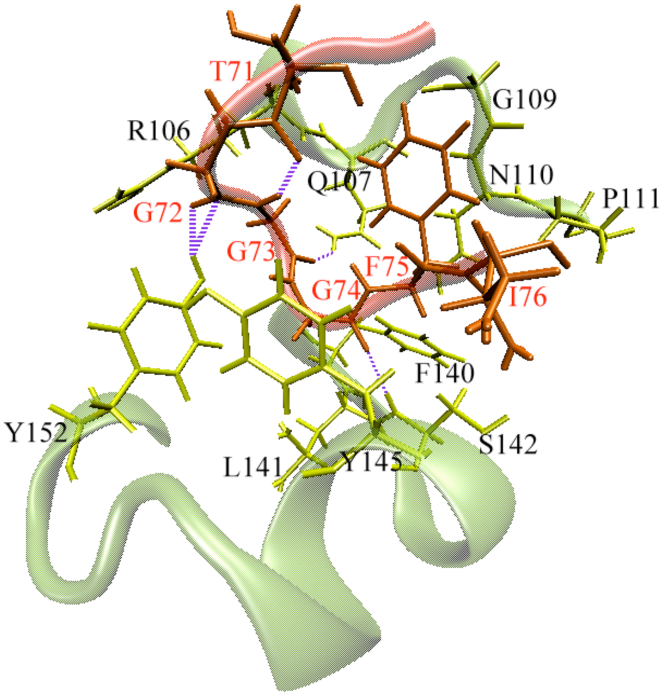


Figure 4-2: XPA-ERCC1 protein-protein interaction.

The binding between ERCC1 (teal) and XPA (red) is primarily mediated by 5 residues from XPA peptide, namely; G72, G73, G74, F75 and I76. On the other hand, the contribution from the ERCC1 binding site is distributed among 10 residues; R106, Q107, G109, N110, P111, F140, L141, S142, Y145 and Y152.

In a recent study, Barbara et al. reported mutations in the central domain of ERCC1 that had a significant impact on NER activity *in vitro* and *in vivo*.[24] These mutations occur at the XPA binding site within ERCC1, preventing the interaction between the two proteins. Due to these mutations, the ERCC1-XPF nuclease was not recruited to the damaged DNA sites after exposing cells to ultra violet (UV) radiation. Consequently, the last incision step that is performed by ERCC1-XPF was never completed leading to a dysfunctional NER mechanism in these cells and, hence, a hypersensitivity to UV radiation. These results agree with previous findings on the importance of XPA in NER, where no cellular function beyond NER has been observed for XPA.[25] Interestingly, these mutations didn't affect the activity of ERCC1-XPF in other DNA repair pathways leading to two distinctive conclusions. First, the XPA-ERCC1 interaction is only necessary for NER but not for other DNA repair pathways that ERCC1-XPF is important for their activity. Second, the involvement and recruitment of ERCC1-XPF to the different DNA repair pathways is coordinated through different and not overlapping protein-protein interactions mediated by ERCC1. Based on these findings, one can selectively disrupt the activity of ERCC1-

XPF within these DNA repair pathways by inhibiting its interactions with the recruitment factors to the damaged sites. These observations, coupled with the available crystal structure of this interaction make ERCC1 and XPA an extremely attractive target for computationally based development of small molecule inhibitors that are targeted for use in combination therapies involving cisplatin. And this protein-protein interaction was the one we chose in this thesis.

4.1.4 Known NER Inhibitors

Although the NER pathway has been recognized as one of the most important factors that increase the resistance against platinum-based therapy little work has been done on regulating its activity. Here, I will point out to the three major studies that identified inhibitors for the NER mechanism. First is the work done by Jean-Marc Barret et al. and their discovery of F11782.[26] Second are the findings of Jiang and Yang on the effects of the cell cycle checkpoint abrogator UCN-01 (7-hydroxystaurosporine) on NER.[27] Finally, is the DNA damaging agent Et743.[28]

F11782

Using the 3D(DNA Damaged Detection) assay (first proposed by Wood *et al.*[29] and then modified by Salles's team[30]), Barret et al. screened for NER inhibitors and identified F11782. The compound was already known as an inhibitor for both the topoisomerases II and I.[31] Moreover, F11782 did not show any activity toward other enzymes such as DNase I or T4 polynucleotide kinase, indicating that the compound targets one of the proteins that are involved in NER. Further investigations on F11782 limited its NER inhibitory activity to one of the earlier steps of the pathway, specifically either the helicase or the incision steps, with more preference to the incision step.[26]

UCN-01

Jiang and Yang analyzed the effects of UCN-01 (a well-known protein kinase C inhibitor and cell cycle checkpoint abrogator[32]) on the NER pathway.[27] Their findings showed that UCN-01 inhibited the repair of cisplatin-induced DNA damage both *in vitro* and *in vivo* and indicated that UCN-01 has a dramatic inhibitory effect on the interaction of NER proteins. The drug enhanced the activity of cisplatin only in NER-proficient cells, but not in the deficient ones. However, they did not report any direct binding of UCN-01 to any of these proteins and speculated that the observed inhibitory activity may result from UCN-01-mediated regulation of the signaling pathway that involves posttranslational modifications of repair proteins. Although Jiang and Yang attributed the lose of NER activity to an attenuation in the ERCC1-XPA protein-protein interaction, their carfeul and detailed binding analysis of the compound to the two proteins revealed that UCN-01 did not interact directly with either of them. However, in this work I used UCN-01 as a positive control, assuming it can bind to the XPA binding site within ERCC1, particularly because the drug can fit within the binding pocket despite its limited interactions with the protein (see Figure 4-3).

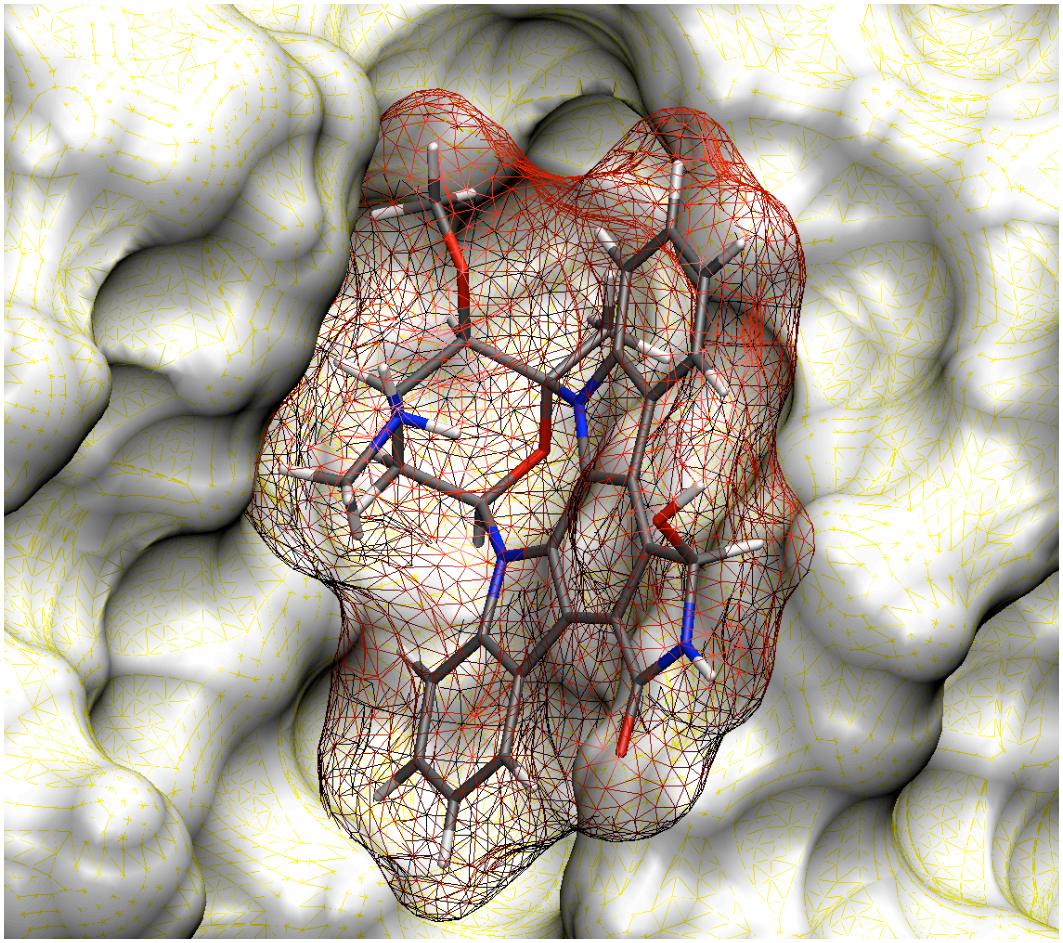


Figure 4-3 UCN01 within the ERCC1 pocket.

The docked structure can fit within the XPA binding site in ERCC1.

Ecteinascidin 743

A final compound that have been shown to interfere with NER is Ecteinascidin 743 (Et743).[28] Et743 is currently in phase II/III clinical development and its main mode of action is as a DNA damaging agent. The drug seems to specifically obstruct the TCR-NER sub-pathway, however, it does not act as an inhibitor for any of the proteins that are involved in the NER mechanism. A model proposed by Gregory et al. suggests that the DNA adducts formed by Et743 are more efficient than those of cisplatin in dealing with NER.[28] They suggest that the Et743-guanine adducts trap the TCR-NER pathway at the incision or ligation steps, preventing the pathway from being completed.

4.2 Results And Discussion

Here, we have utilized the RCS technique (section 3.3.3) to construct a dynamic pharmacophore model (section 2.2.2) targeting the ERCC1-XPA interaction. We utilized a minimized model of the XPA binding site within ERCC1 to employ flexible residue docking as implemented in AutoDock 4.0. This was then followed with RCS docking, where MD simulations and RMSD conformational clustering were used to generate a set of forty-four representative conformations of the binding site within ERCC1. AutoDock was then used to screen the National Cancer Institute Diversity Set (NCIDS) and DrugBank compounds (section 3.2.1) against a set of seven target conformations, composed of the six most dominant cluster-representative structures along with an equilibrated folded conformation for the binding site produced by employing principal component analysis on the ERCC1 trajectory. Top hits were rescored by docking them to the whole set of cluster-representative structures and ranked by their weighted average binding energy (section 3.4.3). The non-redundant hits from these screens were then used to identify a dynamic binding-site pharmacophore that target the ERCC1-XPA interaction. The pharmacophore model was then compared to docking results for the weak inhibitor of NER, UCN-01 (7-hydroxystaurosporine) (section 4.1.4).

4.2.1 Molecular Dynamics Simulations Of The ERCC1-XPA Interaction

To obtain a minimized model for library screenings and a set of flexible residues for docking, the central domain of ERCC1 was subjected to MD simulation, in both its free and XPA bound states.[23] The proper equilibration of these systems was essential in order to perform VS on a set of rigid receptor models that represent, approximatly the whole conformatioal space of the XPA binding site within ERCC1. Moreover, it is generaly required to start with adequately sampled, energetically minimized models in order to eliminate unfavorable atom contacts that may have been introduced as a result of crystal packing in the original structure.

4.2.2 PCA and Completeness of Sampling

As described in section 3.3.4, PCA is used to transform the original space spanned by the MD trajectory into a reduced set of independent variables comprising the essential dynamics of the system. PCA was performed over the entire 50 ns simulation using atoms comprising the 22 residues contained in the ERCC1 binding site with the backbone atoms RMSD fitted to the minimized crystal structure. Covariant analysis of the trajectories from the ERCC1-free MD simulations, successively divided into thirds, was performed using the same procedure used for the PCA (see Methods). Normalized overlaps calculated between each of these thirds are reported in Table 4-1. The high overlap between the thirds indicates that each part of the simulation samples approximately the same conformational space, and it is unlikely that there are unexplored regions missed earlier in the run. Although there is no guarantee that complete equilibrium sampling is given, we have concluded that the observed overlap is acceptable and that adequate sampling within the MD trajectory for the binding site had been obtained.

1st vs. 2nd	0.87
1st vs. 3rd	0.86
2nd vs. 3rd	0.87

Table 4-1: PCA normalized overlap for the binding site within ERCC1.

Plots of the RMSDs for the backbone atoms from the initial co-ordinates of the XPA peptide and ERCC1 free and bound to XPA for the last 10 ns of the simulation illustrate the inherent stability of the complex (Figure 4-4). For the XPA-free simulations, the protein backbone RMSD fluctuated about a mean of 1.8 Å. When XPA was bound to ERCC1, the protein backbone RMSDs fluctuated about a mean of 1.6 Å, indicating a stabilizing effect induced by XPA interactions within ERCC1. Relative to values calculated for the ERCC1 backbone, the XPA backbone RMSD fluctuated around a higher mean of 3.5 Å, illustrating the greater mobility of the XPA peptide as compared to the ERCC1 protein. This observation was also confirmed by results presented in Figure 4-5-a, where unbound ERCC1 main-chain B-factors (averaged over heavy atoms) are generally higher than the corresponding bound values. This, again, suggests the relative flexibility of the model in this region, especially residues 105-119, 140-160 and 168-177, which constitute the XPA binding site. Within the XPA-ERCC1 model, residues 178-183 were shown to have more flexibility than those in the free model suggesting that they are not involved in the protein-protein interaction. Overall, the 22 residues defining the binding site seem to be relatively rigid during the MD simulation in the XPA-ERCC1 models. In particular, residues 72-75 of XPA (see Figure 4-5-b) are more rigid than other XPA residues, suggesting their critical participation in binding to ERCC1.

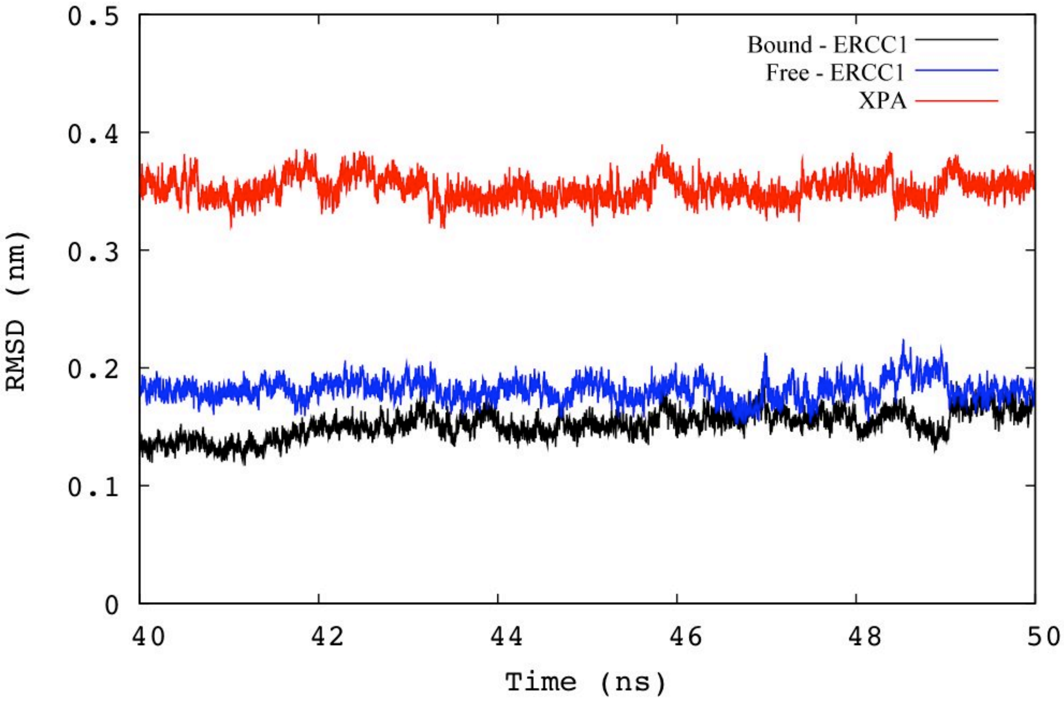


Figure 4-4: RMSD analysis for the ERCC1-XPA MD simulations.

Plot of the RMSD of the backbone atoms from the reference structure as a function of simulation time in XPA-peptide, ERCC1-free and XPA-ERCC1 complex.

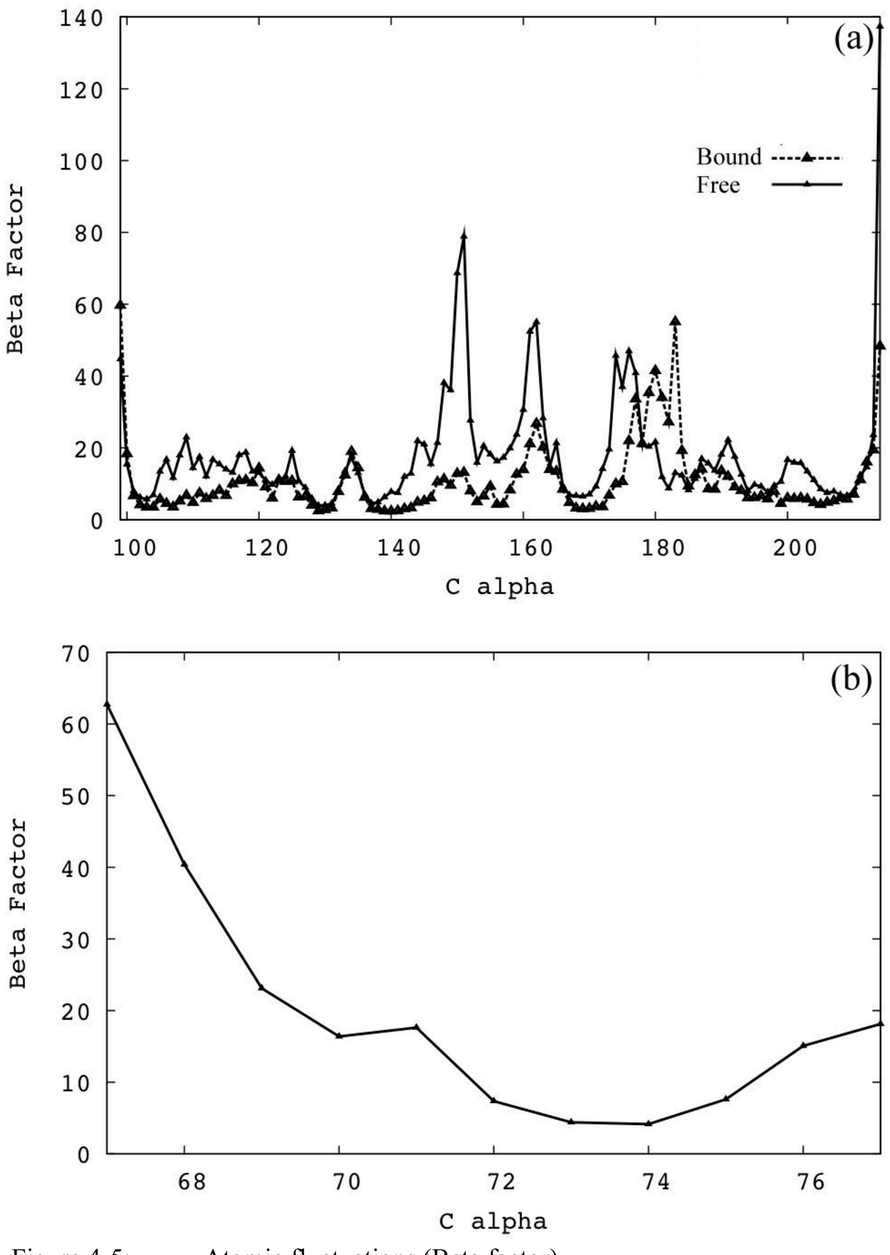


Figure 4-5: Atomic fluctuations (Beta factor).
Plot of the B-factors averaged over the protein backbone atoms as a function of residue number in the simulations of (a) ERCC1-free and ERCC1-bound and (b) XPA peptide. The solid and dotted lines correspond to ERCC1-free and ERCC1-bound, respectively.

4.2.3 Energy Decomposition Of The XPA-ERCC1 Interaction

Current docking methodologies provide a mechanism for the inclusion of flexible receptor side chains within the docking grid (section 2.2.1.2). However, without a clear understanding as to which residues should be introduced as flexible, this can quickly become an intractable problem. As there was no specific information with regards to which residues contribute to the ERCC1-XPA binding interaction, we have calculated the free energy of binding for each residue in ERCC1 that has been shown to interact with XPA. Equilibration of both the holo and apo forms of the ERCC1 binding site allowed us to obtain free energy profiles for each of the amino acids involved in the ERCC1-XPA interaction (Table 4-2 and Figure 4-2).

	Residue *i*	ΔG_i	$\Delta E_{i.gas}^{ele}$	$\Delta E_{i.gas}^{vdw}$	$\Delta G_{i.sol}^{ele}$	$\Delta G_{i.sol}^{nonele}$
ERCC1	ARG106	-1.71	-7.16	-2.16	8.01	-0.41
	GLN107	-2.51	-0.98	-1.76	0.35	-0.12
	GLY109	-1.27	-0.63	-2.11	1.93	-0.46
	ASN110	-1.96	-0.72	-2.50	1.13	-0.13
	PRO111	-1.36	-0.78	-1.23	0.94	-0.29
	PHE140	-1.66	-0.72	-1.95	1.17	-0.16
	LUE141	-1.90	-1.56	-1.17	0.85	-0.02
	SER142	-1.98	-2.59	-0.82	1.53	-0.10
	TYR145	-5.58	-5.19	-2.92	3.12	-0.57
	TYR152	-1.19	-3.61	-1.14	3.68	-0.12
XPA	GLY72	-3.03	-10.72	-1.15	9.28	-0.44
	GLY73	-3.18	-2.52	-3.33	3.06	-0.39
	GLY74	-4.56	-5.02	-4.62	5.78	-0.70
	PHE75	-6.15	0.54	-8.17	2.24	-0.76
	ILE76	-3.82	-0.79	-4.10	1.86	-0.79
Total energy (Kcal/mol)		~ -42	-42.45	-39.13	44.93	-5.46

Table 4-2: Binding energy decomposition into key residues mediating the XPA-ERCC1interaction.

The binding between the ERCC1 binding site and the XPA peptide is due mainly to the favorable solute-solute electrostatic and hydrophobic interactions ($\Delta E_{gas}^{ele} \approx -42$ kcal/mol, $\Delta E_{gas}^{vdW} \approx -39$ kcal/mol), which outweighed the unfavorable solute-solvent electrostatic interaction ($\Delta G_{solv}^{ele} \approx 45$ kcal/mol) by ≈ -36 kcal/mole. Although the non-polar contribution to the solvation energy ($\Delta G_{solv}^{nonele} \approx -6$ kcal/mol) is favorable for binding, it does not contribute significantly to the binding affinity. The most significant binding contributions between ERCC1 and XPA were determined to be mediated primarily by five residues from the XPA peptide, G72, G73, G74, F75 and I76, contributing approximately 50% of the total binding energy. Within the ERCC1 binding site, Y145 contributed approximately -5 kcal/mol to the binding energy with the remainder of the -42 kcal/mol being distributed among other 8 residues (R106, Q107, G109, N110, P111, L141, S142 and Y152). Overall, the main contributors to the binding energy were Y145 from ERCC1 and F75 from XPA, which stacks against N110 from ERCC1, which contributed ~-2 kcal/mol (see Figure 3c). In our model, the hydroxyl group of Y152 in ERCC1 also forms two hydrogen bonds with the backbone carbonyl of G72 in XPA. Two hydrogen bonds were also observed between S142 of ERCC1 and G72 of XPA, and Q107 of ERCC1 and G73 of XPA. An intramolecular hydrogen bond was observed between T71 and G73 within XPA and is in agreement with experimental findings detailing critical residues mediating the XPA-ERCC1 interaction[24]. All of these residues therefore explicitly define the binding site within ERCC1 and therefore the conformation of the potential inhibitor that should mimic the XPA peptide (see Figure 4-2).

4.2.4 Flexible Docking Virtual Screening

Decomposition of the total binding energy from our models into individual residue contributions allowed us to identify key residues that mediate the ERCC1-XPA interaction. As Y145 contributed about 26% of the total ERCC1 binding energy (see Table 4-2), for our initial docking runs, we have used the minimized crystal structure of the ERCC1-XPA complex, with Y145 being the only flexible residue during the virtual screening procedure.

4.2.5 Ensemble-Based Virtual Screening

An obvious drawback when considering only the flexibility of restricted protein fragments is that the collective motion of the complete receptor backbone is neglected. To overcome this deficiency we have used an ensemble of protein conformations as a target for docking as an alternative approach to introduce a feature of global protein flexibility. Such an ensemble at the extreme, is capable of describing the entire conformational space of the binding site, yet must still be represented by a set of limited conformations in order to save computational screening time.

To generate a reduced set of representative models of the ERCC1 binding site, we applied root-mean-square difference (RMSD) conformational clustering to the apo-binding site trajectory obtained from the MD simulation. Using the average-linkage algorithm, we obtained a total of 44 clusters that represent the complete MD trajectory (see Figure 4-6). Of these 44, the six most dominant clusters represented approximately 48%, 8%, 6%, 5.5%, 4% and 3.8% of the entire ensemble. We concluded that these six dominant clusters were sufficient to describe the collective conformational changes in the apo-ERCC1 trajectory for subsequent screening experiments.

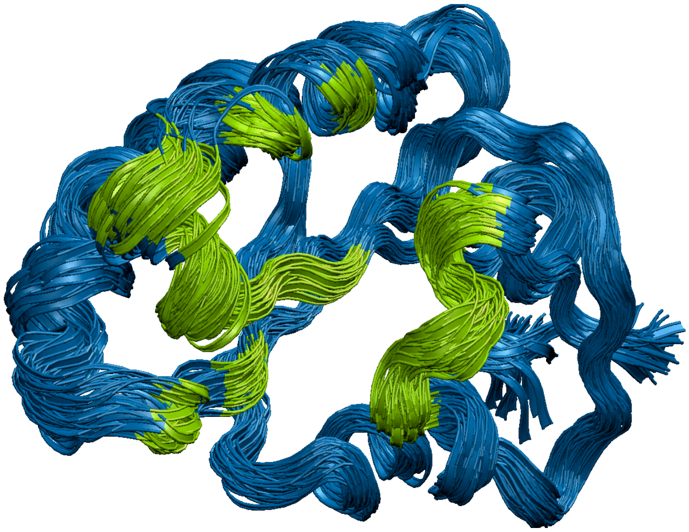


Figure 4-6: Forty-four representative structures for the ERCC1 binding site. The binding site is in green.

The accumulation of approximately half of the MD trajectory conformations into the first dominant cluster motivated us to use PCA in order to extract the lowest energy conformation of the binding site. This conformation was then appended to the six dominant structures to perform an ensemble-based virtual screening against the full set of ligand compounds. Figure 4-7 represents the spatial distributions of occupancies for the conformational states over the planes spanned by the dominant principal components of the binding site. The grouping of conformations into a single cluster suggests the presence of a global minimum and a significant basin of attraction, indicative of a low energy conformation for the binding site. A representative structure for the folded conformation was then calculated by collecting all conformations contained within the three minima. The backbone atoms of the binding site of these conformations were then RMS fit to the reference structure and the centroid of the RMS fit was used as an additional representative conformation to the six dominant structures in virtual screening experiments. It is noteworthy that, while the two equilibrated structures were produced through two different methods, the RMSD between the two models was only 1.12 Å, which is quite low when compared to values calculated previously (see Figure 4-4).

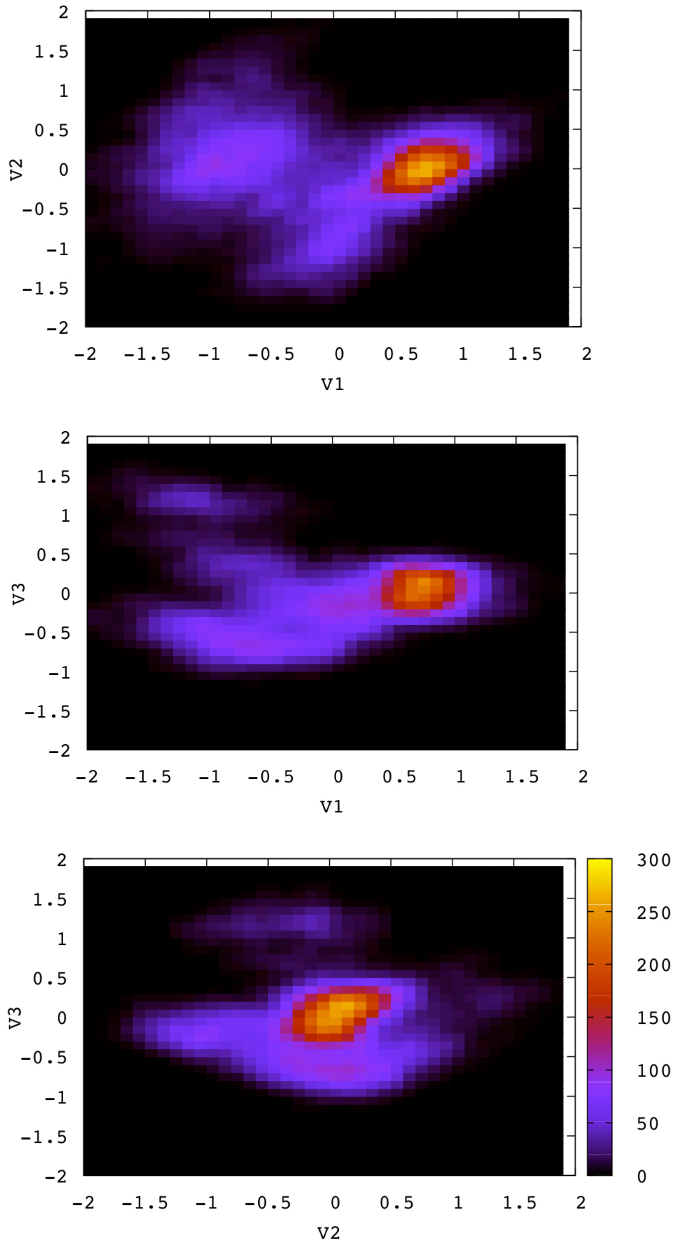


Figure 4-7: PCA of the ERCC1 binding site.

Projections of the ensemble of conformations onto the planes of the three most important principal components. The first and second, the first and third and the second and third principal components are plotted on the x and y axes, respectively. The histograms represent the occupancies of the corresponding conformation states, with lighter colors indicating more frequently visited areas. The three histograms reveal a global minimum indicating the convergence of sampling and folding of the ERCC1 binding site.

4.2.6 Pose Clustering

In this study, we performed eight screening experiments against the full set of database compounds. The first was against the minimized crystal structure with a flexible Y145, while the other seven constituted the ensamble based screen. Screening of all 3450 compounds contained in the NCDIS and Drugbank databases, against the eight target structures, produced a total of 2.76 million distinct poses that required classification. While AUTODOCK is capable of clustering these poses into subgroups depending on RMSD, the total number of clusters and population for each cluster is mostly dependent on the RMSD cutoff that is initially chosen. As such, there is no adequate means to anticipate an optimum cutoff for the RMSD to produce the best quality result. As we are dealing with a diverse set of input ligands, this clustering method does not provide an accurate means of comparing resulting populations and binding energies between ligands, making it difficult to score compounds accurately.

Following the iterative clustering procedure described in chapter 3 (section 3.4.2), docking results were clustered to produce optimal clustering patterns. This is illustrated in Figure 4-8, where the elbow criterion for the top three hits from ensemble screening suggested different cluster counts for the different poses.

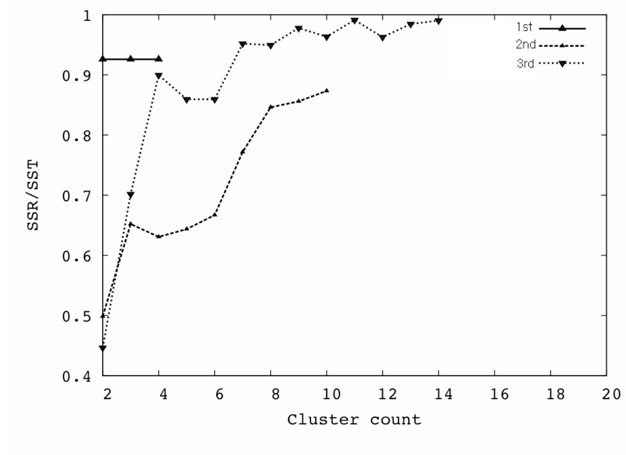


Figure 4-8: Percentage of variance for the top 3 hits from RCS VS experiment. The SSR/SST is expected to plateau for cluster counts exceeding the optimal number of clusters.

This clustering methodology proposed three clusters with three different representative conformations for the planar molecule characterizing the top hit (see), while 14 clusters were suggested for the third hit which includes more torsional degrees of freedom. We propose this clustering method as an alternative dynamic technique to be used for virtual screening as opposed to clustering all the poses with a single RMSD value for all the docked compounds.

4.2.7 Pose Ranking

For each virtual screening experiment, we have ranked significant poses for each of the 3450 molecules contained in the database by using the results from the elbow criterion and the lowest energy that corresponds to the most populated cluster. Once all poses from each ligand entry were clustered, we then filtered all of the clusters so that only those containing at least 25% of the total population are considered as top hits. For the flexible screening experiment, top hits were ranked by their binding energies of the largest cluster. Top hits from the ensemble-based screening experiments were collected from the seven experiments by first extracting the largest cluster from each individual screening flowed by ranking the clusters by their binding energies. This produced a set of non-redundant hits ranked by their binding energies of the most populated cluster.

Table 3 shows the top ten hits from the ensemble-based screening ranked by their binding energies and compared to their docking results from the flexible run. It is clear that the two methods produce dissimilar ranking for most of the compounds, with several hits being excluded from the flexible screening due to the 25% cut off on the largest cluster population. Although the flexible docking showed low binding energies, even lower than the ensemble based calculations, the poor clustering of these hits suggested that they didn’t fit properly into the binding pocket. This observation indicates the importance of backbone dynamics and side-chain movement as compared to allowing only one residue to be flexible during docking. In order to refine the ensemble based screening results and consider all possible target conformations, we docked the top 50 hits obtained from the ensemble docking results to the complete set of receptor representative structures and applied the relaxed complex scheme to re-score the poses.

4.2.8 Rescoring Using The RCS

The non-redundant 50 hits obtained from the ensemble-based screening experiments were re-docked into all of the 44 clusters representing the apo-ERCC1 MD ensemble. For each compound (see Figure 4-9) the RCS weighted average and minimum binding energies were compared to the ensemble screening average and minimum binding energies. To compare how these ligands were docked to the crystal structure, we also included the binding energies of the most populated cluster from the flexible screening.

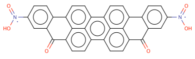
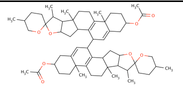
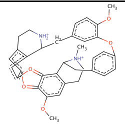
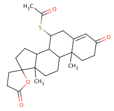
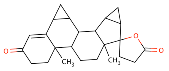
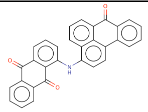
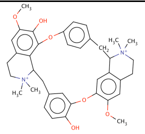
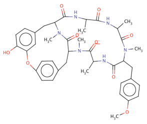
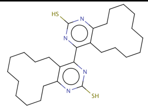
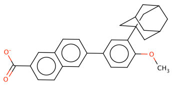
ENS Rank	FLEX Rank	ENS mean Kcal/mol	FLEX BE Kcal/mol	ID	Structure
1	1	-8.79	-11.67	NSC # 51535	
2	EXCLUDED	-8.21	-9.28	NSC # 93352	
3	17	-7.55	-9.83	NSC # 181486	
4	7	-7.51	-10.24	ZINC03861599	
5	57	-7.49	-9.11	ZINC03927200	
6	15	-7.49	-9.88	NSC # 13987	
7	EXCLUDED	-7.46	-9.34	NSC # 36387	
8	76	-7.41	-8.98	NSC # 259969	
9	13	-7.39	-9.95	NSC # 372060	
10	8	-7.38	-10.19	ZINC03784182	

Table 4-3: Top ten hits from the Ensemble screening.

The ranking and binding energies are compared to the Flexible screening.

For most of the compounds, the RCS weighted-mean and the ensemble-mean differ by less than 0.7 kcal/mol, indicating that our selection of only the first six representative structures was sufficient to describe the conformational space of the binding site (see Figure 4-9). Furthermore, those compounds which ranked very low when docked to a number of target structures including the most dominant conformation (-3.52 Kcal/mol), could be excluded in the RCS scoring.

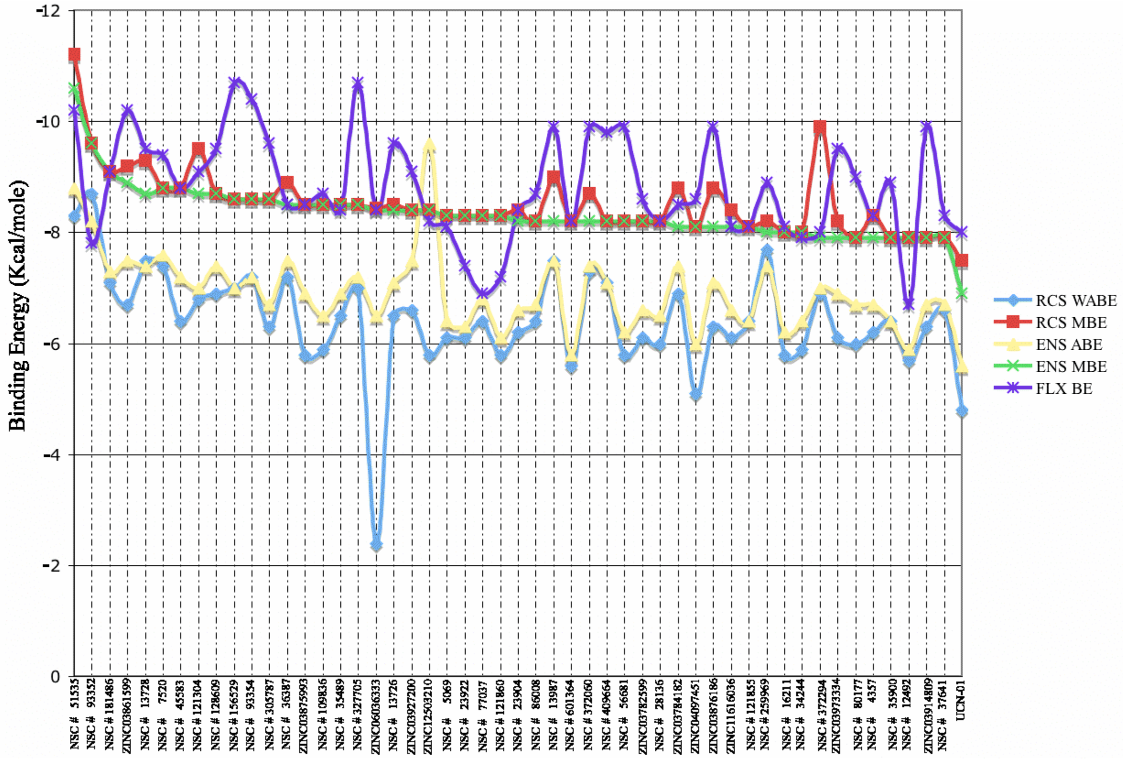


Figure 4-9: Binding energy statistics for the irredundant top 50 hits suggested by the ensemble-based screening.

The RCS weighted-average (RCS WABE) and the ensemble average binding energies (ENS ABE) showed similar behaviour for most of the compounds. The RCS minimum energy (RCS MBE) and the ensemble minimum binding energy (ENS MBE) plateau with small number of hits show lower BE for the RCS screening indicating their binding to infrequent conformations. The Flexible screening binding energies fluctuated around -9 kcal/mol showing lower binding energies than the other methods. The RCS excluded the ZINC06036333 compound from the top hits suggesting that it is not a true binder.

While each conformation in the ensemble screening contributed one seventh of the total binding energy, the relaxed complex scheme evaluates the total energy by scaling individual energies by the percentage population of the docked structures. In this context, for a compound to be ranked high in the RCS score, it must be docked properly against most of the target ensemble, particularly, against the most dominant structures. This scoring technique enables the RCS method to identify and eliminate decoy compounds from the list of ligands. Some hits showed lower RCS minimum-binding energies than their representative values calculated using the ensemble approach, suggesting their binding to a more accepting representative structure. This is the main advantage of using RCS to re-score the top hits; as some compounds may bind to a rarely visited receptor conformation.

Figure 4-10 shows the average clustering of the top hits for the three different methods. Once more, the RCS and the ensemble methods showed the same results for most of the ligands. The average cluster populations for the two methods are more than 30 poses for about 56% of the top hits, indicating their binding for most of the representative conformations. It is noteworthy to indicate that more than 50% of the ensemble screening's hits were excluded from the flexible screening due to the 25% cutoff criterion on the largest cluster population.

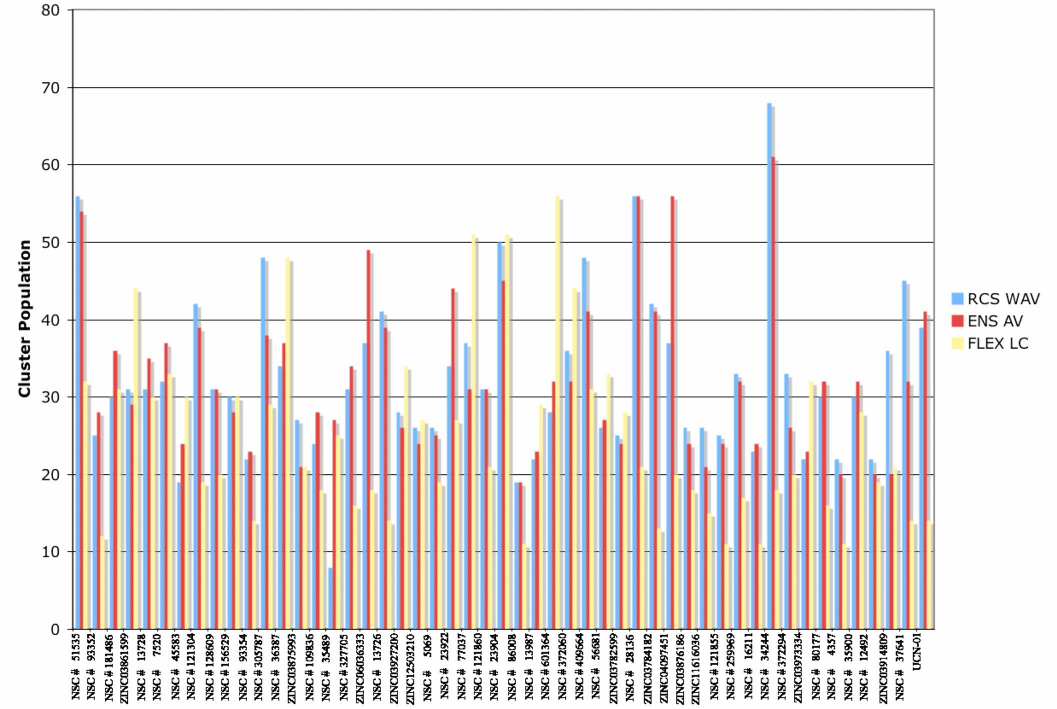


Figure 4-10: Clustering of the irredundant top 50 hits suggested by the ensemble-based screening.

The RCS weighted-average (RCS WAV) and the ensemble average (ENS AV) population showed the same clustering for most of the top hits. For more than 50% of the hits, the flexible largest cluster population (FLEX LC) is lower than the 25% cut off, indicating that they have been excluded from the flexible-screening ranking.

4.2.9 Electrostatic Surface Calculations

The binding mode for three selected top hits within their most favored binding site conformations is shown in Figure 4-11. Electrostatic surface maps are included to provide an additional perspective of the charge distribution in the ERCC1 cavity. The binding cleft is mainly positively charged with small negatively charged spots on boundaries of the binding site. This electrostatic potential distribution indicates that the binding site may exhibit a weak positive electrostatic potential. Although, the charge distribution changed slightly between the two representative binding sites indicating the perseverance of its over all shape, the positive potential is apparent in the closed conformation. Moreover, the charge complementarities between the binding site and the top hits is apparent from Figure 9 and is indicative of a proper binding mode.

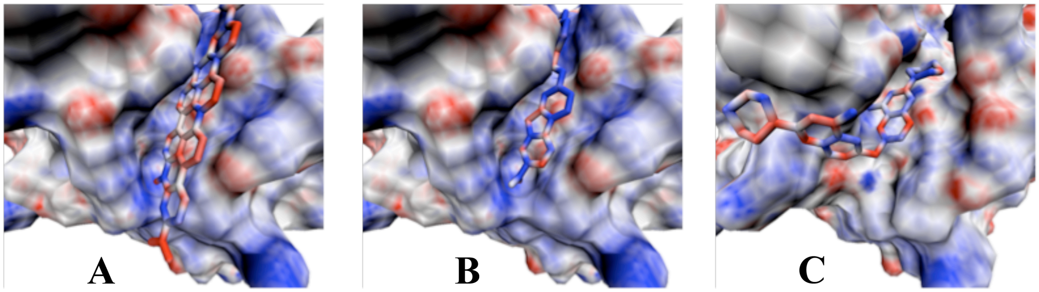


Figure 4-11: Three selected hits within their preferred binding site conformations.

The binding cleft within ERCC1and the top hits are colored by residue electrostatic potential with coloring scale of -10 kT/e (red) to +10 kT/e (blue). The closed structure (a and b) is more positive than the open structure (c).

4.2.10 Pharmacophore Characterization

Having obtained a comprehensive description of the ERCC1-XPA binding interaction and a diverse set of ligand interactions, we turned our attention to the creation of a model describing key chemical features of both the binding site and ligands. These models are commonly known as pharmacophores and represent chemical functions (see section 2.2.2 for more details), valid not only for a currently bound molecule, but also for the putative binding characteristics of unknown molecules. Due to its overall simplicity, this method can be extremely computationally efficient and is exceptionally well suited for interpreting the virtual screening results of large compound libraries. In general, a single ligand bound to a protein's active site is able to provide sufficient information to start the construction of a pharmacophore model. This approach is generally used in the analysis of a known X-ray or NMR structure of a ligand-receptor complex. It is also possible to develop a pharmacophore from a set of ligands that bind to the same region within the target. In the first, structure-based approach, critical chemical features are recognized from the known complex. The second, ligand-based approach extracts a common set of chemical features by exploring properties of the bound ligands. The top hits from NCI diversity set included in the scored compounds were filtered for drug likeness and recorded in Table 4-4 along with the DrugBank top poses.

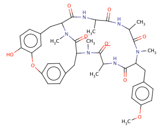
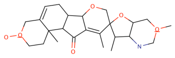
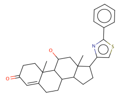
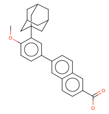
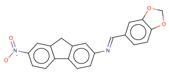
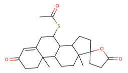
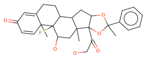
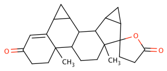
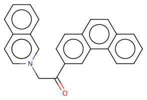
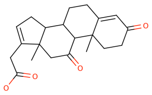
RCS Rank	ENS Rank	RCS WABE Kcal/mol	ID	Structure
3	8	-7.66	259969	
6	3	-7.37	7520	
8	14	-7.21	93354	
15	10	-6.93	ZINC03784182	
17	22	-6.76	121304	
18	4	-6.66	ZINC03861599	
19	27	-6.62	37641	
20	5	-6.57	ZINC03927200	
21	23	-6.51	35489	
23	28	-6.40	86008	

Table 4-4: Drug-like compounds from the NCI set and top hits from the DrugBank ranked by their RCS score and compared to the ensemble screening rank.

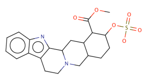
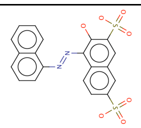
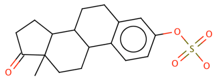
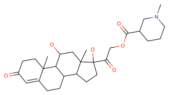
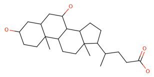
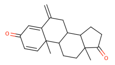
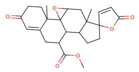
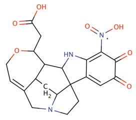
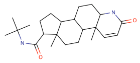
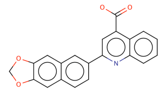
RCS Rank	ENS Rank	RCS WABE Kcal/mol	ID	Structure
25	41	-6.37	121855	
27	16	-6.35	45583	
28	19	-6.31	ZINC03876186	
30	33	-6.24	23904	
31	30	-6.24	ZINC03914809	
33	25	-6.24	ZINC03973334	
34	32	-6.11	ZINC11616036	
36	40	-6.07	5069	
37	35	-6.07	ZINC03782599	
40	42	-5.99	134244	

Table 4-4 Continued.

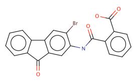
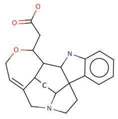
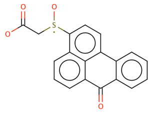
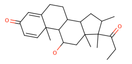
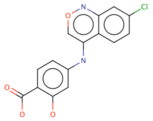
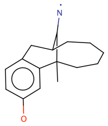
RCS Rank	ENS Rank	RCS WABE Kcal/mol	ID	Structure
42	45	-5.82	56681	
43	47	-5.82	121860	
45	46	-5.79	16211	
46	34	-5.76	ZINC12503210	
47	49	-5.70	12492	
49	48	-5.11	ZINC04097451	

Table 4-4 Continued.

The binding energies for the hits ranged from -7.66 to -5.11 Kcal/mol. Note, at this energy range, the UCN-01 compound is not selected among the top hits since its RCS score was -4.81 Kcal/mol (see Figure 4-9). The top hits showed an overall similar structure including planer hydrophobic rings mostly located on the two edges of the ligands with hydrogen bond donors and acceptors on the middle of the structures. These hits generally mimic the XPA peptide (see Figure 4-2) in its interaction with the ERCC1 binding site. Most of the filtered compounds showed almost the same ranking in the ensemble based calculations. This is a confirmation that the six representative structures were sufficient to substitute the full set of the 44 representative conformations. To further reduce the complexity of the pharmacophore generation, those ligand atoms that did not fall within a cut off of 25% occupancy were removed and remaining atoms were used to construct the excluded shell describing the ligands as bound to ERCC1 (see Figure 4-12-a).

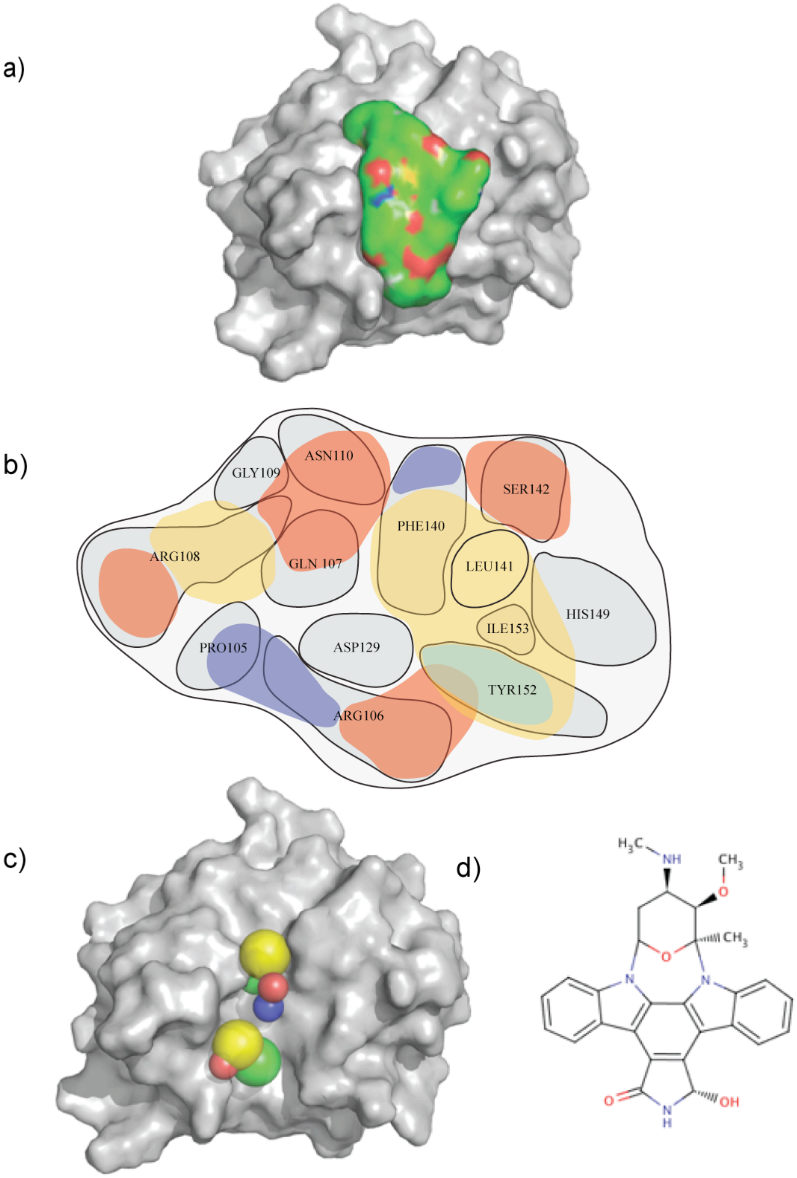


Figure 4-12: Pharmacophore determination. (a) The equilibrated ERCC1 (grey surface) showing the excluded volume occupied by atoms from ligands obtained in the virtual screening experiments (green surface). Atoms included in this image were obtained by clustering the top ligands, from virtual screening experiments, and omitting those that were outside of a 90% RMSD cutoff. (b) Pharmacophores from each of the top 30 ligands were created with their interactions in the ERCC1 binding site. The type of pharmacophore interactions, with each residue were scored and are represented schematically. Yellow patches indicate hydrophobic interactions with the pocket, red and blue patches represent hydrogen bond acceptor and donors respectively, while green patches indicate aromatic interactions. Orientation of the binding site is the same as in panel a. Tyrosine 145 and Histidine 149 (indicated by an asterisk *) do not lie on the bottom of the pocket but are observed within a lip that overhangs the pocket (see panel a). (c) The averaged pharmacophore model obtained from the docked poses from virtual screening Each sphere represents a specific chemical entity with the size being representative of the overall contribution at each position. Coloring is identical to that described for panel a. (d) The chemical structure of UCN-01.

The ERCC1 binding site pharmacophore model (see Figure 4-12-b) consists of two spatially separated areas of hydrophobic interaction encompassing residues R108, F140, L141, Y152 and I153. In addition to the hydrophobic interactions, there are also two regions of possible aromatic stacking with residues Y145 and Y152. We note that residue Y145 was identified as a critical interaction in the ERCC1-XPA binding energy calculations and was set as flexible in the virtual screening experiments. A critical observation was that the Y154 side chain occupies a position on the floor of the binding pocket, while the Y145 side chain sits above the binding pocket, resulting in the formation of a shallow cavity. This configuration of tyrosine side chains presents the likelihood for forming an aromatic sandwich, a feature that is observed in the active site of monoamine oxidases. In addition to hydrophobic and aromatic features, the binding pocket is also defined by three hydrogen bond acceptor regions (residues R108, R106, N110 and S142) and, two smaller hydrogen bond donor regions (residues P105, R106 and a small region of F140).

One of the most interesting findings of this exercise was that the majority of ligands from the virtual screening experiments were extremely symmetrical, a feature that is reflected in the pharmacophore model (see Figure 4-12-c). The most significant chemical feature within the pharmacophore is the naphalene group, which forms an aromatic sandwich between Y145 and Y152 within the ERCC1 binding site. The positioning of hydroxyl groups within the napthalene group also results in the formation of favorable hydrogen bonding interactions with R106 and S142 in ERCC1. The hydrocarbon linker region between the two hydrophobic ring functionalities spans the 6Å seperating hydrophobic patches in the ERCC1 binding site and also provides additional rotational flexibility required for proper ring orientation.

4.3 Conclusion

NER removes bulky DNA damage induced by UV irradiation or by UV-mimetic DNA damaging agents such as cisplatin [33]. One way to target NER is to inhibit the interaction between two of the NER essential elements, ERCC1 and XPA. To date, only one such compound, UCN-01, has been characterized and tested pre-clinically.[27] As a consequence, it is expected to potentiate cisplatin toxicity based on its suspected effect on ERCC1. Accordingly, the purpose of this study was to undertake a computational search for potential inhibitors of this pathway by inhibiting the XPA-ERCC1 protein-protein interaction that is involved in the final stages of this pathway.

Using MD simulations and binding energy analysis we identified the key residues constituting the binding pocket within ERCC1 for its interaction with XPA. Subsequently, we have used conformational RMSD clustering to extract 44 different representative structures that describe the whole conformational space of the ERCC1 pocket. Using the dominant ERCC1 structures, we run eight screening experiments that employed the NCIDS library and DrugBank small-molecules as the set of putative ligands and AutoDock as the docking engine. The docked poses were clustered using a proposed dynamic technique that adapts the number of clusters to the optimal clustering pattern. Ranked by the binding energy of the most populated cluster, the non-redundant top 50 compounds resulted from the ensemble-based screening were rescored using the RCS by re-docking them to the 44 structures that describe the whole MD trajectory. Top hits were used to construct a pharmacophore model that can be used in the subsequent identification of novel ERCC1-XPA inhibitors. This pharmacophore model points out to the important features that must be present for an active ERCC1-XPA inhibitor. It can be employed as the basis for the rational design of specific inhibitors for the XPA-ERCC1 interaction that would ultimately result in the development of a cisplatin-based combination therapy for a broad range of cancers.

Aside from the target presented here, we recently employed the same methodology to identify inhibitors for the XPF-ERCC1 interaction. That is targeting the enzyme heterodimer to eradicate its activity as a direct way of regulating the NER pathway.[34]

4.4 Bibliography

1. Rouillon, C.; White, M. F., The evolution and mechanisms of nucleotide excision repair proteins. *Res Microbiol* **2011,** *162* (1), 19-26.

2. Nouspikel, T., DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility. *Cell Mol Life Sci* **2009,** *66* (6), 994-1009.

3. Koberle, B.; Tomicic, M. T.; Usanova, S.; Kaina, B., Cisplatin resistance: preclinical findings and clinical implications. *Biochim Biophys Acta* **2010,** *1806* (2), 172-82.

4. (a) Metzger, R.; Leichman, C. G.; Danenberg, K. D.; Danenberg, P. V.; Lenz, H. J.; Hayashi, K.; Groshen, S.; Salonga, D.; Cohen, H.; Laine, L.; Crookes, P.; Silberman, H.; Baranda, J.; Konda, B.; Leichman, L., ERCC1 mRNA levels complement thymidylate synthase mRNA levels in predicting response and survival for gastric cancer patients receiving combination cisplatin and fluorouracil chemotherapy. *J Clin Oncol* **1998,** *16* (1), 309-16; (b) Handra-Luca, A.; Hernandez, J.; Mountzios, G.; Taranchon, E.; Lacau-St-Guily, J.; Soria, J. C.; Fouret, P., Excision repair cross complementation group 1 immunohistochemical expression predicts objective response and cancer-specific survival in patients treated by Cisplatin-based induction chemotherapy for locally advanced head and neck squamous cell carcinoma. *Clin Cancer Res* **2007,** *13* (13), 3855-9; (c) Bellmunt, J.; Paz-Ares, L.; Cuello, M.; Cecere, F. L.; Albiol, S.; Guillem, V.; Gallardo, E.; Carles, J.; Mendez, P.; de la Cruz, J. J.; Taron, M.; Rosell, R.; Baselga, J., Gene expression of ERCC1 as a novel prognostic marker in advanced bladder cancer patients receiving cisplatin-based chemotherapy. *Ann Oncol* **2007,** *18* (3), 522-8; (d) Jun, H. J.; Ahn, M. J.; Kim, H. S.; Yi, S. Y.; Han, J.; Lee, S. K.; Ahn, Y. C.; Jeong, H. S.; Son, Y. I.; Baek, J. H.; Park, K., ERCC1 expression as a predictive marker of squamous cell carcinoma of the head and neck treated with cisplatin-based concurrent chemoradiation. *Br J Cancer* **2008,** *99* (1), 167-72.

5. Yano, J. K.; Wester, M. R.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F., The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-A resolution. *J Biol Chem* **2004,** *279* (37), 38091-4.

6. Wood, R. D., DNA damage recognition during nucleotide excision repair in mammalian cells. *Biochimie* **1999,** *81* (1-2), 39-44.

7. de Laat, W. L.; Jaspers, N. G.; Hoeijmakers, J. H., Molecular mechanism of nucleotide excision repair. *Genes Dev* **1999,** *13* (7), 768-85.

8. Sung, P.; Bailly, V.; Weber, C.; Thompson, L. H.; Prakash, L.; Prakash, S., Human xeroderma pigmentosum group D gene encodes a DNA helicase. *Nature* **1993,** *365* (6449), 852-5.

9. Sijbers, A. M.; de Laat, W. L.; Ariza, R. R.; Biggerstaff, M.; Wei, Y. F.; Moggs, J. G.; Carter, K. C.; Shell, B. K.; Evans, E.; de Jong, M. C.; Rademakers, S.; de Rooij, J.; Jaspers, N. G.; Hoeijmakers, J. H.; Wood, R. D., Xeroderma pigmentosum group F caused by a defect in a structure-specific DNA repair endonuclease. *Cell* **1996,** *86* (5), 811-22.

10. Martin, L. P.; Hamilton, T. C.; Schilder, R. J., Platinum resistance: the role of DNA repair pathways. *Clin Cancer Res* **2008,** *14* (5), 1291-5.

11. Kang, S.; Ju, W.; Kim, J. W.; Park, N. H.; Song, Y. S.; Kim, S. C.; Park, S. Y.; Kang, S. B.; Lee, H. P., Association between excision repair cross-complementation group 1 polymorphism and clinical outcome of platinum-based chemotherapy in patients with epithelial ovarian cancer. *Exp Mol Med* **2006,** *38* (3), 320-4.

12. Shirota, Y.; Stoehlmacher, J.; Brabender, J.; Xiong, Y. P.; Uetake, H.; Danenberg, K. D.; Groshen, S.; Tsao-Wei, D. D.; Danenberg, P. V.; Lenz, H. J., ERCC1 and thymidylate synthase mRNA levels predict survival for colorectal cancer patients receiving combination oxaliplatin and fluorouracil chemotherapy. *J Clin Oncol* **2001,** *19* (23), 4298-304.

13. Lord, R. V.; Brabender, J.; Gandara, D.; Alberola, V.; Camps, C.; Domine, M.; Cardenal, F.; Sanchez, J. M.; Gumerlock, P. H.; Taron, M.; Sanchez, J. J.; Danenberg, K. D.; Danenberg, P. V.; Rosell, R., Low ERCC1 expression correlates with prolonged survival after cisplatin plus gemcitabine chemotherapy in non-small cell lung cancer. *Clin Cancer Res* **2002,** *8* (7), 2286-91.

14. Olaussen, K. A.; Dunant, A.; Fouret, P.; Brambilla, E.; Andre, F.; Haddad, V.; Taranchon, E.; Filipits, M.; Pirker, R.; Popper, H. H.; Stahel, R.; Sabatier, L.; Pignon, J. P.; Tursz,

T.; Le Chevalier, T.; Soria, J. C., DNA repair by ERCC1 in non-small-cell lung cancer and cisplatin-based adjuvant chemotherapy. *N Engl J Med* **2006,** *355* (10), 983-91.

15. Cobo, M.; Isla, D.; Massuti, B.; Montes, A.; Sanchez, J. M.; Provencio, M.; Vinolas, N.; Paz-Ares, L.; Lopez-Vivanco, G.; Munoz, M. A.; Felip, E.; Alberola, V.; Camps, C.; Domine, M.; Sanchez, J. J.; Sanchez-Ronco, M.; Danenberg, K.; Taron, M.; Gandara, D.; Rosell, R., Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: a phase III trial in non-small-cell lung cancer. *J Clin Oncol* **2007,** *25* (19), 2747-54.

16. Scheil-Bertram, S.; Tylus-Schaaf, P.; du Bois, A.; Harter, P.; Oppitz, M.; Ewald-Riegler, N.; Fisseler-Eckhoff, A., Excision repair cross-complementation group 1 protein overexpression as a predictor of poor survival for high-grade serous ovarian adenocarcinoma. *Gynecol Oncol* **2010,** *119* (2), 325-31.

17. Altaha, R.; Liang, X.; Yu, J. J.; Reed, E., Excision repair cross complementing-group 1: gene expression and platinum resistance. *Int J Mol Med* **2004,** *14* (6), 959-70.

18. Ozkan, M.; Akbudak, I. H.; Deniz, K.; Dikilitas, M.; Dogu, G. G.; Berk, V.; Karaca, H.; Er, O.; Altinbas, M., Prognostic value of excision repair cross-complementing gene 1 expression for cisplatin-based chemotherapy in advanced gastric cancer. *Asian Pac J Cancer Prev* **2010,** *11* (1), 181-5.

19. Sugasawa, K.; Ng, J. M.; Masutani, C.; Iwai, S.; van der Spek, P. J.; Eker, A. P.; Hanaoka, F.; Bootsma, D.; Hoeijmakers, J. H., Xeroderma pigmentosum group C protein complex is the initiator of global genome nucleotide excision repair. *Mol Cell* **1998,** *2* (2), 223-32.

20. Li, L.; Elledge, S. J.; Peterson, C. A.; Bales, E. S.; Legerski, R. J., Specific association between the human DNA repair proteins XPA and ERCC1. *Proc Natl Acad Sci U S A* **1994,** *91* (11), 5012-6.

21. Buchko, G. W.; Isern, N. G.; Spicer, L. D.; Kennedy, M. A., Human nucleotide excision repair protein XPA: NMR spectroscopic studies of an XPA fragment containing the ERCC1-binding region and the minimal DNA-binding domain (M59-F219). *Mutat Res* **2001,** *486* (1), 1-10.

22. Saijo, M.; Kuraoka, I.; Masutani, C.; Hanaoka, F.; Tanaka, K., Sequential binding of DNA repair proteins RPA and ERCC1 to XPA in vitro. *Nucleic Acids Res* **1996,** *24* (23), 4719-24.

23. Tsodikov, O. V.; Ivanov, D.; Orelli, B.; Staresincic, L.; Shoshani, I.; Oberman, R.; Scharer, O. D.; Wagner, G.; Ellenberger, T., Structural basis for the recruitment of ERCC1-XPF to nucleotide excision repair complexes by XPA. *EMBO J* **2007,** *26* (22), 4768-76.

24. Orelli, B.; McClendon, T. B.; Tsodikov, O. V.; Ellenberger, T.; Niedernhofer, L. J.; Scharer, O. D., The XPA-binding domain of ERCC1 is required for nucleotide excision repair but not other DNA repair pathways. *J Biol Chem* **2010,** *285* (6), 3705-12.

25. Rosenberg, E.; Taher, M. M.; Kuemmerle, N. B.; Farnsworth, J.; Valerie, K., A truncated human xeroderma pigmentosum complementation group A protein expressed from an adenovirus sensitizes human tumor cells to ultraviolet light and cisplatin. *Cancer Res* **2001,** *61* (2), 764-70.

26. Barret, J. M.; Cadou, M.; Hill, B. T., Inhibition of nucleotide excision repair and sensitisation of cells to DNA cross-linking anticancer drugs by F 11782, a novel fluorinated epipodophylloid. *Biochem Pharmacol* **2002,** *63* (2), 251-8.

27. Jiang, H.; Yang, L. Y., Cell cycle checkpoint abrogator UCN-01 inhibits DNA repair: association with attenuation of the interaction of XPA and ERCC1 nucleotide excision repair proteins. *Cancer Res* **1999,** *59* (18), 4529-34.

28. Aune, G. J.; Furuta, T.; Pommier, Y., Ecteinascidin 743: a novel anticancer drug with a unique mechanism of action. *Anticancer Drugs* **2002,** *13* (6), 545-55.

29. Wood, R. D.; Robins, P.; Lindahl, T., Complementation of the xeroderma pigmentosum DNA repair defect in cell-free extracts. *Cell* **1988,** *53* (1), 97-106.

30. Salles, B.; Rodrigo, G.; Li, R. Y.; Calsou, P., DNA damage excision repair in microplate wells with chemiluminescence detection: development and perspectives. *Biochimie* **1999,** *81* (1-2), 53-8.

31. Perrin, D.; van Hille, B.; Barret, J. M.; Kruczynski, A.; Etievant, C.; Imbert, T.; Hill, B. T., F 11782, a novel epipodophylloid non-intercalating dual catalytic inhibitor of topoisomerases I and II with an original mechanism of action. *Biochem Pharmacol* **2000,** *59* (7), 807-19.

32. Wang, Q.; Fan, S.; Eastman, A.; Worland, P. J.; Sausville, E. A.; O'Connor, P. M., UCN-01: a potent abrogator of G2 checkpoint function in cancer cells with disrupted p53. *J Natl Cancer Inst* **1996,** *88* (14), 956-65.

33. (a) Kang, T. H.; Sancar, A., Circadian regulation of DNA excision repair: implications for chrono-chemotherapy. *Cell Cycle* **2009,** *8* (11), 1665-7; (b) Sancar, A.; Lindsey-Boltz, L. A.; Unsal-Kacmaz, K.; Linn, S., Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem* **2004,** *73*, 39-85.

34. Jordheim, L. P.; Barakat, K. H.; Heinrich-Balard, L.; Matera, E.-L.; Cros-Perrial, E.; Bouledrak, K.; Cohen, R.; Tuszynski, J.; Dumontet, C., Modification in cisplatin sensitivity in cancer cells by small molecules inhibiting the interaction between ERCC1 and XPF. *In prepartion* **2012**.

Chapter 5: DNA polymerase beta (pol ß) inhibitors: a comprehensive overview [(1)]

5.1 Introduction

Similar to the problem we discussed in Chapter 4, the target presented here is an important determinant of cancer cells' sensitivity to anticancer agents and is a major mean of acquiring antitumor drug resistance. It is DNA polymerase beta (pol ß), the error-prone polymerase of base excision repair (BER). Here, I will review all currently known inhibitors for pol ß and set the stage for the next Chapter, where the developed VS protocol identified lead compounds that target pol ß activity.

5.1.1 Base Excision Repair As A Therapeutic Target

Among known DNA repair mechanisms, base excision repair (BER) is the major cellular pathway that is responsible for the recovery of single strand breaks (SSB) and removal of damaged bases such as oxidized-reduced, alkylated and deaminated bases.[1] These DNA modifications can occur spontaneously by exposing cells to environmental mutagens, a process that has been estimated to take place at a rate as high as 10,000 alterations per day in a typical mammalian cell.[2] Besides, such modifications can be induced synthetically as a result of anticancer treatments using alkylating agents or ionizing radiation. However, in the latter case, BER constitutes a prevailing way that is usually adopted by cancer cells to reduce the efficacy of and to promote resistance against a growing list of DNA damaging agents including bleomycin,[3] monofunctional alkylating agents,[4] cisplatin[5] and other platinum-based compounds. Therefore, it has been broadly proposed that regulating the BER pathway via small molecule inhibitors can reduce the required dosage of such DNA damaging agents while potentiating their efficacy in eradicating cancer cells.[6] Fortunately, most of the proteins involved in and coordinating the BER process have been

[(1)] A version of this chapter has been submitted to Drug Discovery Today, Jan, 2012.

identified, cloned and crystallized allowing the rational design of small molecule inhibitors for their activity.[7] One of these proteins, DNA polymerase beta (pol ß), has been recognized as a vital element in completing the BER pathway.[8,9]

5.1.2 DNA Polymerase And BER

DNA pol ß, the smallest naturally occurring DNA polymerase enzyme, belongs to the X-family of DNA polymerases, a family that also includes terminal deoxynucleotidyl transferase and DNA polymerases lambda and mu.[10] The uncomplicated small structure of the protein (39 kDa) made it a standard model that helped in understanding the functional mechanisms of other DNA polymerases. In addition, there is a large body of evidence that pol ß plays an important role throughout cell's life. For example, a "knock-out" of the gene that encodes for pol ß in mice results in embryonic lethality, confirming the importance of the protein during fetal development.[11] More importantly, pol ß plays a significant role in chemotherapeutic agent resistance, as its over-expression reduces the efficacy of anticancer drug therapies including cisplatin.[5,12]Furthermore, small-scale studies on different types of cancer showed that pol ß is mutated in approximately 30% of tumors, which in turn reduces pol ß fidelity in DNA synthesis exposing the genome to serious and often deleterious mutations.[13,14] Based on the these findings, pol ß, the error-prone polymerase of BER, has been seriously considered as a promising therapeutic target for cancer treatment.

Many inhibitors of DNA pol ß have been identified during the last two decades. To name but a few, this list includes polypeptides,[15] fatty acids,[16] triterpenoids,[17] sulfolipids,[18] polar lipids,[19] secondary bile acids,[20] phenalenone-derivatives,[21] anacardic acid,[22] harbinatic acid,[23] flavanoid derivatives,[24] and pamoic acid.[25] However, most of these inhibitors are not potent enough or lack sufficient specificity to eventually become approved drugs.

5.2 Base Excision Repair (BER)

A typical mammalian cell is frequently exposed to many factors that can cause severe damage to its genome integrity. For example, reactive oxygen species, alkylating agents and ionizing radiation are generally correlated with the formation of non-bulky damaged nucleotides, abasic sites and single-strand breaks within the DNA molecule.[1,26] Nevertheless, the cell has been endowed with a robust and conserved defense mechanism, namely BER pathway (shown in Figure 5-1) that specifically removes these types of damage and restores the cell to normality.[27,28,29]

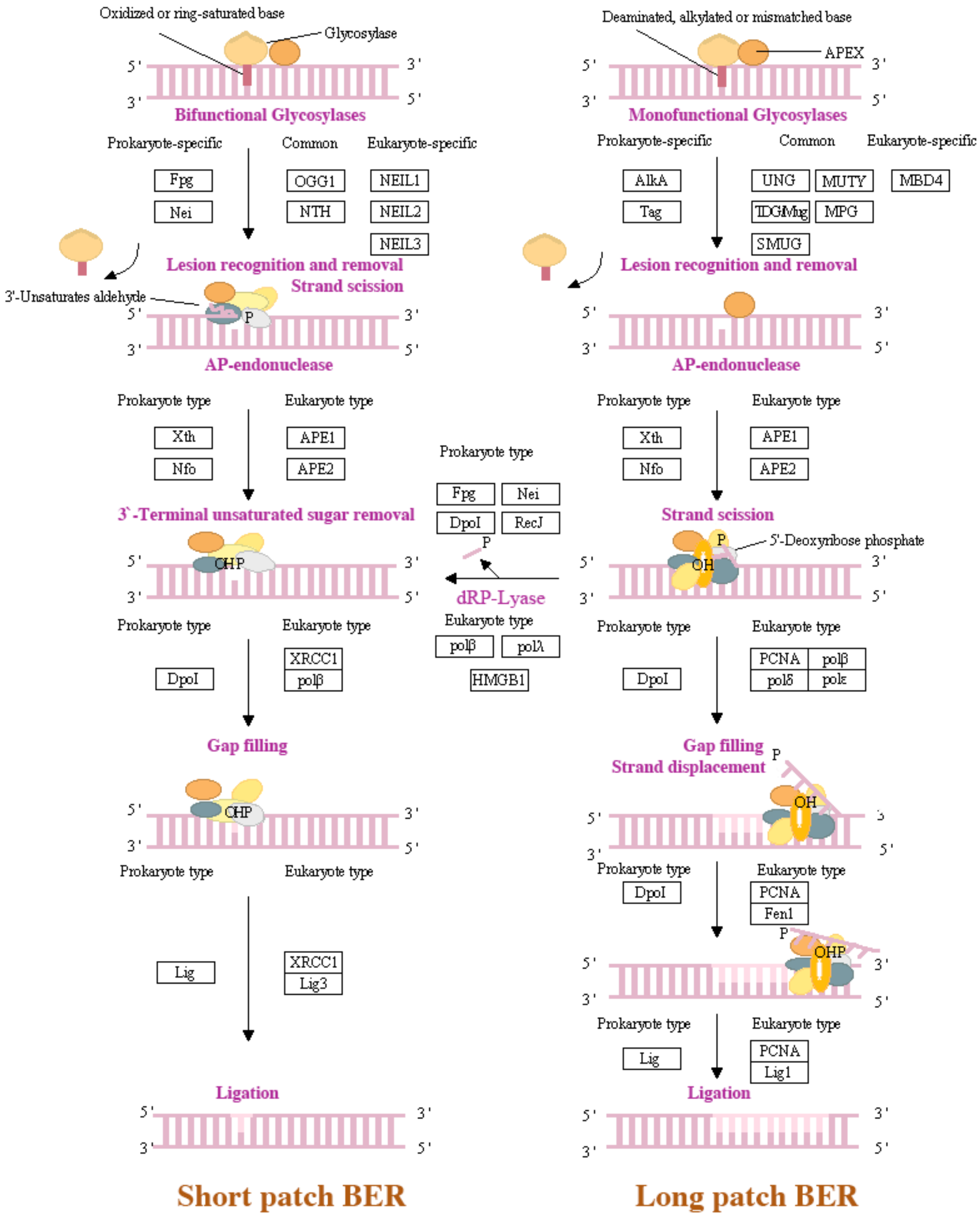


Figure 5-1: Base excision repair (BER). See text for details (adopted from KEGG database).[30]

In fact, BER is a sequential process that is coordinated and processed using at least 30 different proteins. Normally, there are two different types of BER mechanisms that can take place depending on the type of nucleotide modification and the number of incorporated nucleotides within the damaged DNA structure. First is the more frequent single nucleotide BER in which, as the name implies, a single damaged nucleotide is removed and replaced by a new nucleotide satisfying Watson–Crick rules. On the other hand, other studies have identified a sub-BER pathway that can generate repair patches greater than one-nucleotide and is referred to as long-patch BER (see below).

Single nucleotide BER generally starts with damage-specific DNA glycosylases, an assorted family that includes about 11 different species.[29] These enzymes recognize and cleave the N-glycosylic bond between the irregular base and the sugar–phosphate backbone. The outcome of this step is a nucleotide that lacks its base, which is commonly referred to as an abasic or apurinic / apyrimidinic (AP) site. Certainly, an AP site that is generated by the action of a DNA glycosylases enzyme or through a spontaneous hydrolysis of the N-glycosylic bond is considered to be highly mutagenic and risky factor during DNA replication.[31] The next performer in the BER process depends on whether the used DNA glycosylase catalyzes strand incision at the AP site after removing the damaged base (bifunctional glycosylase), or if it can only remove the damaged base (monofunctional glycosylase). In the former case, the lyase activity of pol β removes the 5′-sugar-phosphate residue, creating a one-nucleotide gap and leaving a 5′-phosphate on the downstream DNA strand. While in the latter, the incision step is performed by AP endonuclease, which generates a one-nucleotide gap with 3′-hydroxyl and 5′-deoxyribose phosphate (dRP). It is important to observe that at this stage, single strand breaks (SSB) that may occur within the DNA structure as a result of reactive oxygen species or ionizing radiation are identical to the current BER intermediates that arose form the incision step. Consequently, SSB are commonly repaired through BER and for that reason, the SSB repair mechanism can be considered to be a sub-pathway emerging from the larger BER process.[28] During the next step of BER, the polymerase activity of pol β fills the generated gap with the correct nucleotide leaving the final step to DNA ligase I or III to ligate the nicked DNA termini.[32]

It should be noted that the aforementioned steps simply demonstrate the essential routes that are generally followed during BER. Hence, there are many other enzymes that participate in the process in order to coordinate or modify the DNA termini, making them suitable substrates for pol β or DNA ligases. For example, if a bifunctional DNA glycosylase was employed in the incision step, the 3′-margin of the gap is usually blocked with a sugar-phosphate group. In this case, and because the polymerase activity of pol β requires a 3′-hydroxyl terminus, additional enzymatic activities such as polynucleotide kinase or AP endonuclease are required to modify the 3′-margin and generate a 3′-OH instead.[33] Another example is XRCC1, which acts as a scaffolding protein and interacts with most of the enzymes that are involved in the BER pathway drawing them into the lesion site.[28,34]

Finally, BER can also adopt a different strategy, known as long patch BER,[35] when faced with abnormal modifications within the damaged DNA. These modifications generally occur in the 5´-dRP moiety, which is not a substrate for the lyase activity of pol β, leading to a blockage of the repair process. Examples of such modifications include reduced AP sites, C1 oxidized AP sites[36] or adenine opposite 8-oxoGs.[37,38] In this case, pol β starts the process by adding a single nucleotide to the repaired gap and then is replaced by Pol δ/ε, which extends the repair patch and displaces several nucleotides to create a 50-flap junction.[39]The flap is then detached from the structure with the help of flap endonuclease-1 (FEN1). Alternatively, the modified 5´-dRP group can be removed by FEN1 before the action of pol β takes place. In this case, the generated one-nucleotide gap makes up a perfect substrate for pol β to perform its job. It is also worth mentioning that, under conditions of low energy, poly(ADP-ribose) polymerase 1(PARP1) catalyzes the poly(ADP)ribosylation of proteins and stimulates long-patch BER.[40, 41]

5.3 Structure of DNA pol β

The faithfulness of BER is dependent on the polymerization step, where the major BER DNA pol β, must incorporate the correct Watson–Crick base paired nucleotide into the one-nucleotide repair gap. The enzyme has been identified as a 39-kDa protein with 335 amino acids in its sequence (see Figure 5-2).[8] Its small size compared to other polymerases, made it the smallest and simplest cellular DNA polymerase found. Although pol β lacks the proofreading 3'- or 5'-exonuclease activities, which usually are found in high fidelity enzymes, it possesses 5'-dRP lyase and AP lyase activities instead.[9, 42] The active pol β enzyme is a stable monomer in solution, folded into distinct domains and subdomains that exhibit a variety of functions essential for its activity. These functions include single-stranded (ss) and double-stranded (ds) DNA binding, nucleoside triphosphate (dNTP) binding, and the dRP lyase and nucleotidyl transferase catalytic activities.[43]

Essentially, the full-length enzyme consists of an amino-terminal lyase domain (8 kDa), connected by a short protease-sensitive fragment to a carboxyl-terminal polymerase domain (31 kDa). The 31-kDa domain is further subdivided into C-(catalytic), D- (duplex DNA binding), and N- (nascent base pair binding) subdomains. Interestingly, similar to other DNA polymerases, the overall structure of pol β resembles the shape of a right hand, with fingers (C-subdomain), thumb (D-subdomain) and palm (N-subdomain) arrangements (see Figure 2).[8, 44] The active enzyme requires a single-stranded DNA template and divalent metal ions for its polymerase activity. Moreover, it employs two major substrates, namely, a 2'-deoxynucleoside 5'-triphosphate (dNTP) and a template-primer DNA. Accordingly, the C-subdomain contributes three aspartate residues, namely Asp190, Asp192 and Asp256, which coordinate two divalent metal ions (Mg^{2+}).

Nevertheless, it should be noted that the presence of the nucleotide-binding metal ion as part of the protein structure has been shown to depend on the existence of DNA and the dNTP substrate within their related binding sites in the enzyme.[45]

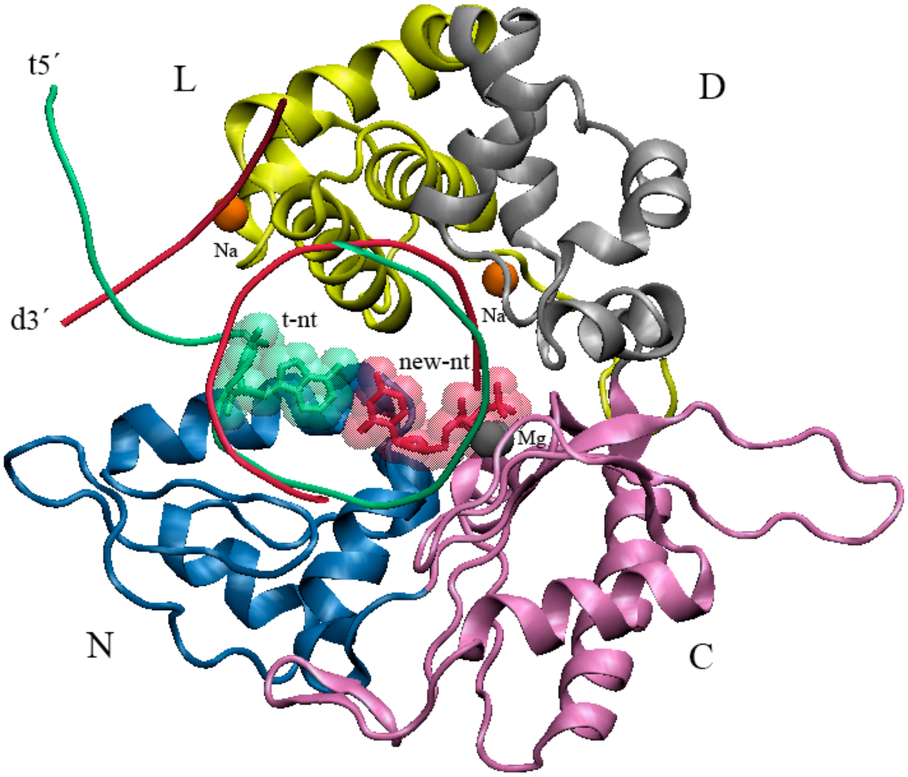


Figure 5-2: Structure of DNA polymerase beta. See text for details.

Several crystal structures of pol β at different stages of the catalytic cycle have been resolved. Considerably, these structures revealed the significant conformational changes that take place within the various subdomains of the protein.[10] These conformational dynamics processes are obvious when comparing the structure of the apo-enzyme[46] to other structures that encompass DNA and the two substrates of the enzyme.[47] The fully loaded pol β structure implied that the 8-kDa domain interacted with the downstream duplex, where the 5'-phosphate on the downstream strand is located close to the dRP lyase active site. Furthermore, the lyase domain cooperates with the N-subdomain in order to form a doughnut-shaped structure that surrounds the DNA molecule (see Figure 2). These notable interactions and functions of the lyase domain indicate that, a small molecule that can bind to the lyase active site, especially, within the ssDNA binding pocket should be able to affect the polymerization activity of pol β as well. Besides the lyase domain dynamics, the N-subdomain seems to exhibit considerable movements once the correct dNTP substrate is bound to the enzyme.[47]Additionally, as illustrated in Figure 2, there is a major conformational change within the structure of the DNA substrate.

Analogous to most DNA-binding proteins, pol β utilizes the well-known helix-hairpin-helix (HhH) motifs that unspecifically interact with the DNA backbone. These HhH motifs are located within the lyase domain (residues 55-79) and the D-subdomain (92-118) and interact with each end of the incised DNA strand, namely, the downstream and primer strands. In addition, like other HhH motifs, they encompass monovalent metals, which in the case of pol β have been identified to be Na^+ ions. However, as mentioned above and similar to the nucleotide-binding metal Mg^{2+} ion, their presence within the structure of the protein is mainly dependant on the existence of the bound DNA substrate.[48]

5.4 DNA pol β inhibitors

In fact, the first attempt to inhibit and understand the activity of pol β employed portions of the protein itself.[15] This earliest *in vitro* study by Husain and his coworkers tried to examine the enzyme as a potential therapeutic target by investigating its interaction with various pol β domains. Interestingly, by evaluating the influence of the 8-, 14-, 27- and 31-kDa N-terminal domains on the active full-length enzyme, they showed that, only the 14-kDa fragment specifically inhibited pol β activity. Although this inhibitory behavior on the isolated protein was eradicated by increasing the concentrations of its substrates, the 14-kDa domain was able to prevent the progress of a BER assay *in vitro*. While this study utilized a large peptide as a potential pol β inhibitor, which may seem to be an impracticable and unreasonable drug candidate, this work provided the much-needed proof of concept that a pol β inhibitor can impede the BER pathway. This, in turn, can provide a means for potentiating therapeutically related DNA damaging agents.

In parallel to this pioneering effort, many attempts were made to isolate and identify a small molecule inhibitor that can specifically bind to pol β. In regard to these endeavors, one can recognize at least two research groups that contributed to a large extent to the discovery of more than sixty molecules that can bind to DNA polymerases in general with a few of them that can target pol β in particular. The main source of these compounds was the screening of natural products for their ability to inhibit the activity of pol β or other polymerases and little work has been done on synthesizing new medicinal chemistry compounds to solve this problem. More surprisingly, despite the large number of crystal structures that have been deposited into the structural databases, the authors could not identify a single study that exploited these structures in building *de novo* small-molecule inhibitors or applying virtual screening techniques to uncover small molecules that can target specific domains of pol β. Table 5-1 and Figure 5-3 list the structure and activities of a selected number of the identified inhibitors (see text below for more details).

In fact, the origin of the studies that were focused on screening for small molecule pol β - inhibitors can be attributed to Mizushina and his co-workers in the mid 1990.[16] This team represents collaboration among research groups in Japan, and their first study was the screening of microbial fermentation for structures that can inhibit DNA polymerases activity. During their analysis, they isolated linoleic acid (LA) (1), a well-known fatty acid as an inhibitor for calf thymus DNA pol α and cloned purified rat DNA pol β. They also examined the effect of a number of commercially available fatty acids on the activity of DNA polymerases. Their findings sat up an important concept; several fatty acids, particularly long chain fatty acids with a *cis-configuration,* interact with DNA polymerases and strongly suppress their activities. More importantly, the same group completed a more detailed study to understand the mode of inhibition of two different fatty acids that showed promising interaction with pol β, namely LA and nervonic acid (NA) (2).[49] Interestingly, NA showed more inhibitory activity than LA with Ic_{50} values of 5.8 and 38 μM, respectively. Comparing the effects of NA and LA on the two DNA polymerases and, the two compounds exhibited different interactions with the two enzymes. That is, the two fatty acids compete with both the dNTP substrate and template-primer for pol β, whereas they bind to pol β without competing with these substances. Moreover, their analysis of the effects of NA and LA on the proteolytic fragments of pol β revealed that the two compounds bind to the 8-kDa DNA-binding domain suppressing its binding to the template-primer DNA. Their results also showed that the binding of the two compounds to the lyase active site of pol β is much stronger than their binding to the polymerase site, to such a degree that ~10 000 times more of either fatty acid was required to inhibit the DNA polymerase activity.[49] Later, in a separate study, Mizushina and his co-workers identified an ergosterol peroxide derivative that can enhance the efficacy of LA in inhibiting the activity of pol β.[50]

Similarly, Tanaka reported the discovery of four triterpenoid compounds isolated from the mycelium of a basidiomycete and found that these compounds selectively inhibit the activities of mammalian DNA polymerase α and β *in vitro*.[17] The four compounds have been termed fomitellic acid (FA) A (3), B (4), C and D and two of them, namely FAs A and B were easier to produce in abundant quantities than the other FAs which were minor components and particularly hard to separate. Accordingly, the authors focused on these FA A and B in a more rigorous study in order to investigate their mode of inhibition of the proteins. Similar to fatty acids, on DNA polβ, the fomitellic acids competed with both the substrate and the template-primer, however, on DNA pol α, their mode of action was not competitive with either the template primer or the substrate. In fact, they found that the two FAs bind strongly to the lyase active site of polβ, but not to the 31-kDa fragment.[51]

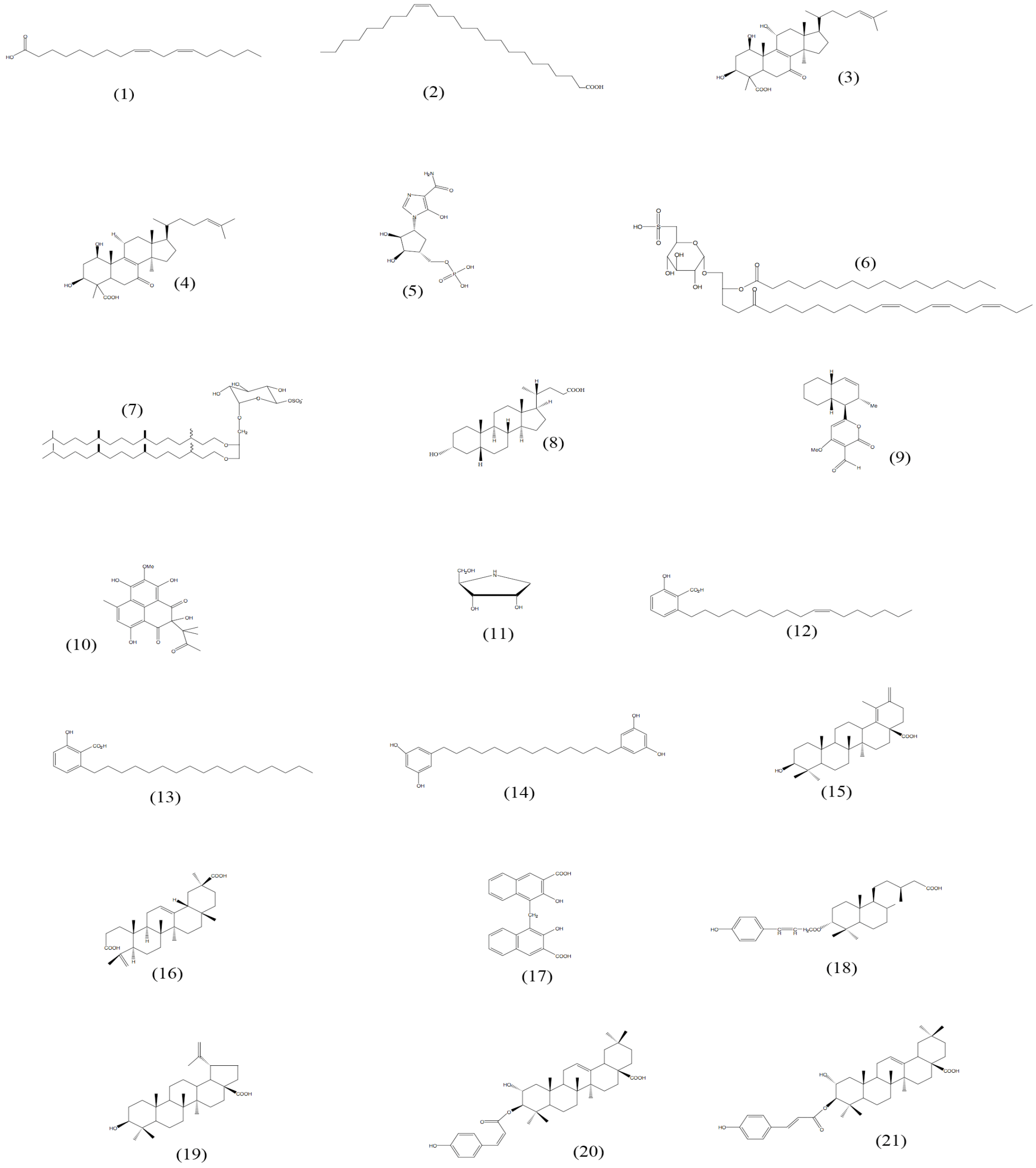
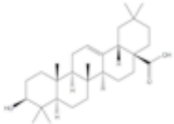
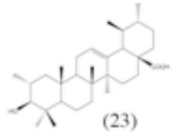
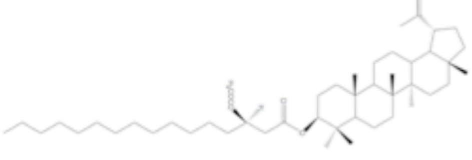
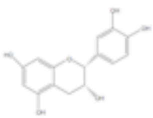


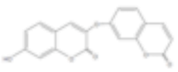
Figure 5-3: Structures of DNA polymerase beta inhibitors listed in Table 5-1.

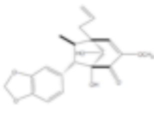

(22)

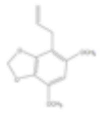

(23)

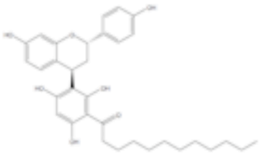

(24)


(25)


(26)


(27)


(28)


(29)

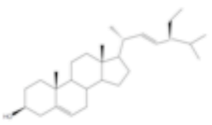

(30)

Figure 5-3 continued.

A comprehensive study of the mode of interaction of breMP with pol β showed that the drug not only competed with the substrate of the polymerase active site but also with the template-primer.[52] These findings were confirmed by examining the interaction of breMP with the individual active sites of pol β, which revealed that the compound essentially binds to the polymerase catalytic domain and could not bind to the template-primer-binding site. These results suggest that breMP directly binds to the substrate-binding site of the catalytic domain, and indirectly, perturbs the template-primer incorporation into its binding domain.[52]

As a continuation of their efforts to isolate novel DNA polymerase inhibitors, Mizushina and his co-workers isolated three sulfolipid compounds from a pteridophyte, Athqrium niprmicum (6 is their most potent structure).[18] They reported these compounds as potent inhibitors of the activities of calf DNA polymerase α and rat DNA polymerase β with IC_{50} values in the range from 1.5 to 3 μg/mL. Analogously to fatty acids, the three inhibitors competed with the DNA template and substrate of DNA polβ, and acted non-competitively on DNA polα. More importantly, these compounds did not affect the activity of a number of other proteins including calf thymus terminal deoxynucleotidle transferase, prokaryotic DNA polymerases such as the

Klenow fragment of DNA polymerase I, T4 DNA polymerase and T*aq* polymerase, the DNA metabolic enzyme DNase I and the human immunodeficiency virus type 1 reverse transcriptase.[18]

ID	Given name	Binding site and affinity	Other targets	Ref
1	Linoleic acid (LA)	8-kDa domain (IC_{50}=38μM)	Pol α	16, 49-50
2	Nervonic acid (NA)	8-kDa domain (IC_{50}=5.8μM)	Pol α	16, 49-50
3	Fomitellic acid (FA) A	8-kDa domain (IC_{50}=125μM)	Pol α	17, 51
4	Fomitellic acid (FA) B	8-kDa domain (IC_{50}=90μM)	Pol α	17, 51
5	BreMP	31-kDa domain (IC_{50}=20μM)	Pol α	52
6	Sulfolipid-derivative 1	Not identified (IC_{50}=3 μg/mL)	Pol α	18
7	KN-208	Not identified (K_i=0.05μM)	Pol α, *E. coli* pol I, HIV RT	19
8	Lithocholic acid (LCA)	8-kDa domain (IC_{50}=11μM)	Pol α	20, 53
9	Solanapyrone A	8-kDa domain (IC_{50}=30μM)	Pol λ	54
10	SCUL-A	Not identified (IC_{50}=17μM)	Pol λ, Pol α	21
11	DRB	8-kDa domain (IC_{50}=28μM)	Polymerases and glycosidases	55
12	Anacardic acid	Not identified (IC_{50}=9μM)	Not studied	22
13	Oleic acid	Not identified (IC_{50}=25μM)	Not studied	22
14	Bis-5-alkylresorcinols derivative	Not identified (IC_{50}=5.8μM)	Not studied	56
15	Triterpenoid-derivative	Not identified (IC_{50}=5.6μM)	Not studied	57
16	Koetjapic acid (KJA)	8-kDa domain (IC_{50}=20μM)	Not studied	25, 58
17	Pamoic acid (PA)	8-kDa domain (K_D =9μM)	Not studied	25
18	Harbinatic acid	Not identified (IC_{50}=2.9μM)	Not studied	23
19	Betulinic acid	Not identified (IC_{50}=14μM)	Not studied	59
20	3-cis-p-coumaroyl Maslinic acid	Not identified (IC_{50}=15μM)	Not studied	59
21	3-trans-p-coumaroyl Maslinic acid	Not identified (IC_{50}=4.2μM)	Not studied	59
22	Oleanolic acid	Not identified (IC_{50}=7.5μM)	Not studied	60
23	2-αhydroxyursolic acid	Not identified (IC_{50}=12.6μM)	Not studied	61
24	Lupane triterpenoids derivative	8 kDa-domain (IC_{50}=3.8 μM)	Not studied	62
25	(-)-epicatechin	8 kDa-domain (IC_{50}=18.5 μM)	Not studied	63
26	Edgeworin	8 kDa-domain (IC_{50}=22.5 μM)	Not studied	64
27	Neolignan-1	8 kDa-domain (IC_{50}=15.5 μM)	Not studied	65
28	Neolignan-3	8 kDa-domain (IC_{50}=18.6 μM)	Not studied	65
29	Myristinin A	Not identified (IC_{50}=2.8μM)	DNA	24
30	Stigmasterol	8 kDa-domain (IC_{50}=60.2 μM)	Not studied	66

Table 5-1: Selected inhibitors of DNA pol. (The structures are shown in Figure 5-3.)

In a similar study, Mizushina extracted prunasin as a weak inhibitor of pol β with an IC_{50} value of 150 μ*M*.[67] The inhibition mode of pol β by the compound was competitive with the substrate, dNTP. This inhibitory behavior was improved to about 40 μ*M* in the presence of fatty acid, indicating that the fatty acid allowed easier access of the compound to the substrate-binding site.

Ogawa et al. isolated sulfated glycoglycerolipid (KN-208) (7), a polar lipid, from an archaebacterium and identified it as an inhibitor for both DNA polymerase α and β.[19] In fact, the same compound also targets *Escherichia coli* DNA polymerase I Klenow fragment (*E. coli* pol I) and human immunodeficiency virus reverse transcriptase (HIV RT), indicating that KN-208 does not selectively inhibit pol α and β as the previously mentioned inhibitors.[19] Moreover, its mode of action on these polymerases was only competitive with the binding of the DNA template primer and not competitive with the binding of the substrate. Although KN-208 can bind to several targets, its binding to pol β was 10-fold stronger than that for pol α, 60-fold stronger than for HIV RT and 140-fold stronger than for *E. coli* pol I, with the sulfate group at the 68-position of the compound was important in its inhibitory activity.[19]

In an important investigation, Ogawa et al. also studied 17 different kinds of bile acids with respect to their inhibition of mammalian DNA polymerases.[20] Intriguingly, their findings revealed that only Lithocholic acid (LCA) (8), one of the major components amongst secondary bile acids, was able to suppress the activity of DNA polymerases. Again, although LCA can bind to several DNA polymerases, its effectiveness against the activity of pol β was the strongest compared to other enzymes. However, in contrast to KN-208, LCA competes with the substrate of pol β and dose not compete with the DNA template-primer binding. Moreover, by comparing the effect of structural variations of LAC to its derivatives, Ogawa and his co-workers found that the C-7 and C-12 positions in the sterol skeleton are important for the inhibitory activity of LCA.[20] The mode of action and specific interactions between LCA and pol β were comprehensively investigated in a different study by Mizushina et al.[53] In this study, the full-length pol β was separated proteolytically into two fragments, namely, the lyase active site (template-primer binding domain 8-kDa) and the polymerase active site (catalytic domain 31-kDa). Binding analysis revealed that LCA tends to bind strongly to the 8-kDa domain, and not to the 31-kDa domain. This important finding was confirmed using NMR analysis, where the 8-kDa domain was shown to associate with LCA as a 1:1 complex with a dissociation constant (K_D) of 1.56 mM. Moreover, NMR chemical shifts were observed only in residues mainly found in helix-3, helix-4, and the 79-87 turn of the same face. Interestingly, the three residues Lys60, Leu77, and Thr79 of pol β exhibited profound interactions with the LCA. It should be also noted that, in this study, Mizushina and his co-workers used molecular docking to understand and illustrate the binding mode of LCA within the NMR-detected binding site and to compare its binding to that of NA, another fatty acid that inhibits pol β (see above).[49] Docking analysis revealed that the binding sites of the two different compounds comprised the DNA binding pocket within pol β as an essential component for their binding. However, the two inhibitors interacted with different residues within the surface of the protein. These residues were close to the DNA binding site on pol β with the two residues Lys35 for NA and Lys60 for LCA showing an important role in the binding of the two compounds, respectively.[53]

Another study by Mizushina et al. identified solanapyrone A (9) as an inhibitor for both DNA polymerase β and λ with IC_{50} values of 30 μM for pol β and 37 μM for pol λ.[54] Since polβ and λ are two similarly structured proteins that are descended from the same family of DNA polymerases, Mizushina focused on polβ in order to examine its interaction with solanapyrone A. Interestingly, solanapyrone A competed with both the DNA template and the nucleotide substrate. They also found that the compound could bind selectively to the N-terminal 8-kDa domain of polβ. In fact, the Mizushina group showed that solanapyrone A inhibits the binding of DNA to the single-stranded DNA-binding site within polβ and does not affect the other two activities of the 8-kDa domain, namely, recognition of the 5'-phosphate in gapped DNA structures and AP lyase activity. Moreover, similarly to the LCA case, the binding interactions between solanapyrone A and the ss-DNA binding region on the surface of polβ were confirmed using molecular docking simulations. According to their results, the two ketone groups of the compound bound strongly to the hydrophilic residue Lys60, and the benzene groups interacted with the hydrophobic amino acids in both helix-3 and helix-4.[54]

In a different screening study conducted by the same group, Perpelescu et al. tested the effects of two phenalenone-skeleton-based compounds, sculezonone-B (SCUL-B) and sculezonone-A (SCUL-A) (10), upon the activity of a number of DNA polymerases.[21] Interestingly, the two compounds were found to exhibit diverse interactions with the different tested polymerases. That is, while both SCUL-B and SCUL-A strongly inhibited bovine polα and λ, the two compounds weakly interacted with polε, and had almost no effect on HIV reverse transcriptase and an *E. coli* DNA polymerase I Klenow fragment. More importantly, SCUL-A was found to be more selective against polβ than SCUL-B. This is apparent by comparing the IC_{50} values of the two compounds with respect to their interaction with polβ (17 μM for SCUL-A and 90 μM for SCUL-B). Similarly to this study, Mizushina and his co-workers showed that a pyrrolidine alkaloid, termed as DRB (11), was able to suppress the activity of a number of eukaryotic DNA polymerases with IC_{50} values of 21–35 μM.[55] Although such compounds are widely known to inhibit other enzymes such as glycosidases, DRB had almost no effect on the activities of prokaryotic DNA polymerases, nor DNA metabolic enzymes such as human immunodeficiency virus type 1 reverse transcriptase, T7 RNA polymerase, and bovine deoxyribonuclease I. Furthermore, the mode of inhibition of DRB against pol β was competitive with both the substrate and the DNA template-primer, while, for pol α, DRB competed only with the substrate. Although the structure of DRB resembles that of dNTP, the affinity of the compound was observed to be higher at the template-primer binding site than at the dNTP substrate-binding site.

More recent studies by the same group revealed a number of compounds that can target DNA polymerases and in some cases interact with a number of other enzymes as well. Examples of such compounds include isosteviol, which targets mammalian polymerases and human DNA topoisomerase II (topo II);[68] a number of sulfolipid derivatives that can interact with both polα

and β;[69] epolactaene derivatives that can target DNA polymerases and topo II;[70] catechin derivatives that can target pol α and λ;[71] and finally the two azaphilone derivatives, kasanosins A and B that specifically found to target pol β and λ.[72]

Starting from the late nineties, one can recognize the emergence of a new team representing the National Cancer Institute-sponsored National Cooperative Drug Discovery Groups (NCDDG) that has entered the field of screening for novel inhibitors of pol β (see ref. [73] for an early review of their work). Similar to the aforementioned attempts, the NCDDG group continued the screening of natural products to search for small molecules that can regulate the activity of pol β and sensitize the cells for several DNA damaging agents. As far as the authors can ascertain from the literature, the first study focusing on pol β inhibitors performed by this group was the work by Chen et al., who isolated five compounds that showed inhibitory activities against pol β with IC_{50} values ranging from 9 to 72 μM.[22] The five compounds were extracted from the plant *Schoepfia californica* using bioassay-guided fractionation techniques and four of these molecules were shown to be anacardic acid (12) and structurally related derivatives, while the fifth was oleic acid (13). In a different study, Deng et al. isolated three bis-5-alkylresorcinols compounds from *Panopsis rubescens*.[56] the three compounds showed strong binding to calf thymus pol β, with IC_{50} values ranging from 5.8 to 7.5 μM (14 is the most potent structure). Moreover, Deng et al. used the same fractionation procedure against *Baeckea gunniana* to separate a methyl ethyl ketone extract, which was identified as a potent inhibitor of rat pol β.[57] This study revealed four active ursane and oleanane triterpenoid compounds that can bind to pol β, with IC_{50} values ranging from 5.3 to 8.5 μM in the presence of bovine serum albumin (BSA) and from 2.5 to 4.8 μM in the absence of BSA (15 is the most potent structure).

Sun et al. isolated three active natural products with IC_{50} values ranging from 20 to 36 μM.[58] They also generated ten more derivatives and examined their interactions with the protein. Only three derivative compounds were active against pol β. What makes this study distinctive from the previous work done by this group is its focus on the mode of action of these inhibitors of pol β. The authors found that the compounds exhibited a mixed-type inhibition pattern for both the substrate, dNTP, as well as the DNA template-primer. That is, when altering the concentrations of both dNTP and the DNA template-primer separately, the inhibition pattern was intermediate between competitive and noncompetitive inhibition for the two substances. Comparing the performance of the three compounds along with their derivatives, one can notice that one of these molecules, known as koetjapic acid (KJA) (16), showed reasonable activity in interacting with pol β and was the subject of a later study that was carried out by Hu and his co-workers from the same NCDDG team.[25] In this study, Hu et al. used NMR analysis to identify the binding interface between KJA and the 8-kDa domain of pol β and decompose its residue contributions. Their findings suggest that the binding pocket of the compound within the surface of polβ is located between the two helices, helix-2 and helix-4 of the 8-kDa domain. Interestingly, the same region has been recognized in different studies to be essential in the DNA binding and deoxyribose phosphate lyase activities of the enzyme[47]. Hu also examined nine structurally related synthetic compounds that are similar to KJA for their activity against polβ. The structures of these compounds involved different categories of functional groups that varied between aromatic and other hydrophobic chains in combination with two carboxylate groups. Intriguingly, these compounds were found to bind to the same or a very similar region on the surface of the enzyme. Moreover, the compounds also were able to enhance the efficacy of methyl methanesulfonate (MMS), a monofunctional methylating agent that targets the DNA whose induced damage is mainly repaired by BER. More importantly, the most potent compound, one of the tested derivatives, also known as pamoic acid (PA) (17), was found to be an inhibitor of the deoxyribose phosphate lyase and DNA polymerase activities of purified polβ on a BER substrate. It should be noted that the inhibition of these two activities of polβ by PA was only observed when the compound has been pre-incubated with the enzyme before initiation of the BER reactions. Moreover, PA was not an effective polβ-inhibitor when pre-incubated with DNA alone. These observations may indicate that the binding reaction between polβ and PA is very slow and requires an apo-enzyme to be completed. These interesting findings were further pursued by a different group from the Centre National de la Recherche Scientifique (CNRS) in France, to understand and identify the precise interactions between PA and the 8-kDa domain of polβ.[74] In this study, Hazan et al. used a combination protocol of blind docking and NMR analysis to identify the binding site of PA within the surface of the lyase domain of polβ and to suggest its binding conformation. These results confirmed the earlier findings of Hu et al.[25] and revealed that PA binds to a site formed by helix 2 and helix 4, which also corresponds to the single-stranded DNA binding site.[47] Particularly, The aromatic groups of pamoic acid formed favorable hydrophobic interactions with the residues Tyr39, Ala42, Gly64 and Gly66 within the identified binding site. Furthermore, the presence of many lysine residues in the binding pocket allowed favorable electrostatic interactions for the two carboxyl groups of PA. In their proposed model, one of the carboxyl groups is oriented towards His34 and Lys35 making close contacts with Ile69 amide proton, while the other carboxyl group formed hydrogen bonds with the amide proton of Lys68 and with the hydroxyl group of Thr67.

In a series of similar studies and using bioassay-guided fractionation techniques, the NCDDG group identified a considerable number of polβ inhibitors that can bind to the enzyme with reasonably high affinities. This list includes harbinatic acid (18) with IC_{50} of 2.9 μM;[23] the three triterpenoid compounds betulinic acid (19), 3-cis-p-coumaroyl maslinic acid (20) and 3-trans-p-coumaroyl maslinic acid (21) with IC_{50} values ranging from 4.2 to 15μM;[59] an additional six pentacyclic triterpenoids compounds extracted from *Freziera sp.* with IC_{50} values ranging from 7.5 to 16μM (22 is the most potent structure);[60]a sesquiterpenoid derivative with IC_{50} of 45.2 μM targeting the lyase activity of polβ;[75]four lyase inhibitors comprising a triterpene, ursolic acid, hydroxyursolic acid, and β-sitosteryl-β-D-galactoside with IC_{50} values ranging from 12.6 to 26.5μM (23 is 2-αhydroxyursolic acid, their most potent inhibitor);[61] four lupane triterpenoids with IC_{50} values ranging from 3.8 to 21.5μM targeting the lyase activity of polβ (24 is their best inhibitor) ;[62] the lyase-inhibitor, (-)-epicatechin (25), with IC_{50} values of 18.5μM which also potentiated the efficacy of monofunctional methylating agent in cultured human cancer cells;[63] the biscoumarin derivative, Edgeworin, which inhibited the lyase activity with IC_{50} of 22.5μM (26);[64] and finally, two neolignan lyase inhibitors with IC_{50} values ranging from 15.3 to 18.6 μM (27, 28). However, for all of these listed inhibitors, the exact binding locations, mode of inhibition against polβ, or the possibility of targeting other DNA polymerases or enzymes within the cell were not identified.

In another important study, Maloney et al. investigated the synthesis of three flavanoids derivatives, namely myristinin A (29), B and C, which exhibited a distinctive characteristic besides their ability to inhibit the activity of polβ.[24] That is, these compounds can also cleave and induce a considerable damage to the DNA, allowing one to exploit their dual-activity as an innovative therapeutic strategy in cancer treatments. As myristinin A showed more potent Cu^{2+}-dependent DNA-damaging activity and polβ inhibition (with IC_{50} values of 2.8 μM) than the inseparable mixture of myristinin B and C, Maloney et al. focused only on the synthesis of the former compound. However, the synthesis of the two other structures B and C has been revealed in a more recent study by the same group.[76] Similarly to the interesting behavior of the abovementioned flavanoid derivatives, Starck et al. identified a number of 5-alkylresorcinols that also mediated Cu^{2+}-dependent DNA damage and also suppressed the ability of polβ to restore the DNA damage that they cause.[77] Interestingly, one of these alkylresorcinols, namely bis(dihydroxyalkylbenzenes), showed potent activity both as an inhibitor for polβ and as a DNA-damaging agent. This compound resulted in the reduction of the number of viable cells when incubated in the presence of bleomycin and a further decrease in the number of viable cells in the presence of both bleomycin and Cu^{2+}.[77]

In their most recent study, the NCDDG group used bioassay-guided fractionation to isolate four polβ inhibitors, namely oleanolic acid, edgeworin, betulinic acid, and stigmasterol.[66] Interestingly, although stigmasterol (30) was not as strong as the other compounds, it was the most specific structure for the lyase activity, which inhibited both the lyase and polymerase activities of the enzyme. More importantly, Gao et al. showed that, the four inhibitors potentiated the efficacy of the anti-cancer drug bleomycin in cultured A549 cells, without any influence on the expression of polβ in these cells. These results were confirmed using an unscheduled DNA synthesis assay, which suggested that that the potentiation of bleomycin cytotoxicity by these compounds was a direct result of an inhibition of DNA repair synthesis.[66]

This concludes the summary of the current state-of-art regarding the search for inhibitors of DNA polβ activity. However, it seems that, in spite of more than twenty years of extensive research in this area and regardless of the high number of various structures that have been isolated and extracted, there is no single molecule that can be pointed out as a polβ-specific inhibitor. This keeps the door wide-open for innovative work that can employ novel techniques such as *de novo* drug design or virtual screening for molecular structures that can target specific domains of the enzyme. Particularly with the considerable number of crystal structures that have been added to the protein data bank database, this can be used to illustrate the various states of the catalytic activity of the protein.

5.5 Conclusion

BER is the primary pathway that removes damaged DNA bases and repairs single strand breaks that are generated spontaneously or produced by many DNA damaging agents.[6] Accordingly, this pathway has been recognized as an important determinant of cancer cells' sensitivity to many anticancer agents including ionizing radiation, bleomycin, monofunctional alkylating agents and cisplatin. Therefore, several investigations have considered the proteins that are evolved in this pathway as promising therapeutic targets.[4, 6]

Family-X of DNA polymerases and polβ in particular are the foremost elements of BER.[42] This is mainly due to their ability to fill short gaps within the damaged DNA molecule. Fortunately, a lot of structural data and biological information about polβ are currently available, making it the first DNA polymerase enzyme whose structural description is complete.[10, 78] These observations attracted researchers to look for regulators of BER through the discovery of inhibitors of the polymerization step of the pathway. The fundamental principle behind this objective is to preserve the ionizing radiation or chemotherapeutic-induced damage within the genome in order to potentiate the efficacy of these DNA-damaging agents and hence, force the cell to undergo apoptosis.[73] Although, these efforts resulted in a large number of DNA polβ-inhibitors listed here in this review, these inhibitors are not specific or potent enough to be persued as drug candidates. This is mainly because most of the identified compounds target other polymerases or enzymes and a considerable number of them cannot enter the cell due to solubility problems.

Therefore, we decided to take an alternative drug discovery avenue that has been only slightly touched upon in a few of the above-mentioned studies. In the next chapter, I will present the outcomes of applying the VS protocol discussed in this thesis and search for novel inhibitors of polβ.

Bibliography

1. Xu, G.; Herzig, M.; Rotrekl, V.; Walter, C. A., Base excision repair, aging and health span. *Mech Ageing Dev* **2008,** *129* (7-8), 366-82.
2. Lindahl, T., Instability and decay of the primary structure of DNA. *Nature* **1993,** *362* (6422), 709-15.
3. Parsons, J. L.; Dianova, II; Dianov, G. L., APE1 is the major 3'-phosphoglycolate activity in human cell extracts. *Nucleic Acids Res* **2004,** *32* (12), 3531-6.
4. Liu, L.; Nakatsuru, Y.; Gerson, S. L., Base excision repair as a therapeutic target in colon cancer. *Clin Cancer Res* **2002,** *8* (9), 2985-91.
5. Hoffmann, J. S.; Pillaire, M. J.; Garcia-Estefania, D.; Lapalu, S.; Villani, G., In vitro bypass replication of the cisplatin-d(GpG) lesion by calf thymus DNA polymerase beta and human immunodeficiency virus type I reverse transcriptase is highly mutagenic. *J Biol Chem* **1996,** *271* (26), 15386-92.
6. Sharma, R. A.; Dianov, G. L., Targeting base excision repair to improve cancer therapies. *Mol Aspects Med* **2007,** *28* (3-4), 345-74.
7. Dianov, G.; Lindahl, T., Reconstitution of the DNA base excision-repair pathway. *Curr Biol* **1994,** *4* (12), 1069-76.
8. Beard, W. A.; Wilson, S. H., Structure and mechanism of DNA polymerase Beta. *Chem Rev* **2006,** *106* (2), 361-82.
9. Wilson, S. H., Mammalian base excision repair and DNA polymerase beta. *Mutat Res* **1998,** *407* (3), 203-15.
10. Uchiyama, Y.; Takeuchi, R.; Kodera, H.; Sakaguchi, K., Distribution and roles of X-family DNA polymerases in eukaryotes. *Biochimie* **2009,** *91* (2), 165-70.
11. Gu, H.; Marth, J. D.; Orban, P. C.; Mossmann, H.; Rajewsky, K., Deletion of a DNA polymerase beta gene segment in T cells using cell type-specific gene targeting. *Science* **1994,** *265* (5168), 103-6.
12. Bergoglio, V.; Canitrot, Y.; Hogarth, L.; Minto, L.; Howell, S. B.; Cazaux, C.; Hoffmann, J. S., Enhanced expression and activity of DNA polymerase beta in human ovarian tumor cells: impact on sensitivity towards antitumor agents. *Oncogene* **2001,** *20* (43), 6181-7.
13. Starcevic, D.; Dalal, S.; Sweasy, J. B., Is there a link between DNA polymerase beta and cancer? *Cell Cycle* **2004,** *3* (8), 998-1001.
14. Chan, K.; Houlbrook, S.; Zhang, Q. M.; Harrison, M.; Hickson, I. D.; Dianov, G. L., Overexpression of DNA polymerase beta results in an increased rate of frameshift mutations during base excision repair. *Mutagenesis* **2007,** *22* (3), 183-8.
15. Husain, I.; Morton, B. S.; Beard, W. A.; Singhal, R. K.; Prasad, R.; Wilson, S. H.; Besterman, J. M., Specific inhibition of DNA polymerase beta by its 14 kDa domain: role of single- and double-stranded DNA binding and 5'-phosphate recognition. *Nucleic Acids Res* **1995,** *23* (9), 1597-603.
16. Mizushina, Y.; Tanaka, N.; Yagi, H.; Kurosawa, T.; Onoue, M.; Seto, H.; Horie, T.; Aoyagi, N.; Yamaoka, M.; Matsukage, A.; Yoshida, S.; Sakaguchi, K., Fatty acids selectively inhibit eukaryotic DNA polymerase activities in vitro. *Biochim Biophys Acta* **1996,** *1308* (3), 256-62.
17. Tanaka, N.; Kitamura, A.; Mizushina, Y.; Sugawara, F.; Sakaguchi, K., Fomitellic acids, triterpenoid inhibitors of eukaryotic DNA polymerases from a basidiomycete, fomitella fraxinea. *J Nat Prod* **1998,** *61* (9), 1180.
18. Mizushina, Y.; Watanabe, I.; Ohta, K.; Takemura, M.; Sahara, H.; Takahashi, N.; Gasa, S.; Sugawara, F.; Matsukage, A.; Yoshida, S.; Sakaguchi, K., Studies on inhibitors of mammalian DNA polymerase alpha and beta: sulfolipids from a pteridophyte, Athyrium niponicum. *Biochem Pharmacol* **1998,** *55* (4), 537-41.
19. Ogawa, A.; Murate, T.; Izuta, S.; Takemura, M.; Furuta, K.; Kobayashi, J.; Kamikawa, T.; Nimura, Y.; Yoshida, S., Sulfated glycoglycerolipid from archaebacterium inhibits eukaryotic DNA polymerase alpha, beta and retroviral reverse transcriptase and affects methyl methanesulfonate cytotoxicity. *Int J Cancer* **1998,** *76* (4), 512-8.

20. Ogawa, A.; Murate, T.; Suzuki, M.; Nimura, Y.; Yoshida, S., Lithocholic acid, a putative tumor promoter, inhibits mammalian DNA polymerase beta. *Jpn J Cancer Res* **1998,** *89* (11), 1154-9.

21. Perpelescu, M.; Kobayashi, J.; Furuta, M.; Ito, Y.; Izuta, S.; Takemura, M.; Suzuki, M.; Yoshida, S., Novel phenalenone derivatives from a marine-derived fungus exhibit distinct inhibition spectra against eukaryotic DNA polymerases. *Biochemistry* **2002,** *41* (24), 7610-6.

22. Chen, J. Z., Y.; Wang, L.; Sucheck, S.; Snow, A.; Hecht, S., Inhibitors of DNA Polymerase b from Schoepfia Californica. *J C S Chem Commun* **1998**, 2769–2770.

23. Deng, J. Z.; Starck, S. R.; Hecht, S. M.; Ijames, C. F.; Hemling, M. E., Harbinatic acid, a novel and potent DNA polymerase beta inhibitor from Hardwickia binata. *J Nat Prod* **1999,** *62* (7), 1000-2.

24. Maloney, D. J.; Deng, J. Z.; Starck, S. R.; Gao, Z.; Hecht, S. M., (+)-Myristinin A, a naturally occurring DNA polymerase beta inhibitor and potent DNA-damaging agent. *J Am Chem Soc* **2005,** *127* (12), 4140-1.

25. Hu, H. Y.; Horton, J. K.; Gryk, M. R.; Prasad, R.; Naron, J. M.; Sun, D. A.; Hecht, S. M.; Wilson, S. H.; Mullen, G. P., Identification of small molecule synthetic inhibitors of DNA polymerase beta by NMR chemical shift mapping. *J Biol Chem* **2004,** *279* (38), 39736-44.

26. Wood, R. D.; Mitchell, M.; Sgouros, J.; Lindahl, T., Human DNA repair genes. *Science* **2001,** *291* (5507), 1284-9.

27. Dianov, G. L.; Souza-Pinto, N.; Nyaga, S. G.; Thybo, T.; Stevnsner, T.; Bohr, V. A., Base excision repair in nuclear and mitochondrial DNA. *Prog Nucleic Acid Res Mol Biol* **2001,** *68*, 285-97.

28. Almeida, K. H.; Sobol, R. W., A unified view of base excision repair: lesion-dependent protein complexes regulated by post-translational modification. *DNA Repair (Amst)* **2007,** *6* (6), 695-711.

29. Hitomi, K.; Iwai, S.; Tainer, J. A., The intricate structural chemistry of base excision repair machinery: implications for DNA damage recognition, removal, and repair. *DNA Repair (Amst)* **2007,** *6* (4), 410-28.

30. http://www.genome.jp/kegg/kegg2.html.

31. Loeb, L. A.; Preston, B. D., Mutagenesis by apurinic/apyrimidinic sites. *Annu Rev Genet* **1986,** *20*, 201-30.

32. Tomkinson, A. E.; Chen, L.; Dong, Z.; Leppard, J. B.; Levin, D. S.; Mackey, Z. B.; Motycka, T. A., Completion of base excision repair by mammalian DNA ligases. *Prog Nucleic Acid Res Mol Biol* **2001,** *68*, 151-64.

33. Hegde, M. L.; Hazra, T. K.; Mitra, S., Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. *Cell Res* **2008,** *18* (1), 27-47.

34. Vidal, A. E.; Boiteux, S.; Hickson, I. D.; Radicella, J. P., XRCC1 coordinates the initial and late stages of DNA abasic site repair through protein-protein interactions. *EMBO J* **2001,** *20* (22), 6530-9.

35. Klungland, A.; Lindahl, T., Second pathway for completion of human DNA base excision-repair: reconstitution with purified proteins and requirement for DNase IV (FEN1). *EMBO J* **1997,** *16* (11), 3341-8.

36. Fortini, P.; Dogliotti, E., Base damage and single-strand break repair: mechanisms and functional significance of short- and long-patch repair subpathways. *DNA Repair (Amst)* **2007,** *6* (4), 398-409.

37. Fortini, P.; Pascucci, B.; Parlanti, E.; D'Errico, M.; Simonelli, V.; Dogliotti, E., 8-Oxoguanine DNA damage: at the crossroad of alternative repair pathways. *Mutat Res* **2003,** *531* (1-2), 127-39.

38. Hashimoto, K.; Tominaga, Y.; Nakabeppu, Y.; Moriya, M., Futile short-patch DNA base excision repair of adenine:8-oxoguanine mispair. *Nucleic Acids Res* **2004,** *32* (19), 5928-34.

39. Fortini, P.; Pascucci, B.; Parlanti, E.; Sobol, R. W.; Wilson, S. H.; Dogliotti, E., Different DNA polymerases are involved in the short- and long-patch base excision repair in mammalian cells. *Biochemistry* **1998,** *37* (11), 3575-80.

40. Muiras, M. L., Mammalian longevity under the protection of PARP-1's multi-facets. *Ageing Res Rev* **2003,** *2* (2), 129-48.

41. Petermann, E.; Ziegler, M.; Oei, S. L., ATP-dependent selection between single nucleotide and long patch base excision repair. *DNA Repair (Amst)* **2003,** *2* (10), 1101-14.

42. Beard, W. A.; Wilson, S. H., Structural design of a eukaryotic DNA repair polymerase: DNA polymerase beta. *Mutat Res* **2000,** *460* (3-4), 231-44.

43. (a) Beard, W. A.; Wilson, S. H., Purification and domain-mapping of mammalian DNA polymerase beta. *Methods Enzymol* **1995,** *262*, 98-107; (b) Kumar, A.; Widen, S. G.; Williams, K. R.; Kedar, P.; Karpel, R. L.; Wilson, S. H., Studies of the domain structure of mammalian DNA polymerase beta. Identification of a discrete template binding domain. *J Biol Chem* **1990,** *265* (4), 2124-31.

44. Burgers, P. M.; Koonin, E. V.; Bruford, E.; Blanco, L.; Burtis, K. C.; Christman, M. F.; Copeland, W. C.; Friedberg, E. C.; Hanaoka, F.; Hinkle, D. C.; Lawrence, C. W.; Nakanishi, M.; Ohmori, H.; Prakash, L.; Prakash, S.; Reynaud, C. A.; Sugino, A.; Todo, T.; Wang, Z.; Weill, J. C.; Woodgate, R., Eukaryotic DNA polymerases: proposal for a revised nomenclature. *J Biol Chem* **2001,** *276* (47), 43487-90.

45. Pelletier, H.; Sawaya, M. R.; Wolfle, W.; Wilson, S. H.; Kraut, J., A structural basis for metal ion mutagenicity and nucleotide selectivity in human DNA polymerase beta. *Biochemistry* **1996,** *35* (39), 12762-77.

46. Sawaya, M. R.; Pelletier, H.; Kumar, A.; Wilson, S. H.; Kraut, J., Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism. *Science* **1994,** *264* (5167), 1930-5.

47. Pelletier, H.; Sawaya, M. R.; Kumar, A.; Wilson, S. H.; Kraut, J., Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* **1994,** *264* (5167), 1891-903.

48. Pelletier, H.; Sawaya, M. R., Characterization of the metal ion binding helix-hairpin-helix motifs in human DNA polymerase beta by X-ray structural analysis. *Biochemistry* **1996,** *35* (39), 12778-87.

49. Mizushina, Y.; Yoshida, S.; Matsukage, A.; Sakaguchi, K., The inhibitory action of fatty acids on DNA polymerase beta. *Biochim Biophys Acta* **1997,** *1336* (3), 509-21.

50. Mizushina, Y.; Watanabe, I.; Togashi, H.; Hanashima, L.; Takemura, M.; Ohta, K.; Sugawara, F.; Koshino, H.; Esumi, Y.; Uzawa, J.; Matsukage, A.; Yoshida, S.; Sakaguchi, K., An ergosterol peroxide, a natural product that selectively enhances the inhibitory effect of linoleic acid on DNA polymerase beta. *Biol Pharm Bull* **1998,** *21* (5), 444-8.

51. Mizushina, Y.; Tanaka, N.; Kitamura, A.; Tamai, K.; Ikeda, M.; Takemura, M.; Sugawara, F.; Arai, T.; Matsukage, A.; Yoshida, S.; Sakaguchi, K., The inhibitory effect of novel triterpenoid compounds, fomitellic acids, on DNA polymerase beta. *Biochem J* **1998,** *330 (Pt 3)*, 1325-32.

52. Mizushina, Y.; Matsukage, A.; Sakaguchi, K., The biochemical inhibition mode of bredinin-5'-monophosphate on DNA polymerase beta. *Biochim Biophys Acta* **1998,** *1403* (1), 5-11.

53. Mizushina, Y.; Ohkubo, T.; Sugawara, F.; Sakaguchi, K., Structure of lithocholic acid binding to the N-terminal 8-kDa domain of DNA polymerase beta. *Biochemistry* **2000,** *39* (41), 12606-13.

54. Mizushina, Y.; Kamisuki, S.; Kasai, N.; Shimazaki, N.; Takemura, M.; Asahara, H.; Linn, S.; Yoshida, S.; Matsukage, A.; Koiwai, O.; Sugawara, F.; Yoshida, H.; Sakaguchi, K., A plant phytotoxin, solanapyrone A, is an inhibitor of DNA polymerase beta and lambda. *J Biol Chem* **2002,** *277* (1), 630-8.

55. Mizushina, Y.; Xu, X.; Asano, N.; Kasai, N.; Kato, A.; Takemura, M.; Asahara, H.; Linn, S.; Sugawara, F.; Yoshida, H.; Sakaguchi, K., The inhibitory action of pyrrolidine alkaloid, 1,4-dideoxy-1,4-imino-D-ribitol, on eukaryotic DNA polymerases. *Biochem Biophys Res Commun* **2003,** *304* (1), 78-85.

56. Deng, J. Z.; Starck, S. R.; Hecht, S. M., bis-5-Alkylresorcinols from Panopsis rubescens that inhibit DNA polymerase beta. *J Nat Prod* **1999,** *62* (3), 477-80.

57. Deng, J. Z.; Starck, S. R.; Hecht, S. M., DNA polymerase beta inhibitors from Baeckea gunniana. *J Nat Prod* **1999,** *62* (12), 1624-6.

58. Sun, D. A.; Starck, S. R.; Locke, E. P.; Hecht, S. M., DNA polymerase beta inhibitors from Sandoricum koetjape. *J Nat Prod* **1999,** *62* (8), 1110-3.

59. Ma, J.; Starck, S. R.; Hecht, S. M., DNA polymerase beta inhibitors from Tetracera boiviniana. *J Nat Prod* **1999,** *62* (12), 1660-3.

60. Deng, J. Z.; Starck, S. R.; Hecht, S. M., Pentacyclic triterpenoids from Freziera sp. that inhibit DNA polymerase beta. *Bioorg Med Chem* **2000,** *8* (1), 247-50.

61. Chaturvedula, V. S.; Gao, Z.; Jones, S. H.; Feng, X.; Hecht, S. M.; Kingston, D. G., A new ursane triterpene from Monochaetum vulcanicum that inhibits DNA polymerase beta lyase. *J Nat Prod* **2004,** *67* (5), 899-901.

62. Chaturvedula, V. S.; Zhou, B. N.; Gao, Z.; Thomas, S. J.; Hecht, S. M.; Kingston, D. G., New lupane triterpenoids from Solidago canadensis that inhibit the lyase activity of DNA polymerase beta. *Bioorg Med Chem* **2004,** *12* (23), 6271-5.

63. Feng, X.; Gao, Z.; Li, S.; Jones, S. H.; Hecht, S. M., DNA polymerase beta lyase inhibitors from Maytenus putterlickoides. *J Nat Prod* **2004,** *67* (10), 1744-7.

64. Li, S. S.; Gao, Z.; Feng, X.; Hecht, S. M., Biscoumarin derivatives from Edgeworthia gardneri that inhibit the lyase activity of DNA polymerase beta. *J Nat Prod* **2004,** *67* (9), 1608-10.

65. Prakash Chaturvedula, V. S.; Hecht, S. M.; Gao, Z.; Jones, S. H.; Feng, X.; Kingston, D. G., New neolignans that inhibit DNA polymerase beta lyase. *J Nat Prod* **2004,** *67* (6), 964-7.

66. Gao, Z.; Maloney, D. J.; Dedkova, L. M.; Hecht, S. M., Inhibitors of DNA polymerase beta: activity and mechanism. *Bioorg Med Chem* **2008,** *16* (8), 4331-40.

67. Mizushina, Y.; Takahashi, N.; Ogawa, A.; Tsurugaya, K.; Koshino, H.; Takemura, M.; Yoshida, S.; Matsukage, A.; Sugawara, F.; Sakaguchi, K., The cyanogenic glucoside, prunasin (D-mandelonitrile-beta-D-glucoside), is a novel inhibitor of DNA polymerase beta. *J Biochem* **1999,** *126* (2), 430-6.

68. Mizushina, Y.; Akihisa, T.; Ukiya, M.; Hamasaki, Y.; Murakami-Nakai, C.; Kuriyama, I.; Takeuchi, T.; Sugawara, F.; Yoshida, H., Structural analysis of isosteviol and related compounds as DNA polymerase and DNA topoisomerase inhibitors. *Life Sci* **2005,** *77* (17), 2127-40.

69. Mizushina, Y.; Kasai, N.; Iijima, H.; Sugawara, F.; Yoshida, H.; Sakaguchi, K., Sulfo-quinovosyl-acyl-glycerol (SQAG), a eukaryotic DNA polymerase inhibitor and anti-cancer agent. *Curr Med Chem Anticancer Agents* **2005,** *5* (6), 613-25.

70. Mizushina, Y.; Kuramochi, K.; Ikawa, H.; Kuriyama, I.; Shimazaki, N.; Takemura, M.; Oshige, M.; Yoshida, H.; Koiwai, O.; Sugawara, F.; Kobayashi, S.; Sakaguchi, K., Structural analysis of epolactaene derivatives as DNA polymerase inhibitors and anti-inflammatory compounds. *Int J Mol Med* **2005,** *15* (5), 785-93.

71. Mizushina, Y.; Saito, A.; Tanaka, A.; Nakajima, N.; Kuriyama, I.; Takemura, M.; Takeuchi, T.; Sugawara, F.; Yoshida, H., Structural analysis of catechin derivatives as mammalian DNA polymerase inhibitors. *Biochem Biophys Res Commun* **2005,** *333* (1), 101-9.

72. Kimura, T.; Nishida, M.; Kuramochi, K.; Sugawara, F.; Yoshida, H.; Mizushina, Y., Novel azaphilones, kasanosins A and B, which are specific inhibitors of eukaryotic DNA polymerases beta and lambda from Talaromyces sp. *Bioorg Med Chem* **2008,** *16* (8), 4594-9.

73. Hecht, S. M., Inhibitors of the Lyase Activity of DNA Polymerase b. *Pharmaceutical Biology* **2003,** *41*, 10.

74. Hazan, C.; Boudsocq, F.; Gervais, V.; Saurel, O.; Ciais, M.; Cazaux, C.; Czaplicki, J.; Milon, A., Structural insights on the pamoic acid and the 8 kDa domain of DNA polymerase beta complex: towards the design of higher-affinity inhibitors. *BMC Struct Biol* **2008,** *8*, 22.

75. Cao, S.; Gao, Z.; Thomas, S. J.; Hecht, S. M.; Lazo, J. S.; Kingston, D. G., Marine sesquiterpenoids that inhibit the lyase activity of DNA polymerase beta. *J Nat Prod* **2004,** *67* (10), 1716-8.

76. Maloney, D. J.; Chen, S.; Hecht, S. M., Stereoselective synthesis of the atropisomers of myristinin B/C. *Org Lett* **2006,** *8* (9), 1925-7.

77. Starck, S. R.; Deng, J. Z.; Hecht, S. M., Naturally occurring alkylresorcinols that mediate DNA damage and inhibit its repair. *Biochemistry* **2000,** *39* (9), 2413-9.

78. Bebenek, K.; Kunkel, T. A., Functions of DNA polymerases. *Adv Protein Chem* **2004,** *69*, 137-65.

Chapter 6: DNA Polymerase Beta Inhibitors: The Last Puzzle[(1)]

6.1 Introduction

This chapter builds upon the information presented in the previous chapter. As was described in chapter 7, DNA polβ is the major DNA polymerase of BER. It plays an important role in chemotherapeutic agent resistance, as its over-expression reduces the efficacy of anticancer drug therapies including bleomycin and cisplatin.[1,2] The enzyme is mutated in approximately 30% of tumors, which in turn reduces pol fidelity in DNA synthesis exposing the genome to serious mutations.[3,4] The previous chapter also described the history of considering polβ as an anticancer target. It listed all DNA-polβ-inhibitors discovered so far and reached the conclusion that most of these inhibitors are not potent enough or lack sufficient specificity to eventually become approved drugs.

6.1.1 Pamoic acid as a promising polβ-inhibitor

Among the compounds listed in Table 5-1, pamoic acid (PA) was one of the few compounds that had promising activity against polβ and a well-defined binding mode. The structure of the compound is shown in Figure 5-3-compound 17, and was initially discovered by Hu and his co-workers.[7] Their NMR analysis revealed that PA binds to the 8-kDa domain of polβ and suggested that the binding pocket is located between the two helices, helix-2 and helix-4 of the 8-kDa domain. Interestingly, the same region has been recognized in different studies to be essential in the DNA binding and deoxyribose phosphate lyase activities of the enzyme.[5] The precise interactions between PA and the lyase domain of pol were further investigated in a different study.[6] In this study, Hazan et al. used a combination protocol of blind docking and NMR analysis to confirm the earlier findings of Hu et al (see Figure 6-1).[7]

[(1)] This chapter has been published in Barakat K, Tuszynski J. J Mol Graph Model. 2011 Feb;29(5):702-16.

6.2 Results And Discussion

In the present work, we focused our search space on the binding site of PA, using it as a positive control. The aim was to discover more potent drug candidates through filtering a library of ~12,500 structures via the VS protocol described in chapter 3. The molecules we tested included the NCI diversity set, the DrugBank set of small-molecules and more than 9,000 fragment structures with drug-like properties extracted from ZINC database (see section 3.2.1). The top 300 hits that showed strong affinity for pol have been validated and rescored using a more robust scoring function, the MM-PBSA method.

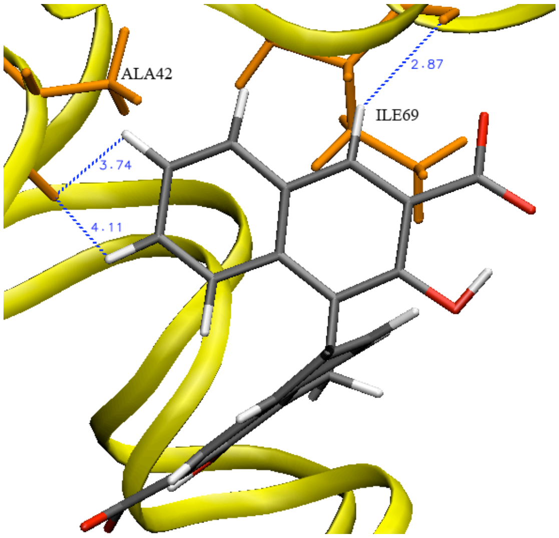


Figure 6-1: Docked structure of PA within the lyase active site of polβ. Our docking analysis confirmed the two studies by Hu[7] and Hazan.[6]

6.2.1 MD Simulations On The Lyase Domain

To generate an ensemble of equilibrated pol-models for chemical library screenings, the 8-kDa domain of polβ (PDB code 1DK3)[8] was subjected to MD simulations. The proper equilibration of these systems was essential in order to perform virtual screening on a set of rigid receptor models that represent approximately the whole conformational space of the PA binding site (which concides with the DNA binding site)[5] within the lyase domain of polβ. It should be noted that, as we used docking as a preliminary filtering step in screening the full set of compounds for polβ-inhibitors, it was essential to generate an ensemble of polβ structures in order to partly incorporate protein flexibility during docking. In this context, we selected the top 300 compounds that can bind to their appropriate polβ conformation for post-docking analysis using MD simulations to introduce the full flexibility for both the ligand and its selected target structure.

6.2.2 PCA and completeness of sampling

As described in section 3.3.4, we used PCA to transform the MD trajectory into a reduced set of independent variables comprising the essential dynamics of the system.[9,10] PCA was performed on the coordinates of the residues forming the binding site (residues 30 to 35; 37 to 43; and 63 to 70). Resulting eigenvectors were sorted by descending eigenvalues, which represent the variance of the motion along the principal components. The ten dominant eigenvalues are shown in Figure 6-2-a. The first eigenvalue has a magnitude that is significantly higher than those of the other eigenvalues. The components with the largest eigenvalues represent correlated motions of the binding site with the most significant standard deviations of the motion along the corresponding orthogonal directions.

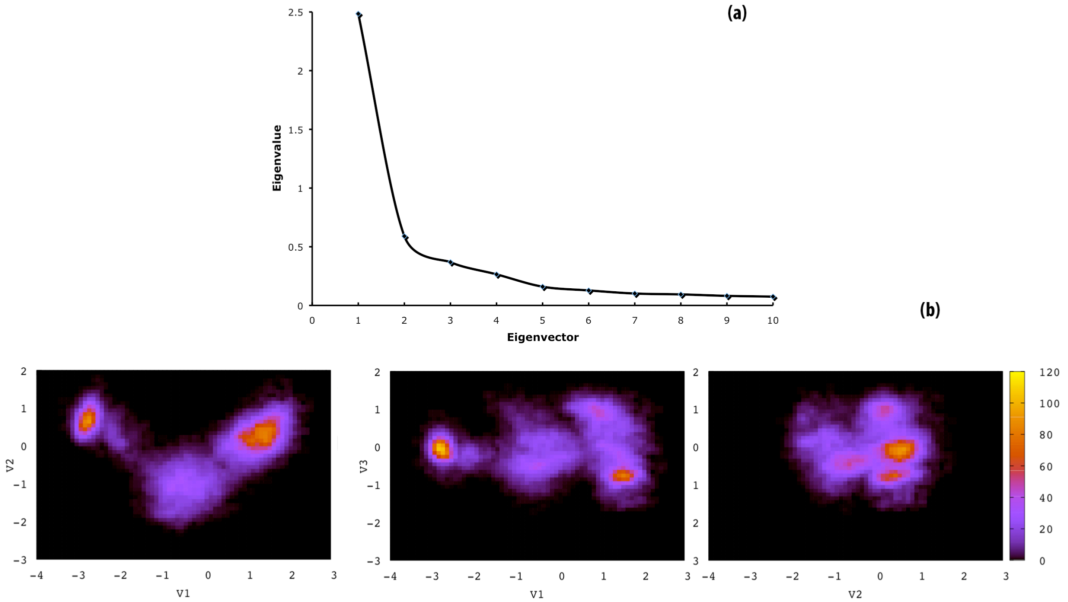


Figure 6-2: PC analysis.

(a) The ten most dominant eigenvalues. (b) Projection of the MD trajectory of the dominant three eigenvectors.

The PA (DNA) binding site within pol β adopted limited conformations throughout the MD simulations, indicating the rigidity of the protein (see Figure 6-2-b). The grouping of MD trajectories into a limited number of clusters suggests the presence of favored folded conformations with significant basins of attraction. This observation was also confirmed by results presented in Figure 3, where the main-chain B-factors (averaged over heavy atoms) for almost all residues constructing the 8-kDa domain being rigid except for a number of residues that are located on the protein termini or which have no direct influence on the binding site. These are residues 1 to 9; 28 to 31; and 80 to 87.

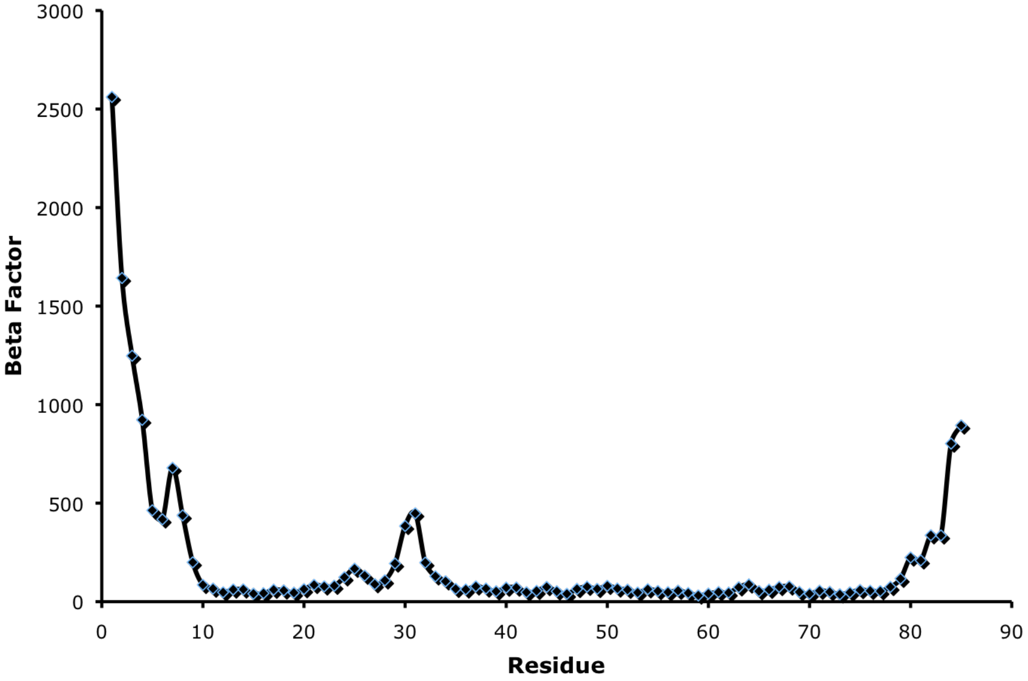


Figure 6-3: Plot of the B-factors averaged over the protein backbone atoms as a function of residue number in the simulations of pol β.

Normalized overlaps calculated between each of these thirds are ranged from 0.79 to 0.81. The high overlap between the thirds indicates that each part of the simulation samples approximately the same conformational space, and it is unlikely that there are unexplored regions missed earlier in the runs. Although there is no guarantee that complete equilibrium sampling is given, we have concluded that the observed overlap is acceptable and adequate sampling within the MD trajectories for the binding site had been obtained.

6.2.3 Ensemble-based Virtual Screening

Following the iterative clustering described in section 3.3.5, we generated a reduced set of representative models of the PA(DNA) binding site. Figure 6-4 shows the evaluation of the Davies-Bouldin index (DBI) and the percentage of variance explained by the data (SSR/SST) for different cluster counts (see Methodology). DBI for the apo-system exhibited local minima at cluster counts of 45, 60, 80 and 105. However, as the percentage of variance explained by the data started to plateau after 35 clusters, we concluded that 45 clusters is a reasonable cut-off for pol β structures.

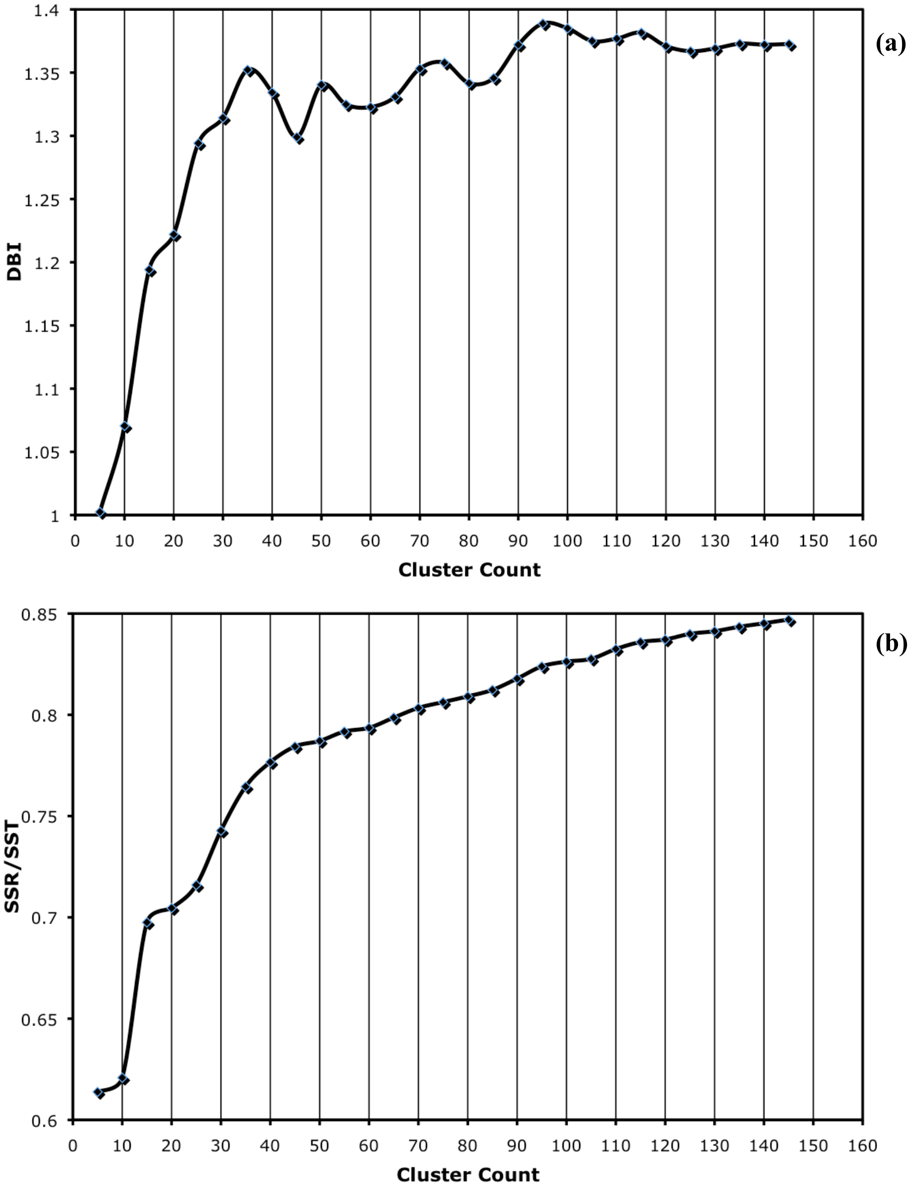


Figure 6-4: Clustering analysis for the pol β trajectory. A high-quality clustering is obtained when a local minimum in DBI correlates with saturation in the SSR/SST ratio (cluster count of 45).

Therefore, in this study, we constructed an ensemble of eleven distinct conformations to perform ensemble-based virtual screening on pol β against the full set of ligand compounds. This ensemble incorporated the ten most dominant structures that comprised ~85% of trajectory in addition to the relaxed NMR polβ conformation. The ultimate goal was to reduce the number of representative structures included in the ensemble-based screening and concurrently comprise most of the conformational space of the binding site. Figure 4 represents the eleven structures used in this work. The 8-kDa structure adopted different conformational changes demonstrating the significance of introducing receptor flexibility during the docking procedure.

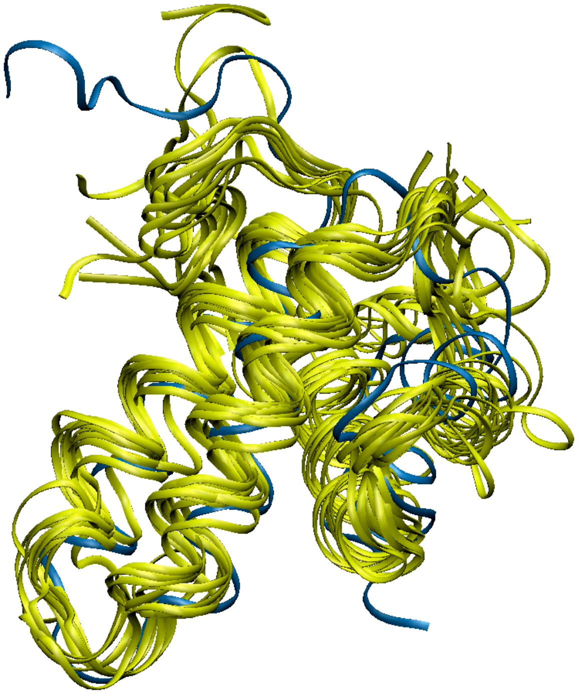


Figure 6-5: Eleven dominate conformations for pol β. This ensemble comprised the NMR structure (blue) and ten structures extracted from the MD trajectory (yellow).

6.2.4 Pose Clustering

Clustering of docked poses followed the same strategy as described in sections 3.4.2 and 4.2.6.

6.2.5 Preliminary Ranking Of Hits

For each virtual screening experiment, we have ranked significant poses for each of the molecules contained in the database by using the results from the elbow criterion and the lowest energy that corresponds to the most populated cluster. Once all poses from each ligand entry were clustered, we then filtered all of the clusters so that only those containing at least 25% of the total population were considered as top hits. Top hits were collected from the 11 experiments by first extracting the largest cluster from each individual screening followed by ranking the clusters according to their binding energies. This produced a set of non-redundant hits ranked by their binding energies of the most populated cluster.

The apparent K_D value for PA binding to pol β is 9 μM.[7] Using the AutoDock scoring function, we obtained a value of -6.2 kcal/mol for the binding energy of PA to the 8-kDa domain. Although this value is in excellent agreement with the experimental value (-6.9 kcal/mol) as calculated using K_D value, it has been widely reported in the literature that empirical scoring functions including the AutoDock scoring function (see 2.2.1.3), are not efficient in discriminating false positives in VS experiments and are biased toward their training set of compounds.[11,12] Consequently, in this work, the top 300 hits were rescored using the MM-PBSA method (see below) to validate their docking results and confirm their binding to the protein.

6.2.6 Ranking Using The MM-PBSA Scoring Function

Following the discussion in section 3.7, we used the MM-PBSA method, introduced by Kollman et al.[13] to measure the binding energies of the top 300 hits relative to that of the positive control, i.e. PA, and compare their MM-PBSA-ranking to that of AutoDock calculations. As has been reported earlier by other groups who used the method, the most computationally demanding step is the calculation of the solute entropy using the normal mode (NMODE) method (see section 3.7). Although this component can be neglected if only relative binding (relative ranking) of compounds is required,[13] we calculated the entropy contributions for all the top 300 hits using 200 snapshots extracted from their 2ns MD trajectories (see Methodology). In our calculations, this part ranged from 10 to 15 kcal/mol, indicating its significance in predicating the overall binding energies. The top 34 hits according to MM-PBSA calculations are shown in Table 6-1. As expected, the MM-PBSA analysis predicted the binding energy of PA more accurately than the AUTODOCK scoring function. Furthermore, AutoDock ranking of the top hits has been partially altered when compared to MM-PBSA calculations (see Table 6-1). These results also illustrate the limitations of the AutoDock scoring function in eliminating from the set of active compounds false positive ligands, i.e. compounds that are predicted to bind the target but fail to do so in validation assays. We noticed this in our calculations where a number of compounds highly ranked by AutoDock exhibited very low, and in some cases positive, binding energies using the MM-PBSA analysis, indicating their weak binding to the protein (see Table 6-1). The identified binding mode of PA is shown in Figure 6-1. The atomic distances between PA and the two residues ALA42 and

ILE69 are in excellent agreement with what was shown by Hu et al.[7] and Hazan et al.,[6]confirming the successful docking of PA within the DNA binding site of pol β.

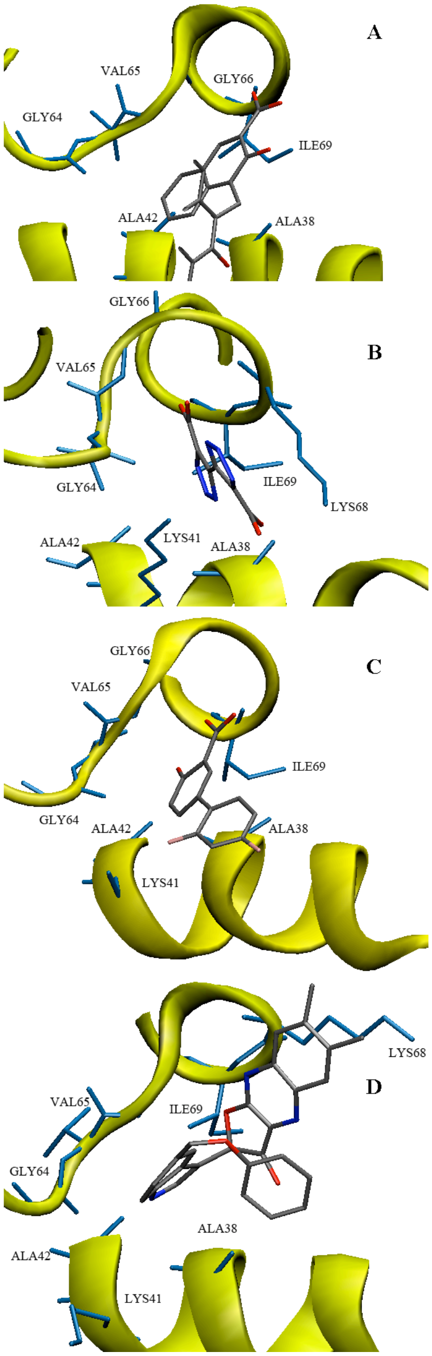


Figure 6-6: Binding modes of selected hits.

The binding mode of PA (A) and the top three hits from the MM-PBSA ranking (B-D). Pol β is shown in yellow, important protein residues are shown in blue, and the different atoms of the bound compounds are shown by their representative colors (carbon in gray, oxygen in red, nitrogen in blue and hydrogen in white).

Figure 6-6 demonstrates the binding modes of the top three hits of the MM-PBSA ranking. Similar to a substantial number of our suggested top hits, the shown compounds are small in size, however, they are occupying a considerable portion of the DNA-binding pocket. These lead compounds can be employed as the basis for a further fragment-based drug design step, in order to construct potent and more specific pol β inhibitors.

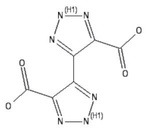
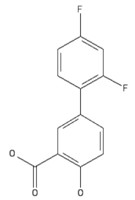
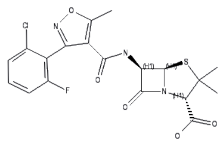
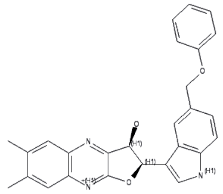
MM-PBSA scoring		Ensemble-based scoring			
Rank	BE ± 1.5 (kcal/mol)	Rank	BE ± 2.2 (kcal/mol)	ID	Chemical structure
1	-12.5	180	-7.2	ZINC19229065	
2	-12.4	263	-7.0	ZINC00020243	
3	-11.9	33	-8.2	ZINC04102187	
4	-11.3	24	-8.4	NSC#372280	

Table 6-1: Top 30 hits according to MM-PBSA ranking. Compounds are ranked by their binding energies as were calculated using the MM-PBSA method and compared to their ranking using the AUTODOCK scoring function.

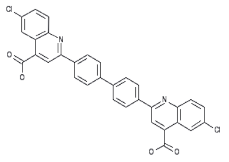
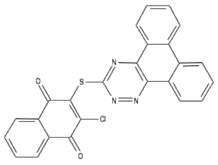
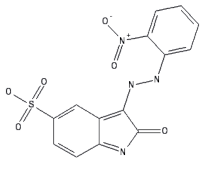
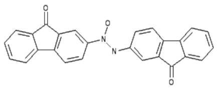
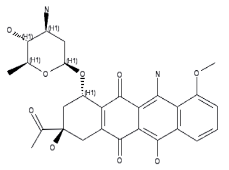
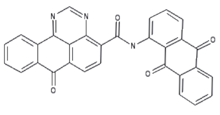
5	-10.6	5	-9.6	NSC#210627	
6	-10.4	19	-8.6	NSC#327705	
7	-9.9	14	-8.8	NSC#116654	
8	-9.8	40	-8.1	NSC#12363	
9	-9.6	23	-8.4	NSC#254681	
10	-9.3	22	-8.5	NSC#299137	

Table 6-1 Continued.

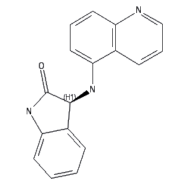
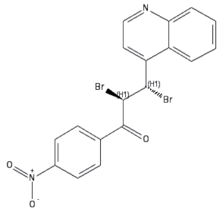
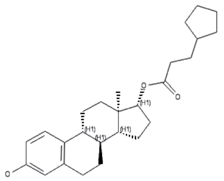
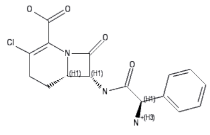
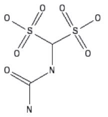
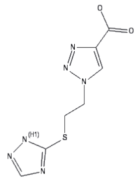
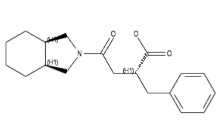
11	-9.2	202	-7.2	NSC#117198	
12	-9.2	88	-7.6	NSC#150289	
13	-9.0	77	-7.7	NSC#3354	
14	-8.9	264	-7.0	ZINC01530992	
15	-8.8	13	-8.8	ZINC05368838	
16	-8.6	245	-7.0	ZINC20596577	
17	-8.4	87	-7.6	ZINC11616579	

Table 6-1 Continued.

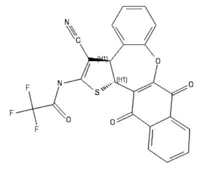
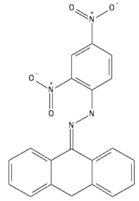
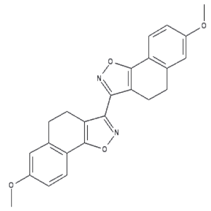
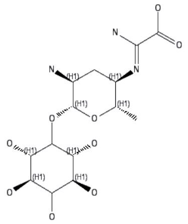
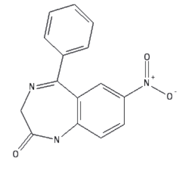
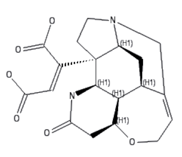
MM-PBSA scoring		Ensemble-based scoring			
Rank	BE (kcal/mol)	Rank	BE (kcal/mol)	ID	Chemical structure
18	-8.2	226	-7.1	NSC#686365	
19	-7.9	6	-9.5	NSC#201873	
20	-7.8	79	-7.7	NSC#371688	
21	-7.7	160	-7.3	NSC#100858	
22	-7.6	258	-7.0	ZINC03812992	
23	-7.5	9	-9.0	NSC#123420	

Table 6-1 Continued.

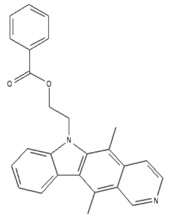
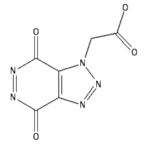
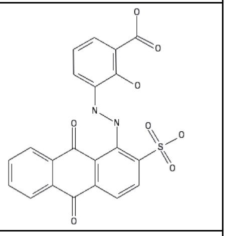
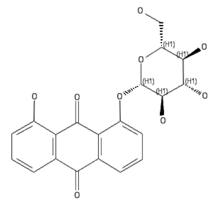
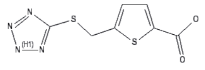
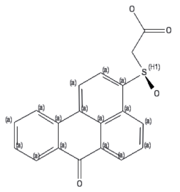
24	-7.5	214	-7.1	NSC#125908	
25	-7.5	123	-7.4	ZINC16958839	
26	-7.4	3	-9.7	NSC#125908	
27	-7.1	18	-8.7	NSC#255980	
28	-7.0	142	-7.3	ZINC20395500	
29	-7.0	155	-7.3	NSC#16211	

Table 6-1 Continued.

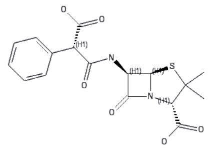
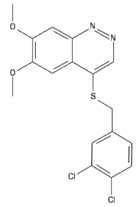
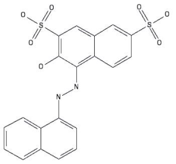
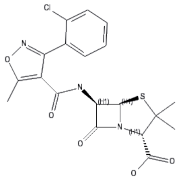
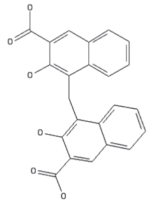
30	-6.9	143	-7.3	ZINC03978033	
31	-6.9	139	-7.3	NSC#64814	
32	-6.8	4	-9.6	NSC#45583	
33	-6.7	28	-8.3	ZINC03875417	
34	-6.7	278	-6.2	PA	

Table 6-1 Continued.

6.3 Conclusions

Building upon the information presented in the previous chapter, we applied the RCS technique to account for the full receptor flexibility in screening for inhibitors of the lyase activity of DNA polβ. Our library of screening compounds comprised of the National Cancer Institute

(NCI) diversity set, DrugBank small molecules and a set of ~9,000 small fragments with drug-like properties. The full set of compounds (~12,500) has been screened against ten polβ structures. AutoDock was used to place the compounds within the specified binding site and to search for their minimal energy conformations. Then, the irredundant top 300 hits from AutoDock screening were rescored using the MM–PBSA method. We used pamoic acid (PA), a well-known polβ-inhibitor, as our positive control. This is because it was the subject of recent extensive studies and the information on its binding to polβ has been determined with a high degree of accuracy. Although, more than 12,500 compounds have been screened in this study, we suggest a future use of larger libraries (on the order of 100,000 to 1,000,000 compounds) may be even more successful in discovering higher affinity hits.

Our results confirmed the experimental findings concerning the binding of PA to the DNA binding cleft within the 8-kDa domain of polβ with an affinity that is close to the reported experimental data. Furthermore, we suggested a set of compounds that can target the DNA-binding site within the 8-kDa with higher affinity than PA.

6.4 Bibliography

1. Hoffmann, J. S.; Pillaire, M. J.; Garcia-Estefania, D.; Lapalu, S.; Villani, G., In vitro bypass replication of the cisplatin-d(GpG) lesion by calf thymus DNA polymerase beta and human immunodeficiency virus type I reverse transcriptase is highly mutagenic. *J Biol Chem* **1996,** *271* (26), 15386-92.
2. Bergoglio, V.; Canitrot, Y.; Hogarth, L.; Minto, L.; Howell, S. B.; Cazaux, C.; Hoffmann, J. S., Enhanced expression and activity of DNA polymerase beta in human ovarian tumor cells: impact on sensitivity towards antitumor agents. *Oncogene* **2001,** *20* (43), 6181-7.
3. Starcevic, D.; Dalal, S.; Sweasy, J. B., Is there a link between DNA polymerase beta and cancer? *Cell Cycle* **2004,** *3* (8), 998-1001.
4. Chan, K.; Houlbrook, S.; Zhang, Q. M.; Harrison, M.; Hickson, I. D.; Dianov, G. L., Overexpression of DNA polymerase beta results in an increased rate of frameshift mutations during base excision repair. *Mutagenesis* **2007,** *22* (3), 183-8.
5. Pelletier, H.; Sawaya, M. R.; Kumar, A.; Wilson, S. H.; Kraut, J., Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* **1994,** *264* (5167), 1891-903.
6. Hazan, C.; Boudsocq, F.; Gervais, V.; Saurel, O.; Ciais, M.; Cazaux, C.; Czaplicki, J.; Milon, A., Structural insights on the pamoic acid and the 8 kDa domain of DNA polymerase beta complex: towards the design of higher-affinity inhibitors. *BMC Struct Biol* **2008,** *8*, 22.
7. Hu, H. Y.; Horton, J. K.; Gryk, M. R.; Prasad, R.; Naron, J. M.; Sun, D. A.; Hecht, S. M.; Wilson, S. H.; Mullen, G. P., Identification of small molecule synthetic inhibitors of DNA polymerase beta by NMR chemical shift mapping. *J Biol Chem* **2004,** *279* (38), 39736-44.
8. Maciejewski, M. W.; Liu, D.; Prasad, R.; Wilson, S. H.; Mullen, G. P., Backbone dynamics and refined solution structure of the N-terminal domain of DNA polymerase beta. Correlation with DNA binding and dRP lyase activity. *J Mol Biol* **2000,** *296* (1), 229-53.
9. Garcia, A. E., Large-amplitude nonlinear motions in proteins. *Phys Rev Lett* **1992,** *68* (17), 2696-2699.
10. Amadei, A.; Linssen, A. B.; Berendsen, H. J., Essential dynamics of proteins. *Proteins* **1993,** *17* (4), 412-25.
11. Tondi, D.; Slomczynska, U.; Costi, M. P.; Watterson, D. M.; Ghelli, S.; Shoichet, B. K., Structure-based discovery and in-parallel optimization of novel competitive inhibitors of thymidylate synthase. *Chem Biol* **1999,** *6* (5), 319-31.
12. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002,** *47* (4), 409-43.
13. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd, Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* **2000,** *33* (12), 889-97.

Chapter 7: Dual-Inhibitors For The P53-MDM2/MDM4 Interactions: Solving Two Puzzles At Once[(1)]

7.1 Introduction

For the last two decades, the tumor suppressor protein p53 has been called the "guardian of the genome". P53 earned this character due to its vital roles in cell cycle, apoptosis, DNA repair and senescence[1,2,3,4] In these processes, p53 responds to cellular stresses, such as hypoxia and DNA damage, by accumulating in the nucleus and activating various pathways to maintain the cell's functional normality.[5] As such, tumor cells have developed numerous ways to disable its function. Inactivation of the p53 pathway occurs in most human cancers. Certainly, The gene *TP53*, encoding for p53, is mutated or deleted in ~50% of human cancers.[6] In the rest of human cancers, although p53 retains its wild type structure, its activity is eradicated by its main cellular inhibitors, murine double minute 2/4 (MDM2/4).[7,8,9] The potential to restore the p53 activity, and hence annihilate cancer, made the p53 protein a viable therapeutic target.[10]

7.1.1 A brief history of p53 research

Almost two decades of careful research has been focused on the nature and function of p53. The story began in 1979, when Lane and Crawford reported a 53 kDa polypeptide forming a complex with a viral protein.[11] The newly discovered peptide named protein 53 (p53) became the subject of many subsequent studies.[12] The complex studied by Land and Crawford was initially noticed only in cancer transformed cells, indicating that p53 has a certain rule in tumor development.[13] Their findings along with others' placed p53 as a possible oncogene.[14] This observation was supported by further studies that showed over-expression of p53 in many tumor cells. In the 1980s new complexes were identified involving p53 and many viral proteins. By that time, it became clear that forming these complexes is a common strategy for tumor transformation, especially, when the same procedure was followed with the tumor suppressor protein, retinoblastoma (Rb).[14] In the mid 1990s the missing parts of the puzzle were completed. Levine's

[(1)] A version of this chapter has been published in Barakat K, Mane J, Friesen D, Tuszynski J. J Mol Graph Model. 2010 Feb 26;28(6):555-68.

group made the first step toward the currently accepted paradigm of p53.[15] Their experiments on p53 cDNA clones suggested that the wild type protein suppresses tumors, while mutants are transforming normal cells into cancerous ones. The second large step was the discovery of many frequent point mutations of p53 in various tumors that lack the other p53 allele.[16] These findings led to one conceivable conclusion that p53 is, in fact, a tumor suppressor protein.

In the years that followed, the whole picture for the different functions of p53 was constructed piece by piece. P53 is no longer considered a solitary protein. This protein is at the heart of the cellular network that protects cells from cancer. It plays a pivotal role in controlling cell cycle progression and inducing cell death by apoptosis. It is a transcription factor that is normally activated by stress signals that may harm the cell.[17] Depending on the type of damage, it binds specifically to precise locations in the genome and activates several genes. The transcribed proteins may stop the cell cycle, giving the cell time to fix the damage, or lead to apoptosis when the damage is severe. As a result of these critical roles and capabilities, the p53 protein is under continued supervision by its two cellular inhibitors MDM2 and MDM4, and many other proteins, most of which have not been identified yet.[10] Therefore it is not surprising that p53 is inactive in most human cancers. The pathway is mostly disabled by genetic mutations or deletions;[6] defective components in post-translational modification; and finally, over-expression of the two proteins, MDM2[18] and MDM4.[9]

7.1.2 MDM2 and MDM4 Regulate P53

Originally, MDM2 was discovered as the main regulator for p53. Looking at the critical functions of p53, this regulation is important to allow normal cell proliferation and/or maintenance of cell viability. In this context, MDM2 and p53 regulate each other through a feedback loop (see Figure 7-1).[7] In this mechanism, p53 transcribes for MDM2, while MDM2 acts as an E3 ubiquitin ligase that exports p53 out of the nucleus and promotes its degradation. Moreover, by binding to the transactivation domain of p53 within the nucleus, MDM2 inhibits p53 function as a transcription factor for other proteins. Thus, when p53 is activated, the transcription of MDM2 is also induced, resulting in higher MDM2 protein levels and more control on p53 functions. Consequently, MDM2 is envisaged as an effectual inhibitor for p53. Over-expression of MDM2 reduces the cellular ability to activate the p53 pathway under stress conditions. This abnormality of p53 regulation was initially discovered in sarcomas retaining wild-type p53, and it was later observed in several cancers as a common mechanism to disable p53 activity.[19,20,18] In fact, MDM2 over expression has been reported in ~10% of 8000 human cancers from various sites, including lung or stomach (for a review, see [21]).

Structurally related to MDM2, MDM4 (also known as MDMX or HDMX) is a second cellular regulator of p53.[9] Although MDM4 lacks the intrinsic E3 ligase activity of MDM2,[22] current models suggest that it acts as a major p53 transcriptional antagonist independent of MDM2.[23] The two proteins form a heterodimeric complex through their C-terminal RING domain interaction which, in turn, increases the ability of MDM2 to promote p53 degradation.[24] In fact, the MDM4-MDM2 interaction can also lead to ubiquitination and degradation of MDM4 leading to the elimination of MDM4 during DNA damage response.[25] MDM4 is also over expressed in many types of cancer.[21] In some cases, MDM4 was identified as a specific chemotherapeutic target for many tumors, for example for the treatment of retinoblastoma.[26] The binding domains of p53 within MDM2 and MDM4 are very similar,[27]offering promise for the discovery of new small molecule compounds that can target the two proteins simultaneously. Surprisingly, although MDM2 and MDM4 share similar structures (see below) and functions, MDM4 is not a target gene of p53 transcriptional activities.

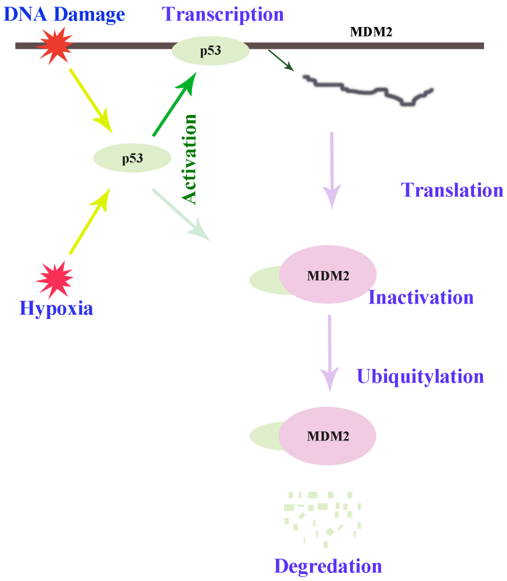


Figure 7-1: p53-MDM2 feedback loop. Adopted from Hardcastle.28

By noting that p53 binds to the same region within MDM2/4, and only MDM2 can promote for p53 degradation, one can reach a possible conclusion regarding their distinct functions. That is, MDM4 regulates p53 activity and MDM2 regulates p53 stability. Recently, Toledo et al. supported this idea by showing that the loss of MDM4 increased the levels of MDM2, which, in turn reduced the concentration of p53.[21] On the other hand, the loss of MDM2 increased the levels of p53, but did not have a significant effect on p53 activity. Francoz also speculated that MDM2 and MDM4 might have these separate but cooperative duties in regulating p53.[29]

7.1.3 Structural similarities between MDM2 and MDM4

MDM2 and MDM4 have similar structures. The two proteins are 492 and 490 polypeptides, respectively. As shown in Figure 7-2, their structures possess three conserved domains: an N-terminal domain that interacts with p53, a Zinc finger domain and finally, a C-terminal ring domain. The unconserved parts within MDM2 and MDM4 are comprised of regions of acidic residues. Earlier genetic and biochemical studies on the p53-MDM2 complex limited their interaction to the 106-N terminal domain of MDM2 and the N-terminus of the transactivation domain of p53.[8]

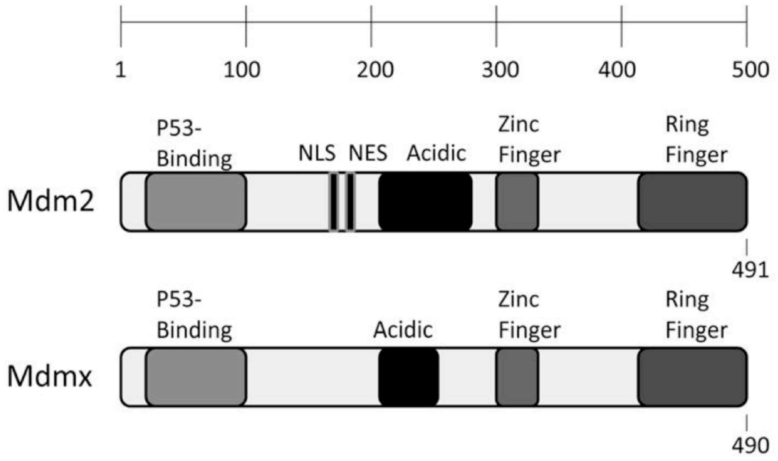


Figure 7-2: Schematic representation of the domains of Mdm2, Mdm4. Adopted from Waning et al.[30]

The high-resolution crystal structure of the complex (see Figure 7-3) demonstrated the essential interacting regions located in the MDM2-p53 interface.[8] Essentially, p53 forms an amphipathic-helix peptide (residues 15-29) that is partly buried inside a small but deep, hydrophobic groove on the surface of the MDM2 N-terminal domain (residues 19-102). This interaction involves four key residues from p53, namely F19, L22, W23 and L26 and at least 13 residues from MDM2 (L54, L57, I61, M62, Y67, Q72, V75, F86, F91, V93, I99, Y100 and I103).[31] Interestingly, 10 out of the 13 most important MDM2 residues described above are conserved in MDM4, which indicates that the binding site of p53 within the surface of MDM4 is similar to, but not identical with, that of MDM2.

7.1.4 Current Inhibitors for the MDM2/4-P53 Interaction

As MDM2 was discovered first, most of the preceding efforts have been exclusively focused on uncovering small molecule MDM2-inhibitors and little work has been done on targeting MDMX. The main concept was that a small molecule that mimics the p53-hot spot residues would ultimately disrupt this interaction and is assumed to completely re-activate the p53 pathway and restore the cell's functional normality. No one was thinking about MDM4, especially, when several studies verified the original concept of targeting the p53-MDM2 interaction. These studies showed that it is possible to activate p53 in cancer cells that retain the wild type structure and promised for an ultimate therapeutic strategy for many cancers.[32,33,34]

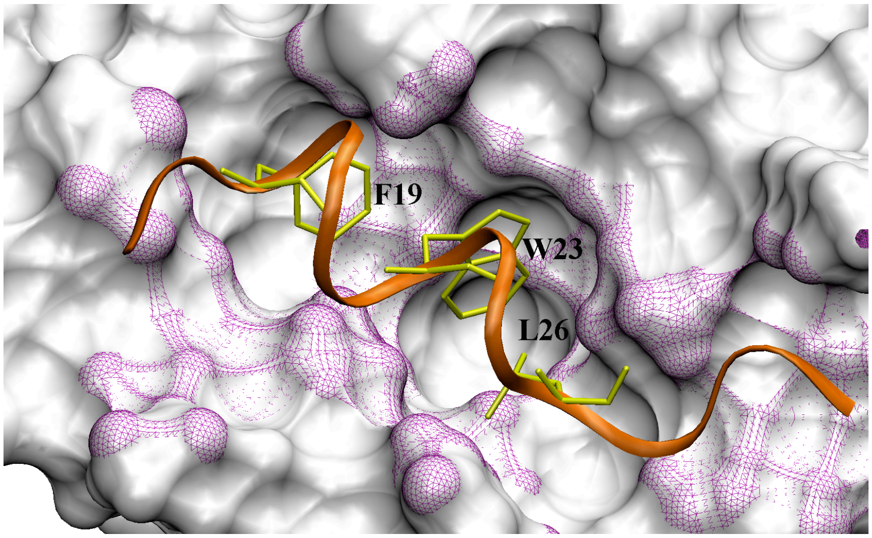


Figure 7-3: P53-MDM2 interaction.

The p53-binding site within MDM2 is shown in molecular surface representation with the residues constituting the binding site are highlighted in purple. P53 (orange) is shown in ribbon representation. The interaction is mainly hydrophobic. P53 residues F19, W23 and L26 point together toward a cleft at the surface of the MDM2 protein. The three p53 residues are surrounded by hydrophobic MDM2 residues L54, L57, I61, M62, Y67, V75, F86, F91, V93, I99, Y100 and I103 (shown in purple).

In particular, the last decade has witnessed the identification of an increasing number of non-peptide, small-molecule MDM2 inhibitors with promising binding affinities.[35] These are analogs of *cis*-imidazoline (Nutlins),[36] spiro-oxindole (MI-63 and MI219)[,37,38,39,37] benzodiazepinedione (TDP665759)[,40,41,42] terphenyl,[43] quilinol,[44] chalcone,[45] and sulfonamide.[46] Of these molecules only three compounds, namely, Nutlin-3, MI-219 and TDP665759 showed sufficiently high binding affinity, and desirable pharmacokinetic profiles in cells.[32] However, these compounds are more highly selective for MDM2 than for its homolog MDM4. In particular, MI-219 showed a greater than10,000-fold selectivity for MDM2 relative to MDM4. These findings were very unexpected and disappointing for many research groups. Because after looking into the similarities between MDM2 and MDM4, it was assumed that inhibitors that target the MDM2-p53 interaction should function in the same way to disrupt the MDM4-p53 binding. Surprisingly, this was not the case, as most of MDM2-inhibitors, including Nutlin-3 have been shown to be inactive in cancer cells overexpressing MDM4,[47] opening a new avenue in p53 research and requiring a new generation of MDM2-inhibitors that can target its homolog, MDM4, as well.

Based on the abovementioned discussion, it is clear that the development of novel compounds that are MDM4-specific or optimized for dual-inhibition of MDM2 and MDM4 is a necessary step to achieve full activation of p53 in tumor cells. Recently, Pazgier et al. reported the development of a potent peptide inhibitor, termed PMI (p53-MDM2/MDM4 inhibitor) that can target the interactions of p53 with both MDM2 and MDMX.[48] This peptide inhibitor provided the proof of concept for this strategy and opened the door for the discovery of novel small molecule inhibitors that can mimic its function. And from here, we start.

7.2 Results And Discussion

Here, we screened the NCI diversity set, the DrugBank set of small-molecules (see section 3.2.1) and more than 3,168 derivative structures extracted from the known MDM2- inhibitors against twenty-eight different MDM2 models that represent the apo- and holo-structure's collective conformational dynamics. The top 300 hits that showed strong affinity for MDM2 have been used in a second round of screening against the p53 binding-site within MDM4 Figure 7-4 depicts the basic strategy that was followed in this work. Results described herein represent the identification of dual-inhibitors that are predicted to disrupt the MDM2/MDMX- p53 interaction and allow for the full activation of the p53 pathway.

7.2.1 MD Simulations of MDM2-p53 and PMI-MDM4

The N-terminal domain of MDM2 was subjected to MD simulations, in both its free and p53-bound states. We also used MD simulations to generate an equilibrated model for the PMI-peptide/MDM4 complex introduced by Pazgier et. al to filter the top hits from the MDM2 screening for those compounds that can mimic the characteristics of this peptide inhibitor. It should be noted that, as we used docking as a preliminary filtering step in screening the full set of compounds for MDM2-inhibitors, it was essential to generate an ensemble of MDM2 structures in order to partly incorporate protein flexibility during docking. In this context, we selected the top 300 compounds that can bind to their appropriate MDM2 conformation for post-docking analysis using MD simulations to introduce the full flexibility for both the ligand and its selected target structure. On the other hand, we did not generate such ensemble of structures for the MDM4 screening exercise because of the following reasons; First, docking runs were not used to filter the compounds for best binders, they were used to place each ligand from the 300 MDM2-top hits within the MDM4 pocket and assemble a minimum energy protein-ligand conformation required for MD simulations. Second, the full flexibility of this complex will be established using a fairly long MD simulations (2ns) that can reasonably explore the conformational space of the protein-ligand complex and simulate their induced fit interaction.

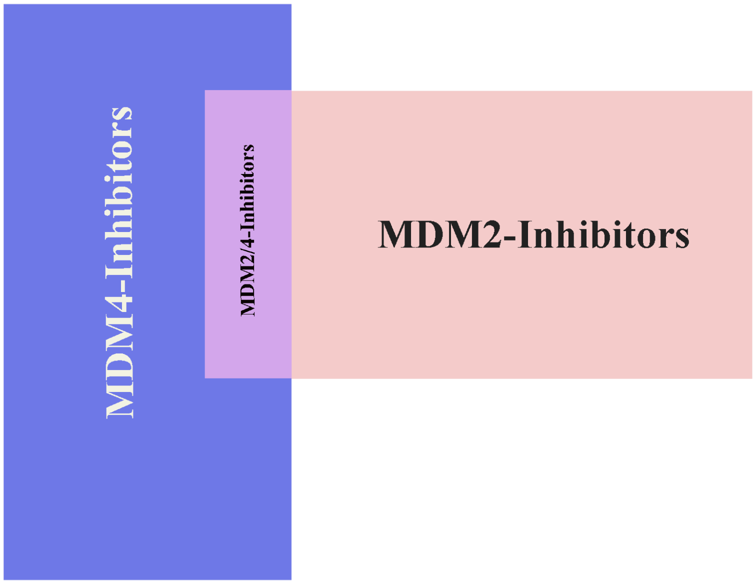


Figure 7-4: Searching strategy for dual MDM2/MDM4 inhibitors. In this work, we first screened the compound databases for MDM2-inhibitors, The results from this search were screened against MDM4. Inhibitors that can bind to the two proteins are represented by the intersection of the two groups.

7.2.2 PCA and Convergence

As described in section 3.3.4, PCA was performed over the entire MD simulations of both the holo- and apo-MDM2 structures using atoms comprising the 18 residues contained in the

MDM2 binding site (residues numbered: 25, 26, 50, 51, 54, 58, 61, 62, 67, 72, 73, 93, 94, 96, 97, 99, 100, 104) with the backbone atoms RMSD fitted to the minimized crystal structure of the two starting configurations. We defined the binding site as comprised of the MDM2-residues that are located within 10 Å from p53-atoms. Resulting eigenvectors were sorted by descending eigenvalues, which represent the variance of the motion along the principal components. The ten dominant eigenvalues for the two simulated systems are shown in Figure 7-5-a. The first eigenvalue has a magnitude that is significantly higher than those of the other eigenvalues. The components with the largest eigenvalues represent correlated motions of the binding site with the most significant standard deviations of the motion along the corresponding orthogonal directions.

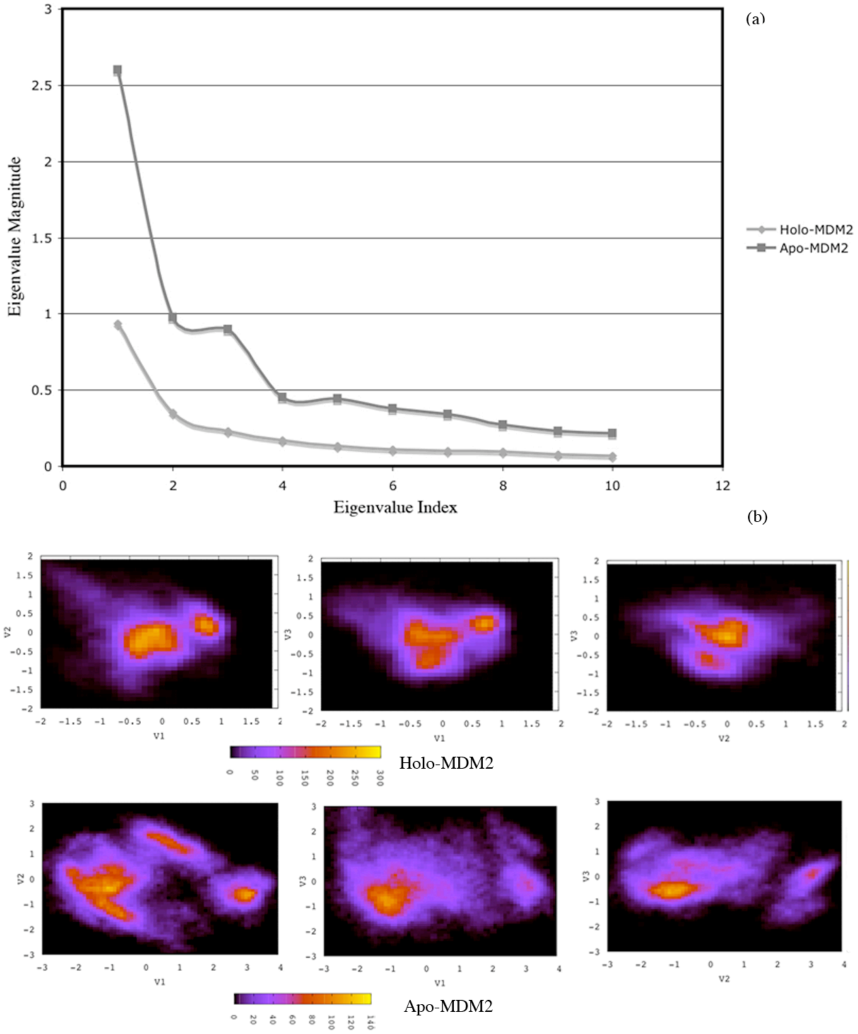


Figure 7-5: PCA for the MDM2 binding site.

a) The dominant ten eigenvalues for the apo and holo trajectories. b) Projections of the ensemble of conformations onto the planes of the three most important principal components.

Figure 7-5-b represents the spatial distributions of occupancies for the conformational states over the planes spanned by the three dominant principal components of the binding site for the two systems. The p53-binding site within MDM2 adopted several conformations throughout the MD simulations, indicating the flexibility of the protein. The grouping of MD trajectories into a limited number of clusters suggests the presence of favored folded conformations and significant basins of attraction.

Covariant analysis of the trajectories from the holo- and apo-MD simulations, successively divided into thirds, was performed using the same procedure used for PCA. Normalized overlaps calculated between each of these thirds are reported in Table 7-1. The high overlap between the thirds indicates that each part of the simulation samples approximately the same conformational space, and it is unlikely that there are unexplored regions missed earlier in the runs. Although there is no guarantee that complete equilibrium sampling is given, we have concluded that the observed overlap is acceptable and adequate sampling within the MD trajectories for the binding site had been obtained.

apo-MDM2	holo-MDM2	
0.761	0.754	1st and 2nd
0.732	0.811	1st and 3rd
0.734	0.792	2nd and 3rd

Table 7-1: PCA normalized overlap for the p53-binding site within MDM2. Covariance analysis has been performed for the three thirds of the MD trajectories for the apo (free) and holo (bound) systems followed by calculating the overlap between their covariance matrices.

The two different systems were stable during the MD simulations as indicated by the plots of the RMSDs for the backbone atoms from the initial co-ordinates for the last 10 ns (see Figure 7-6). For the apo-MDM2 simulations, the protein backbone RMSD fluctuated about a mean of 1.9 Å. On the other hand, the holo-system fluctuated around lower RMSD values for both MDM2 and the p53-peptide of 1.3 Å and of 1.5 Å, respectively, indicating a mutually stabilizing effect induced by protein-protein interactions. Atomic fluctuations predicted the key residues that are important for binding (see Figure 7-7). The main-chain B-factors (averaged over heavy atoms) for the residues constructing the p53-binding site within MDM2 (residues 68-73, 85-102 and 104-105) are higher than the corresponding holo-MDM2 values. This suggests the relative flexibility of the model in this region where the 18 residues defining the binding site seem to be relatively rigid during the MD simulation in the p53-MDM2 models. On the p53 side, residues 19-26 (see Figure 7-7-b) are more rigid than other p53 residues, suggesting their critical participation in binding to MDM2.

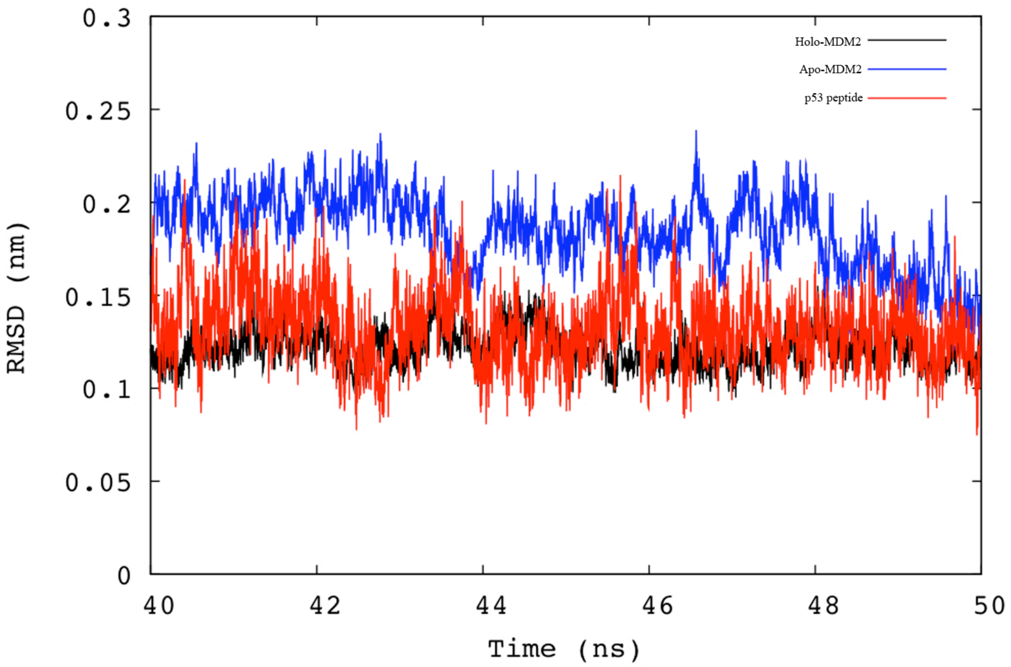


Figure 7-6: RMSD analysis.

Plot of the RMSD of the backbone atoms from the reference structure as a function of simulation time in p53-peptide, MDM2-free and MDM-p53 complex.

7.2.3 Ensemble-based Virtual Screening

As discussed in chapter 2 (section 2.2.1.2), protein flexibility is crucial and must be introduced during the docking process. In this context, we have used an ensemble of protein conformations for docking as a practicable alternative to introduce a feature of global protein flexibility. To generate a reduced set of representative models of the MDM2 binding site, we applied the RMSD conformational clustering to the apo-MDM2 and holo-MDM2 binding site trajectories. Figure 7-8 shows the evaluation of the Davies-Bouldin index (DBI) and the percentage of variance explained by the data (SSR/SST) for different cluster counts (see methodology). DBI for the apo-system exhibited local minima at cluster counts of 10, 20 and 60. However, as the percentage of variance explained by the data started to plateau after 45 clusters for the apo-system, we concluded that 60 clusters is a reasonable cut-off for the free-MDM2 structures. On the other hand, the correlation between these two criteria nicely occurred at a cluster count of 30 for the holo-structure.

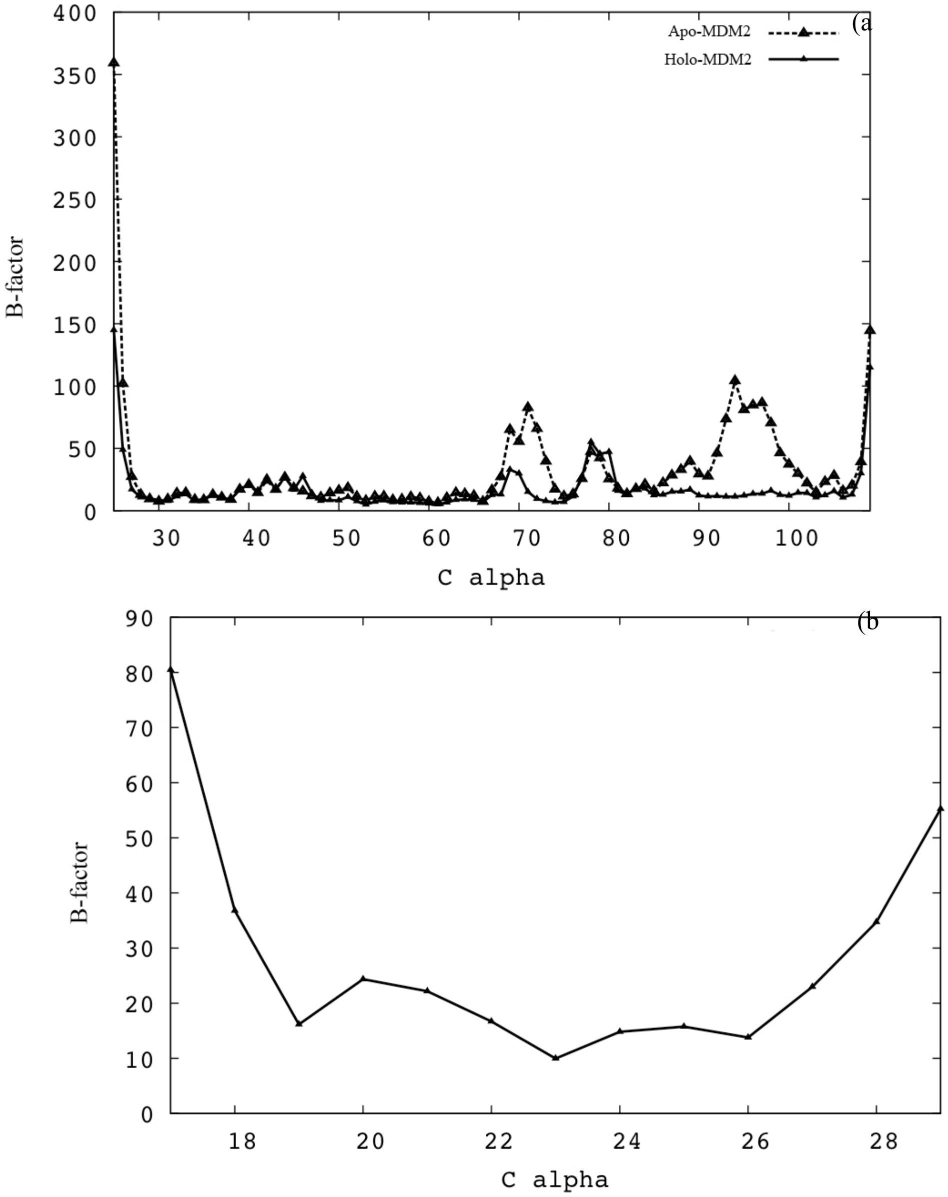


Figure 7-7: Atomic fluctuations.

Plot of the B-factors averaged over the protein backbone atoms as a function of residue number in the simulations of (a) MDM2-free and MDM2-bound and (b) p53 peptide. The solid and dotted lines correspond to MDM2-bound and MDM2-free, respectively.

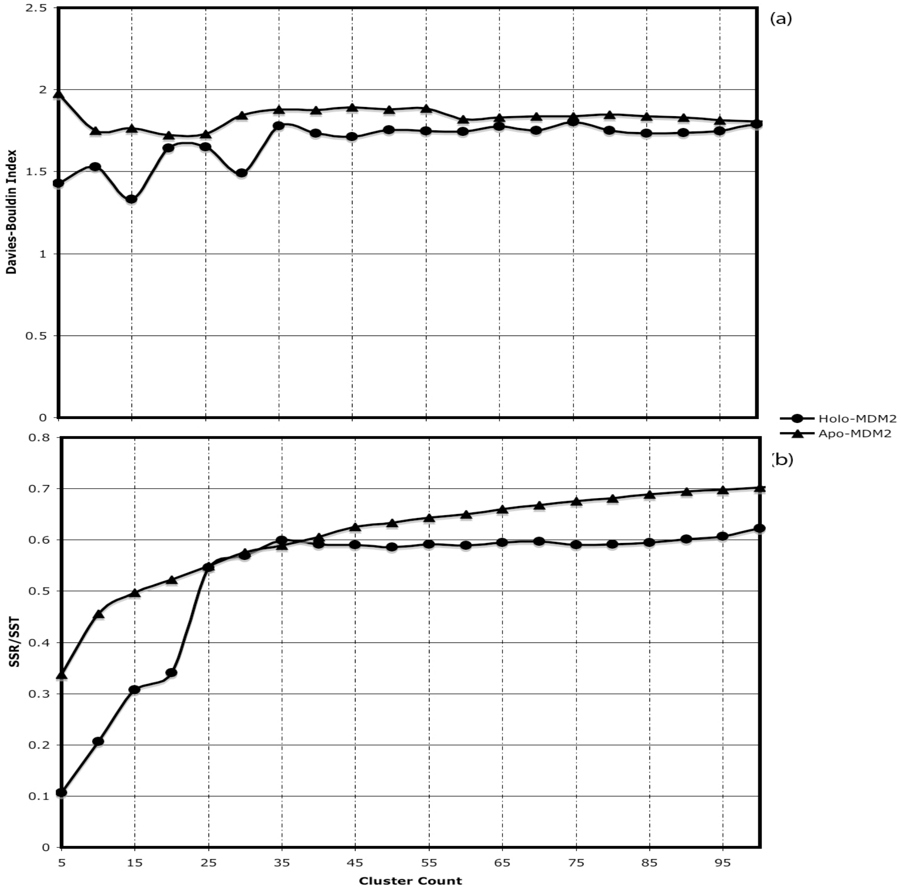


Figure 7-8: Clustering analysis for the two MDM2 trajectories.

A high-quality clustering is obtained when a local minimum in DBI correlates with saturation in the SSR/SST ratio. This is clear at cluster count of 60 for the apo-structure and 30 clusters for the holo-structure.

In this study, we constructed an ensemble of twenty-eight distinct conformations to perform ensemble-based virtual screening on MDM2 against the full set of ligand compounds. This ensemble incorporated the most dominant twenty-two structures that comprised ~75% of apo-trajectory, the most dominant five holo-structures that represented ~80% of the bound conformations (data not shown) and finally the MDM2 conformation extracted from the p53-bound crystal structure. The ultimate goal was to reduce the number of representative structures included in the ensemble-based screening and concurrently comprise most of the conformational space of the binding site.

7.2.4 Pose Clustering

As mentioned above, twenty-eight independent virtual screening experiments were performed against the full set of database compounds. Screening of the full set of compounds contained in the NCIDS, DrugBank and the inhibitor-derivatives databases (more than 6,000 molecules), against the twenty-eight target structures, produced a total of ~ 19 million distinct poses that required classification. Using the iterative clustering technique described in chapter 3 (see section 3.4.2 for details), all docking results were automatically clustered to properly extract their optimal clusters.

7.2.5 Preliminary Ranking

Following the procedure described in section 3.4.3, we ranked significant poses for each of the 6,617 molecules contained in the database by using the results from the elbow criterion and the lowest energy that corresponds to the most populated cluster. Once all poses from each ligand entry were clustered, we then filtered all of the clusters so that only those containing at least 25% of the total population were considered as top hits. Top hits were collected from the 28 experiments by first extracting the largest cluster from each individual screening followed by ranking the clusters according to their binding energies. This produced a set of non-redundant hits ranked by their binding energies of the most populated cluster. Top 300 hits were rescored using the MMPBSA method (see below) and were used in the subsequent MDM4-screening.

7.2.6 MM-PBSA Ranking

The apparent IC_{50} values for Nutlin3, MI-219, TDP665759 and PMI in binding to MDM2 are 90 nM[18], 5 nM[18], 704 nM[18,29] and 3.4 nM35 at 25^0C, respectively. We did not find explicit values for the binding affinities of the three non-peptide molecules regarding their binding to MDMX, hoverer, it has been experimentally confirmed that these compounds are weak binders to MDMX[18,29,35]. The IC_{50}values can be converted to the observed free energy change of binding using the relation:

$$\Delta G = RT \ln K_i \qquad \text{Equ. 7-1}$$

where R is the gas constant, R =1.987 and T is the absolute temperature. Table 2 lists the estimated binding energies for the three compounds compared to the experimentally expected values. Due to the vast number of torsional degrees of freedom in the peptide structure, we did not use AUTODOCK scoring function to calculate its binding energies to the two proteins. However, we used the MM-PBSA method (see section 3.7) to rescore the top 300 hits from the MDM2 screening along with the PMI peptide and compare the predicted binding energies to AUTODOCK scoring and the experimental values.

Compound	MDM2 Ranking (kcal/mol)			MDMX Ranking (kcal/mol)		
	MM/PBSA	AUTODOCK	Experimental	MM/PBSA	AUTODOCK	Experimental
MI-219	-10.6 ± 1.5	-9.1 ± 2.2	-11.4[32]	-5.3 ± 1.5	-6.8 ± 2.2	-5.9[32]
Nutlin-3	-9.3 ± 1.3	-8.2 ± 2.2	-9.7[32,48]	-6.1 ± 1.6	-5.8 ± 2.2	Negligible[48]
TDP665759	-9.5 ± 1.5	-9.1 ± 2.2	-8.4[32,42]	-5.6 ± 1.4	-8.2 ± 2.2	Negligible[42]
PMI	-10.4 ± 1.6	N/A	-11.6[48]	-12.8 ± 1.5	N/A	-11.5[48]

Table 7-2: Relative ranking of positive controls using the two scoring methods compared to experimental data.

Although the discrepancy in the MM-PBSA calculations for the interactions of the four inhibitors with MDM2 is about 1 kcal/mol, the predicted values are in an excellent agreement with the experimental data compared to the values obtained by AutoDock scoring function (see Table 7-2). This observation is evident in the calculated values for their interactions with the MDM4 target, predicting their weak binding to the protein. These results also illustrate the limitations of AutoDock scoring function in eliminating false positive ligands, i.e. compounds that cannot practically bind but predicted to bind, from active compounds. This is shown in Table 7-2, where the TDP665759 compound is predicted to bind to MDM4 with a relatively high binding energy compared to the rest of the compounds. On the other hand, the MM-PBSA approach selected the real binders for the two protein targets. For MDM2, the four ligands can bind strongly to the protein, while, for MDM4, only the PMI peptide can bind with a very high binding energy. This also explains the variations between the two scoring methods in ranking the compounds (see tables Table 7-3 to Table 7-5).

MM-PBSA scoring		Ensemble-based scoring		ID
Rank	BE (kcal/mol)	Rank	BE (kcal/mol)	
1	-14.1	97	-9.2	ZINC08552001
2	-14.1	101	-9.2	NSC#82892
3	-14.0	42	-9.7	Pub#11952783
4	-13.8	187	-9.0	Pub#11375913
5	-13.2	285	-8.5	Pub#10312264
6	-13.0	175	-8.9	Pub#25055003
7	-12.8	191	-8.8	NSC#59276
8	-12.6	293	-8.5	Pub#456323
9	-11.7	110	-9.1	Pub#11855975
10	-11.4	80	-9.3	Pub#11952782
11	-11.3	115	-9.1	Pub#22721132
12	-11.3	150	-9.0	NSC#409664
13	-11.3	81	-9.3	Pub#21060012
14	-11.0	232	-8.7	Pub#20726116
15	-10.9	127	-9.0	Pub#22632481
16	-10.9	267	-8.6	Pub#11272250
17	-10.8	62	-9.5	Pub#22720968
18	-10.8	180	-8.8	NSC#77037
19	-10.6	109	-9.1	MI-219
20	-10.6	93	-9.2	Pub#22721012

Table 7-3: MDM2 top hits. The top 20 hits from MDM2 screening ranked by their binding energies as were calculated using the MMPBSA method and compared to their ranking using the AUTODOCK scoring function.

MM-PBSA scoring		Ensemble-based scoring		ID
Rank	BE (kcal/mol)	Rank	BE (kcal/mol)	
1	-13.2	277	-5.5	ZINC12503171
2	-13.1	74	-7.8	NSC#72254
3	-12.9	178	-7.1	Pub#20726118
4	-12.8	N/A	N/A	PMI
6	-11.6	111	-7.5	Pub#11455269
7	-11.5	13	-8.5	Pub#22721034
8	-11.3	69	-7.8	Pub#10196974
9	-11.0	100	-7.6	Pub#22720998
11	-10.7	43	-8.1	Pub#10290053
13	-10.4	130	-7.4	Pub#22721115
16	-9.9	30	-8.2	Pub#22721175
17	-9.9	50	-8.1	Pub#11541499
18	-9.8	226	-6.7	Pub#11614489
19	-9.7	41	-8.2	Pub#22721095
21	-9.5	19	-8.4	Pub#22721184
22	-9.53	25	-8.3	Pub#216345
23	-9.41	191	-7.0	Pub#24788704
24	-9.35	121	-7.5	Pub#20726117
25	-9.29	57	-7.9	ZINC08552003
29	-9.19	184	-7.1	Pub#17754804

Table 7-4: MDM4-specific top hits. Compounds are ranked by their binding energies as were calculated using the MMPBSA method and compared to their ranking using the AUTODOCK scoring function. The listed compounds have been predicted to bind to MDMX and not to MDM2

As the MM-PBSA confirmed the experimental findings, our subsequent step was to use this technique to re-score the top hits obtained by AutoDock scoring function. Table 7-3 shows the top 20 hits obtained from the ensemble-based screening after rescoring their interactions using the MM-PBSA method. Although Nutlin-3 and TDP665759 are not shown in this table, the ranks of the two compounds were 41 and 36, respectively. It should be also mentioned that a considerable number of the compounds showed positive binding energies after rescoring them using the MM-PBSA method (data not shown), supporting the ability of this technique to discriminate inactive compounds from the AutoDock suggested hits. Although most of the top 20 hits are derivatives of the three positive controls, in particular the benzodiazepinedione scaffold (TDP665759), 5 compounds from both the NCI diversity set and DrugBank libraries showed strong binding energies compared to the positive controls.

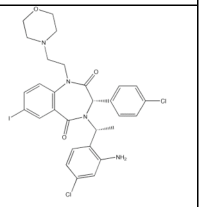
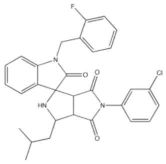
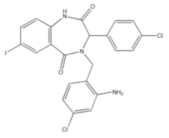
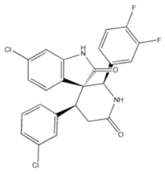
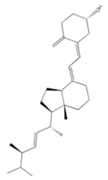
7.2.7 Identifying MDM2/4 dual inhibitors

The top 300 hits that resulted from the MDM2-ensemble-based screening were then docked to an equilibrated MDM4 structure that was extracted from the PMI/MDMX complex (see methodology). The docking step was essential in order to place the compounds within the p53-binding site with their minimum energy conformation. Following the procedure described above, we used the MM-PBSA method to predict the absolute binding energies for each compound. Reassuringly, our calculations confirmed the experimental findings, where the three non-peptidic positive controls showed weak binding to MDM4 compared to MDM2 while the PMI peptide showed very high binding energy (see Table 7-2). Analogously to the experimental results concerning the high specificity of these molecules to the MDM2 target, our calculations predict that a number of compounds can also bind more strongly to MDM4 than to MDM2. The top 20 hits selected from these compounds are expected to be specific MDM4 inhibitors and are shown in Table 7-4.

Table 7-5 lists the compounds that are suggested to function as dual-MDM2/MDM4 inhibitors obtained from screening the top MDM2 hits against the p53-binding site within the MDM4 target. Here, we used MM-PBSA energies to compare the binding of these hits to the two target proteins. As we are only interested in compounds that can bind to MDM2 with affinities as good as those of the known MDM2-inhibitors, we limited our selection to the 16 compounds shown below (see Table 7-5).

Although the binding sites are fairly similar, the MDM4 pocket seems to be more compact than that of MDM2. This is mainly due to the three residues Pro95, Ser96 and Pro97 in MDM4 that have been replaced by His96, Arg97 and Lys98 in MDM2. These substitutions are located on one of the alpha helices that comprise the p53 binding site within the two proteins. Consequently, the proline residues (Pro95 and Pro97) in MDM4 shift this helical domain in

MDM4 relative to MDM2 and cause Lys98 and Tyr99 to protrude into the p53-binding cleft within MDM4, making it shallower and less accessible to many of the MDM2 top hits we found. Moreover, we noticed very minor differences in the electrostatic potential distributions around the surfaces of the two proteins (data not shown), where MDM2 is more positively charged in certain regions deeply located within the binding site.

MDMX rank		MDM2 rank		ID	Structure
Rank	BE (kcal/mol)	Rank	BE (kcal/mol)		
5	-12.5	11	-11.3	Pub#11952782	
10	-10.9	45	-8.8	Pub#5039349	
12	-10.5	67	-7.5	Pub#11284279	
14	-10.2	69	-7.4	Pub#24788253	
15	-10.0	62	-7.7	ZINC04629876	

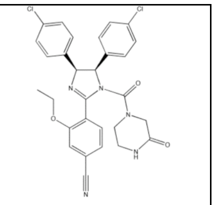
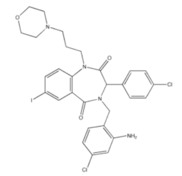
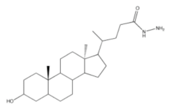
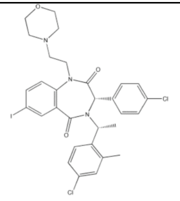
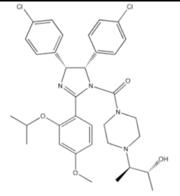
20	-9.6	10	-11.7	Pub#11855975	
26	-9.3	50	-8.3	Pub#11953191	
28	-9.2	45	-8.7	NSC#73109	
31	-9.0	39	-9.0	Pub#11952569	
32	-8.7	65	-7.5	Pub#10240227	

Table 7-5: MDMX/MDM2 Inhibitors. The listed compounds are predicted to bind to the MDM2 and MDMX proteins. The compounds are ranked by their binding energies as were calculated using the MMPBSA method.

These slight variations in both shape and electrical properties of the two proteins played a considerable role in governing the final conformation adopted by the ligands. This observation is clear when comparing the binding modes of nutlin within the two pockets (see Figure 7-9-a). While Tyr100 and Leu99 of MDM2 extend the binding site allowing nutlin to intimately bind to MDM2, the same residues in MDMX clash with the drug preventing it from taking the normal conformation that was adopted within MDM2. On the other hand, Figures 6b-c show how two compounds from the list of proposed MDM2/MDMX inhibitors were able to tolerate the structural variations between the two binding sites (see Figure 7-9).

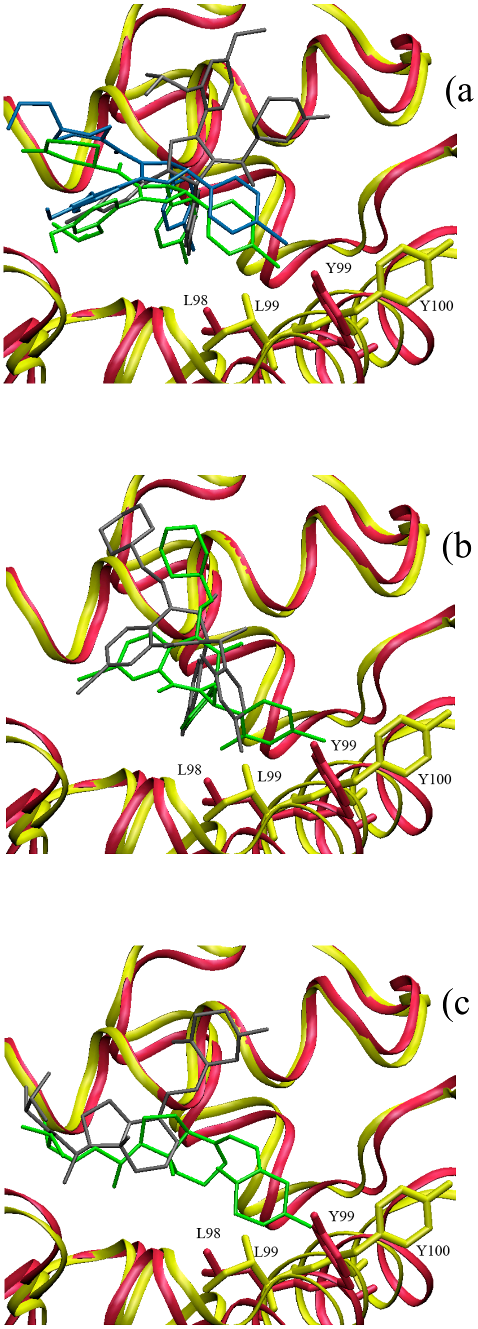


Figure 7-9: Structural variations between MDM2 (yellow) and MDMX (red) and their effect on the binding modes of nutlin-3 (a) and two selected hits form the predicted MDM2/MDMx inhibitors (b and c). Tyr100 and Leu99 of MDM2 and the same residues in MDMX are shown in Licorice representations with the same color as that of the two proteins. For each compound, the binding mode within MDM2 is shown in green and within MDMX is shown in gray. Tyr99 and Leu98 prevent nutlin-3 from binding to MDMX with the same binding conformation adopted by nutlin-2 within the MDM2-pocket (blue). The conformation of nutlin-2 was extracted from the MDM2-nutlin crystal structure 1RV1. On the other hand, compounds Pub#11952782 (b) and ZINC04629876 (c) from the suggested MDM2/MDMX inhibitor list can tolerate the structural variations in the two binding sites in order to maximize their interactions with the proteins.

7.3 Conclusion

The tumor suppressor p53 is one of the most frequently inactivated proteins in human cancers. Direct gene modifications in p53 gene, Tp53, or the interaction between p53 and its two major cellular inhibitors, MDM2 and MDM4, are two fundamental mechanisms employed by cancer cells to block the p53 pathway.[1,2,3,4] Over a number of years, leading efforts in p53 research have been focused on restoring the activity of the mutant protein as a precursor to developing a novel cancer treatment. Although these studies revealed the prospects of inducing tumor cell death, the development of a non-peptide small-molecule p53-activator is still a particularly challenging problem.[49,50] Other significant efforts in this area have been aimed at the discovery of small-molecule inhibitors that can disrupt the interaction of p53 with its main cellular regulator, MDM218-21. This led to the development of Nutlin3[36] and MI-219,[51] the most potent and specific non-peptide MDM2-inhibitors discovered so far.

Recently, MDM4, a protein homologous to MDM2, was found to reduce the efficacy of MDM2-inhbitors including Nutlin3.[32,47] This suggested MDM4 as a new attractive therapeutic target and indicated a need to develop MDM4-specific or MDM2/MDM4-dual inhibitors to fully activate the p53 pathway in tumor cells expressing wild type p53.

Here, we used an improved relaxed complex scheme by combining MD simulations and molecular docking with binding energy analysis to filter a set of 6,617 compounds for effective inhibitors of MDM2 and MDM4. These compounds included the NCI diversity set, DrugBank small molecules and a newly generated set of ~ 3000 derivative structures similar to known MDM2-inhibitors. The derivative library of compounds was included among the docked structures because the structural similarity between the two proteins would imply that an MDM4-inhibitor should be a derivative structure based on one of the known MDM2-inhibitors. Although, more than 6,000 compounds have been screened in this study, we suggest the use of larger libraries (in the order of 100,000 to 1,000,000 compounds) will be more effectual in discovering more active hits in future work. In this context, we used MD simulations, principle component analysis and an iterative clustering technique to generate an ensemble of 28 MDM2 structures that characterize the collective dynamics of the MDM2 protein. Then, we used molecular docking to explore the conformational space of the ligands and to search for their minimal energy configuration within the MDM2 binding site. All docking poses were clustered using the same iterative procedure that we used in extracting the protein structures and then sorted with the minimal energy of the largest cluster. The top 300 hits were rescored using the MM-PBSA procedure and prepared for a second round of screening against the MDM4 target. Following the docking of MDM2-hits to MDMX4 we used the MM-PBSA technique to rescore their binding affinities to MDM4 and suggest a set of MDM2/MDM4 dual inhibitors.

Our results confirmed the experimental findings concerning the weak binding of the MDM2 inhibitors Nutlin-3, MI219 and TDP to its homolog structure MDM4. Moreover, as we anticipated, the top hits from our screening are primarily derivative structures of the three known inhibitors. However, we also suggested a few structures from the NCI diversity set and DrugBank compounds. The molecules we have proposed in the present study, can fit within the two binding sites and adopt different conformations to maximize their interactions with the two proteins. We have also validated our top hit list by comparing their estimated binding energies to the PMI peptide, an MDM2/MDM4 dual-inhibitor proposed by Pazgier et. al. and, reassuringly, our top hits are predicted to have comparable performance to this peptide. It is hoped that our findings will facilitate the development of a new generation of MDM2/MDM4 dual inhibitors that would fully-activate the p53 pathway and aid in the offer new hope in the fight against a broad range of cancers.

7.4 Bibliography

1. Teodoro, J. G.; Evans, S. K.; Green, M. R., Inhibition of tumor angiogenesis by p53: a new role for the guardian of the genome. *J Mol Med* **2007,** *85* (11), 1175-86.

2. Fridman, J. S.; Lowe, S. W., Control of apoptosis by p53. *Oncogene* **2003,** *22* (56), 9030-40.

3. Vousden, K. H.; Lu, X., Live or let die: the cell's response to p53. *Nat Rev Cancer* **2002,** *2* (8), 594-604.

4. Bourdon, J. C.; Laurenzi, V. D.; Melino, G.; Lane, D., p53: 25 years of research and more questions to answer. *Cell Death Differ* **2003,** *10* (4), 397-9.

5. Vogelstein, B.; Lane, D.; Levine, A. J., Surfing the p53 network. *Nature* **2000,** *408* (6810), 307-10.

6. Feki, A.; Irminger-Finger, I., Mutational spectrum of p53 mutations in primary breast and ovarian tumors. *Crit Rev Oncol Hematol* **2004,** *52* (2), 103-16.

7. Kubbutat, M. H.; Jones, S. N.; Vousden, K. H., Regulation of p53 stability by Mdm2. *Nature* **1997,** *387* (6630), 299-303.

8. Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. P., Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **1996,** *274* (5289), 948-53.

9. Shvarts, A.; Steegenga, W. T.; Riteco, N.; van Laar, T.; Dekker, P.; Bazuine, M.; van Ham, R. C.; van der Houven van Oordt, W.; Hateboer, G.; van der Eb, A. J.; Jochemsen, A. G., MDMX: a novel p53-binding protein with some functional properties of MDM2. *EMBO J* **1996,** *15* (19), 5349-57.

10. Chen, F.; Wang, W.; El-Deiry, W. S., Current strategies to target p53 in cancer. *Biochem Pharmacol* **2010,** *80* (5), 724-30.

11. Lane, D. P.; Crawford, L. V., T antigen is bound to a host protein in SV40-transformed cells. *Nature* **1979,** *278* (5701), 261-3.

12. Oren, M., The p53 cellular tumor antigen: gene structure, expression and protein properties. *Biochim Biophys Acta* **1985,** *823* (1), 67-78.

13. Rotter, V.; Wolf, D., Biological and molecular analysis of p53 cellular-encoded tumor antigen. *Adv Cancer Res* **1985,** *43*, 113-41.

14. Bishop, J. M., Viral oncogenes. *Cell* **1985,** *42* (1), 23-38.

15. Harvey, D. M.; Levine, A. J., p53 alteration is a common event in the spontaneous immortalization of primary BALB/c murine embryo fibroblasts. *Genes Dev* **1991,** *5* (12B), 2375-85.

16. Lowe, S. W., Cancer therapy and p53. *Curr Opin Oncol* **1995,** *7* (6), 547-53.

17. Joerger, A. C.; Fersht, A. R., The tumor suppressor p53: from structures to drug discovery. *Cold Spring Harb Perspect Biol* **2010,** *2* (6), a000919.

18. Momand, J.; Jung, D.; Wilczynski, S.; Niland, J., The MDM2 gene amplification database. *Nucleic Acids Res* **1998,** *26* (15), 3453-9.

19. Fakharzadeh, S. S.; Trusko, S. P.; George, D. L., Tumorigenic potential associated with enhanced expression of a gene that is amplified in a mouse tumor cell line. *EMBO J* **1991,** *10* (6), 1565-9.

20. Oliner, J. D.; Kinzler, K. W.; Meltzer, P. S.; George, D. L.; Vogelstein, B., Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature* **1992,** *358* (6381), 80-3.

21. Toledo, F.; Wahl, G. M., Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer* **2006,** *6* (12), 909-23.

22. Jackson, M. W.; Berberich, S. J., MdmX protects p53 from Mdm2-mediated degradation. *Mol Cell Biol* **2000,** *20* (3), 1001-7.

23. Toledo, F.; Krummel, K. A.; Lee, C. J.; Liu, C. W.; Rodewald, L. W.; Tang, M.; Wahl, G. M., A mouse p53 mutant lacking the proline-rich domain rescues Mdm4 deficiency and provides insight into the Mdm2-Mdm4-p53 regulatory network. *Cancer Cell* **2006,** *9* (4), 273-85.

24. Sharp, D. A.; Kratowicz, S. A.; Sank, M. J.; George, D. L., Stabilization of the MDM2 oncoprotein by interaction with the structurally related MDMX protein. *J Biol Chem* **1999,** *274* (53), 38189-96.

25. Pan, Y.; Chen, J., MDM2 promotes ubiquitination and degradation of MDMX. *Mol Cell Biol* **2003,** *23* (15), 5113-21.

26. Laurie, N. A.; Donovan, S. L.; Shih, C. S.; Zhang, J.; Mills, N.; Fuller, C.; Teunisse, A.; Lam, S.; Ramos, Y.; Mohan, A.; Johnson, D.; Wilson, M.; Rodriguez-Galindo, C.; Quarto, M.; Francoz, S.; Mendrysa, S. M.; Guy, R. K.; Marine, J. C.; Jochemsen, A. G.; Dyer, M. A., Inactivation of the p53 pathway in retinoblastoma. *Nature* **2006,** *444* (7115), 61-6.

27. Bottger, V.; Bottger, A.; Garcia-Echeverria, C.; Ramos, Y. F.; van der Eb, A. J.; Jochemsen, A. G.; Lane, D. P., Comparative study of the p53-mdm2 and p53-MDMX interfaces. *Oncogene* **1999,** *18* (1), 189-99.

28. Hardcastle, I., Inhibitors of the MDM2-p53 interaction as anticancer drugs. *Drugs Fut* **2007,** *32* (10), 883.

29. Francoz, S.; Froment, P.; Bogaerts, S.; De Clercq, S.; Maetens, M.; Doumont, G.; Bellefroid, E.; Marine, J. C., Mdm4 and Mdm2 cooperate to inhibit p53 activity in proliferating and quiescent cells in vivo. *Proc Natl Acad Sci U S A* **2006,** *103* (9), 3232-7.

30. Waning, D. L.; Lehman, J. A.; Batuello, C. N.; Mayo, L. D., Controlling the Mdm2-Mdmx-p53 Circuit. *Pharmaceuticals (Basel)* **2010,** *3* (5), 1576-1593.

31. Bottger, A.; Bottger, V.; Garcia-Echeverria, C.; Chene, P.; Hochkeppel, H. K.; Sampson, W.; Ang, K.; Howard, S. F.; Picksley, S. M.; Lane, D. P., Molecular characterization of the hdm2-p53 interaction. *J Mol Biol* **1997,** *269* (5), 744-56.

32. Shangary, S.; Wang, S., Targeting the MDM2-p53 interaction for cancer therapy. *Clin Cancer Res* **2008,** *14* (17), 5318-24.

33. Vassilev, L. T., MDM2 inhibitors for cancer therapy. *Trends Mol Med* **2007,** *13* (1), 23-31.

34. Buolamwini, J. K.; Addo, J.; Kamath, S.; Patil, S.; Mason, D.; Ores, M., Small molecule antagonists of the MDM2 oncoprotein as anticancer agents. *Curr Cancer Drug Targets* **2005,** *5* (1), 57-68.

35. Patel, S.; Player, M. R., Small-molecule inhibitors of the p53-HDM2 interaction for the treatment of cancer. *Expert Opin Investig Drugs* **2008,** *17* (12), 1865-82.

36. Vassilev, L. T., Small-molecule antagonists of p53-MDM2 binding: research tools and potential therapeutics. *Cell Cycle* **2004,** *3* (4), 419-21.

37. Ding, K.; Lu, Y.; Nikolovska-Coleska, Z.; Wang, G.; Qiu, S.; Shangary, S.; Gao, W.; Qin, D.; Stuckey, J.; Krajewski, K.; Roller, P. P.; Wang, S., Structure-based design of spiro-oxindoles as potent, specific small-molecule inhibitors of the MDM2-p53 interaction. *J Med Chem* **2006,** *49* (12), 3432-5.

38. Dastidar, S. G.; Lane, D. P.; Verma, C. S., Multiple peptide conformations give rise to similar binding affinities: molecular simulations of p53-MDM2. *J Am Chem Soc* **2008,** *130* (41), 13514-5.

39. Shangary, S.; Qin, D.; McEachern, D.; Liu, M.; Miller, R. S.; Qiu, S.; Nikolovska-Coleska, Z.; Ding, K.; Wang, G.; Chen, J.; Bernard, D.; Zhang, J.; Lu, Y.; Gu, Q.; Shah, R. B.; Pienta, K. J.; Ling, X.; Kang, S.; Guo, M.; Sun, Y.; Yang, D.; Wang, S., Temporal activation of p53 by a specific MDM2 inhibitor is selectively toxic to tumors and leads to complete tumor growth inhibition. *Proc Natl Acad Sci U S A* **2008,** *105* (10), 3933-8.

40. Grasberger, B. L.; Lu, T.; Schubert, C.; Parks, D. J.; Carver, T. E.; Koblish, H. K.; Cummings, M. D.; LaFrance, L. V.; Milkiewicz, K. L.; Calvo, R. R.; Maguire, D.; Lattanze, J.; Franks, C. F.; Zhao, S.; Ramachandren, K.; Bylebyl, G. R.; Zhang, M.; Manthey, C. L.; Petrella, E. C.; Pantoliano, M. W.; Deckman, I. C.; Spurlino, J. C.; Maroney, A. C.; Tomczuk, B. E.; Molloy, C. J.; Bone, R. F., Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. *J Med Chem* **2005,** *48* (4), 909-12.

41. Parks, D. J.; Lafrance, L. V.; Calvo, R. R.; Milkiewicz, K. L.; Gupta, V.; Lattanze, J.; Ramachandren, K.; Carver, T. E.; Petrella, E. C.; Cummings, M. D.; Maguire, D.; Grasberger, B. L.; Lu, T., 1,4-Benzodiazepine-2,5-diones as small molecule antagonists of the HDM2-p53 interaction: discovery and SAR. *Bioorg Med Chem Lett* **2005,** *15* (3), 765-70.

42. Koblish, H. K.; Zhao, S.; Franks, C. F.; Donatelli, R. R.; Tominovich, R. M.; LaFrance, L. V.; Leonard, K. A.; Gushue, J. M.; Parks, D. J.; Calvo, R. R.; Milkiewicz, K. L.; Marugan, J. J.; Raboisson, P.; Cummings, M. D.; Grasberger, B. L.; Johnson, D. L.; Lu, T.; Molloy, C. J.; Maroney, A. C., Benzodiazepinedione inhibitors of the Hdm2:p53 complex suppress human tumor cell proliferation in vitro and sensitize tumors to doxorubicin in vivo. *Mol Cancer Ther* **2006,** *5* (1), 160-9.

43. Chen, L.; Yin, H.; Farooqi, B.; Sebti, S.; Hamilton, A. D.; Chen, J., p53 alpha-Helix mimetics antagonize p53/MDM2 interaction and activate p53. *Mol Cancer Ther* **2005,** *4* (6), 1019-25.

44. Lu, Y.; Nikolovska-Coleska, Z.; Fang, X.; Gao, W.; Shangary, S.; Qiu, S.; Qin, D.; Wang, S., Discovery of a nanomolar inhibitor of the human murine double minute 2 (MDM2)-p53 interaction through an integrated, virtual database screening strategy. *J Med Chem* **2006,** *49* (13), 3759-62.

45. Stoll, R.; Renner, C.; Hansen, S.; Palme, S.; Klein, C.; Belling, A.; Zeslawski, W.; Kamionka, M.; Rehm, T.; Muhlhahn, P.; Schumacher, R.; Hesse, F.; Kaluza, B.; Voelter, W.; Engh, R. A.; Holak, T. A., Chalcone derivatives antagonize interactions between the human oncoprotein MDM2 and p53. *Biochemistry* **2001,** *40* (2), 336-44.

46. Galatin, P. S.; Abraham, D. J., A nonpeptidic sulfonamide inhibits the p53-mdm2 interaction and activates p53-dependent transcription in mdm2-overexpressing cells. *J Med Chem* **2004,** *47* (17), 4163-5.

47. Hu, B.; Gilkes, D. M.; Farooqi, B.; Sebti, S. M.; Chen, J., MDMX overexpression prevents p53 activation by the MDM2 inhibitor Nutlin. *J Biol Chem* **2006,** *281* (44), 33030-5.

48. Pazgier, M.; Liu, M.; Zou, G.; Yuan, W.; Li, C.; Li, J.; Monbo, J.; Zella, D.; Tarasov, S. G.; Lu, W., Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX. *Proc Natl Acad Sci U S A* **2009,** *106* (12), 4665-70.

49. Bullock, A. N.; Fersht, A. R., Rescuing the function of mutant p53. *Nat Rev Cancer* **2001,** *1* (1), 68-76.

50. Fojo, T., p53 as a therapeutic target: unresolved issues on the road to cancer therapy targeting mutant p53. *Drug Resist Updat* **2002,** *5* (5), 209-16.

51. Ding, K.; Lu, Y.; Nikolovska-Coleska, Z.; Qiu, S.; Ding, Y.; Gao, W.; Stuckey, J.; Krajewski, K.; Roller, P. P.; Tomita, Y.; Parrish, D. A.; Deschamps, J. R.; Wang, S., Structure-based design of potent non-peptide MDM2 inhibitors. *J Am Chem Soc* **2005,** *127* (29), 10130-1.

Chapter 8: Toward The Activation Of The R248Q P53 Mutant: A Never Solved Puzzle[(1)]

8.1 Introduction

This chapter extends the preceding p53-exercise to a further complicated problem. As p53 is the most mutated protein in human cancers,[1] and mutations of p53 alone account for more than half of invasive types of cancer,[2] a simple idea to cure these types of cancer is to reactivate the mutated p53 variants. However, realizing this idea has turned out to be far more complicated. According to the latest version (R15) of the *TP53* mutation database,[3] 27 580 different somatic mutations have been identified in the full-length protein and the overwhelming majority of alterations are located within the core DNA-binding domain (DBD). More importantly, ~75% of the resulting mutants are, fundamentally, full-length proteins with single amino acid substitutions in the DBD. In addition, about 40% of the DBD mutations are concentrated at six particular hot-spots: Arg-175, Gly-245, Arg-248, Arg-249, Arg-273 and Arg-282.[4]

8.1.1 Contact Vs Structural Mutations

Of the six hot-spot residues listed above, alterations at Arg-248 and Arg-273 are classified as DNA contact mutations whereas substitutions at the remaining sites are structural mutations. Contact mutants are characterized by the direct loss of the sequence–specific transactivation activity while retaining the wild-type (WT) conformation.[5] Structural mutations, on the other hand, involve residues primarily responsible for maintaining the conformational integrity of the DBD and stabilizing the p53 DNA–binding surface. Such alterations generate local structural defects, which in turn transfer to critical regions of the DBD, causing indirect loss of

[(1)] A version of this chapter has been published in Barakat K, Issack BB, Stepanova M, Tuszynski J. PLoS One. 2011; 6(11): e27651.

DNA binding.[6] Failure to bind DNA prevents p53-dependent transcription and hence inhibits p53-mediated tumor suppression.

8.1.2 Successful P53 Activators

More than 50% of human tumors exist because of a defect in p53 as a result of single-site mutations. Most of these mutations are enormously frequent. Is it possible to develop a small molecule activator that reverses this misfolding behavior in tumors? A number of experimental studies suggest that tumor cells would be highly sensitive to such activators. On one hand, mutants are more stable and highly abundant than wild type protein.[7] Except for Li-Fraumeni patients who, essentially carry germline p53 mutations, mutant p53 proteins are not abundant in normal cells, however they are over-expressed in tumor cells. This is due to the occurrence of p53-activating signals in cancer cells and inability of mutants of p53 to induce expression of their own inhibitors MDM2 and MDM4. On the other hand, mere synthesis of p53 in normal cells does not activate the p53 pathway, while restored mutants in tumor cells activate the pathway. This indicates that cancer cells have a particular mechanism that is ready to activate newly expressed wild type p53.[8] Based on these facts, one can regard p53 mutants as a “loaded gun”, whose trigger is obstructed by genetic alterations.[9] Consequently, inhibiting mutant proteins using small molecules would be specific and effective in treating cancers, while avoiding the most damaging and unfavorable effects associated with the majority of current cancer therapies which harm normal cells. Many research groups attempted to devise small peptides or molecules that could restore the activity of mutant p53. These results are summarized below.

8.1.2.1 9-hydroxy-ellipticine (9HE)

Ellipticine (EPC) is a relatively old anti-cancer agent extracted from Aspidosperma williansii (Apocynaceae) and purified in the late 1960s. The compound is too toxic to be used clinically. In 1999, Sugikawa and his team discovered 9-hydroxy-ellipticine (9HE), a derivative compound of ellipticine, as an activator for mutants on the three sites 143, 175 and 273.[10] The compound induced G1 arrest and triggered G1 phase-restricted apoptosis in several tumor cells. The exact restoration mechanism of 9HE is still unknown.

8.1.2.2 CP-31398

The profound proof-of-concept for the mutant reactivation strategy was first introduced by Foster and collaborators.[11] Foster, experimentally, screened a chemical library of compounds against the DBD of p53 and the two monoclonal antibodies PAB1620 and PAB240. They identified CP-31398, a small molecule that restored mutants at the two sites 173 and 249, up-regulated p53 target genes and suppressed tumor growth in mice. Further studies by Demma et al. showed that CP-31398 is specific to p53 and does not target its homologous proteins p73 and p63.[12]No precise information exists on how CP-31398 induces its activity and restores these mutants.

8.1.2.3 CDB3

Fersht's group made another advance toward discovering activators for p53 mutants.[13] They proposed a rational strategy to identify such molecules, called the chaperone strategy. Their theory states that: "conformationally compromised oncogenic mutants of the tumor suppressor protein p53 can, in principle, be rescued by small molecules that bind the native, but not the denatured state". Based on this idea, they derived a nine-residue peptide from a p53 binding protein, CDB3. Interestingly, the peptide activated the R249S mutant. What was unique to this project is that NMR data showed that CDB3 binds approximately around the edge of the p53 binding surface, stabilizes the complex and increases its melting temperature. Fersht also speculated that CP-31398 has the same chaperone mechanism as CDB3. Perhaps it binds to the same location as their peptide.

8.1.2.4 PRIMA-1

P53 reactivation and induction of massive apoptosis (PRIMA-1), is a small molecular activator for the His273 mutant.[14] Bykov et al. ran a cell-based screening assay on a cell line that possesses the His273 mutation, using the NCI diversity set as their tested compounds. This led to PRIMA-1, a compound that inhibited cell growth in a mutant p53-dependent manner. It also inhibited cell growth in cell lines that have a His175 mutation. Recently, Lambert et al.[15] suggested that PRIMA-1, covalently, forms adducts with thiols in mutant DBD, which in turn could be the reason for reactivating the mutant protein and inducing apoptosis in tumor cells.

8.1.2.5 MIRA-1

Mutant p53-dependent induction of rapid apoptosis or MIRA-1 was identified from the same screening assay that discovered PRIMA-1.[16] However, it was more potent and structurally different than PRIMA-1. The compound retained DNA binding, mended the mutant conformation in vitro and restored the p53 trans-activation ability in living cells. One interesting observation from this work is that DNA binding assays showed that MIRA-1 stimulated DNA binding of some but not all mutant forms of p53. They noticed also that for one mutant, Gln248, MIRA-1 activated

DNA binding in Namalva cells but not in BL41 cells. Likewise, for the His273 mutant, DNA binding was retained in Saos-2 or SKOV cells, while no effect of the compound was found in the SW80 cell. These finding lead to the conclusion that many cellular factors are involved in the ability of MIRA-1 to reactivate mutant p53.

8.1.2.6 PhiKan083

The work presented in this study is, in fact, more supportive and similar to our proposed research. Fersht's team targeted the Y220C mutant as a rational drug design case. This mutation forms a cavity on a non-functional site of the protein, reducing the protein stability by ~4 kcal/mol. Consequently, they used virtual screening to identify compounds that can complement this cavity in terms of shape, charge and other chemical and physical properties. This search led to a number of small complementing molecules. One of them, PhiKan083, was a potent binder to the cavity. It increased the melting temperature of the mutant and slowed down its rate of degradation. They also confirmed the biding of PhiKan083 through X-ray crystallography showing how the compound interacted with the cavity and induced a conformational change in the protein structure.

8.1.3 Thermodynamic Stability Of The DBD

Investigations of the thermodynamic stability of the DBD have revealed the destabilizing nature of hot-spot mutations relative to WT p53[17,18,19] and highlighted the temperature-dependence of their DNA-binding affinity.[18,20] Other studies focused on different thermodynamical parameters that can determine the ultimate stability of the protein and its mutants. Such experiments included measuring pressure-stability at different temperatures [21,22,23,] different pHs[24] and studying the effect of DNA-binding on the core domain stability.[25]

The first insightful evidence for the importance of temperature in proper DNA binding was reported by Zhang and collaborators in 1994 for Ala-143, which was considered a hot-spot mutation at the time.[20] The mutant p53 exhibited high DNA-binding affinity at temperatures of 306 K and below, as well as stronger transcriptional activity than WT p53. At physiological temperature both the DNA-binding and transcriptional activation functions of the mutant were significantly reduced. These observations were rationalized in terms of a two-conformational state model: a mutant conformation at physiological temperatures, and a wild-type conformation at lower temperatures. Friedlander *et al.* examined the effects of temperature on a wide range of p53 mutants.[26] This included Ala-143, His-175, Trp-248, Ser-249 and His-273. With the exception of His-175, all mutants were able to bind DNA at sub-physiological temperatures (298-306 K). At 310 K, however, their binding to DNA was defective. Numerous other temperature-sensitive mutations were later identified and targeted for restoration.[27] Ishioka's group alone assessed a collection of over 2,000 p53 mutants for temperature sensitivity and identified 113 mutants with activity at 303 K.[28] This represents about ~10% of all reported single amino acid alterations of the DBD in human cancers.[3] Here, we focused on the R248Q mutant, which is the most frequently occurring mutation in human cancer. It is mostly associated with breast, colon, head, neck and skin cancers. Moreover, it ranks as the second most mutants in esophageal, gastric, lung, ovarian, and prostate cancers.

8.1.4 Temperature-Dependence of the Arg-248 Mutants

Mutations at the Arg-248 residue of p53 have been of substantial interest to a large group of cancer researchers. Many experiments were conducted in order to better understand their roles. With the objective of restoring the activity of mutated p53 proteins, many researchers employed various experimental and theoretical techniques aimed at understanding why they are inactive in cancer cells. The work presented here was inspired by many experimental studies that focused on the effects of temperature on the stability, structure and transcriptional activity of p53 and its Arg-248 mutants. For example, Bullock *et al.* investigated the wild-type stability along with a number of its mutants including R248Q at both low and high temperatures [17]. Their work revealed that the R248Q mutant is stable at sub-physiological low temperatures. The R248Q stability was less than that of the wild-type protein by ~2 kcal/mol. The mutant structure also retained a two-stage unfolding transition, similar to the wild-type protein [21], which indicated well-defined structures at low temperature. Interestingly, the addition of a 22-mer double-stranded DNA p53 consensus sequence raised the melting temperature of the tested proteins, signifying a stabilizing effect due to DNA-binding [17]. The effect of DNA on stabilizing the core domain was recently confirmed by Ishimaru *et al.* [25]. An interesting study by Wong *et al.* investigated the structural changes introduced by five hot spot mutations including R248Q at low temperature using chemical shift changes [29]. Their findings indicate that the R248Q mutation induces structural changes in L2 and L3 regions of the core domain at 310 K. That is, the R248Q mutation has the dual capacity of being both a contact and a structural mutation. These structural changes lower the binding affinity to the DNA without significantly destabilizing the protein [30]. In fact, at high pressure and low temperature, WT p53 can adopt the R248Q mutant structure [21]. Benoit *et al.* investigated the transcriptional activation of cyclooxygenase-2 (Cox-2) by p53 at low temperature [31]. They also examined Cox-2 transcription induced by different p53 mutants including the R248Q variant. Cooperating with nuclear factor-kappaB (NF-kappaB), R248Q produced a significant increase in Cox-2 expression similar to the wild-type protein.

Other common mutations of the Arg-248 residue (e.g. R248W and R248A) also expressed a profound dependence on temperature. The most perceptible behavior was noticed in the case of R248W [26], [28]. Friedlander *et al.* [22] showed that R248W can effectively bind to DNA at low temperatures and this binding activity is significantly diminished at physiological temperatures. A kinetic stability experiment on a number of different p53 mutants revealed that R248A had a half-life time ($t_{1/2}$) of 128 minutes at low temperature compared to less than 3 minutes at 310 K. The analysis in this study revealed an important concept in understanding the stability of p53 mutants. Namely, there is a remarkable correlation between the thermodynamic and kinetic instability of the mutants. The more unstable the mutant, the shorter its half-life time. This means that R248A is more stable at low temperatures than at physiological temperatures. All of the above-mentioned experimental data reveal a clear connection between temperature and the stability and activity of p53 R248 mutants in general and the R248Q mutant in particular.

8.2 Results And Discussion

In this chapter, we report on the results of molecular dynamics (MD) simulations that have been carried out for the DBD of WT and R248Q p53 molecules in the presence or absence of a DNA duplex at 300, 305 and 310 K. A comprehensive assessment of the influence of temperature on p53-DNA intermolecular interactions has been performed in terms of structural, dynamical and thermodynamic properties. The main aims of this work are to determine the effects of temperature on the conformations of WT and mutant p53 complexes and to identify key residues or regions of the complexes, which modulate changes in DNA-binding at the different temperatures. Our results indicate that temperature plays an essential role in the stability of the hydrogen bond network and binding properties of p53-DNA complexes over both short and long time-scales. The outcome of our study provides new insights into the way towards restoring apoptosis in the above-mentioned types of cancer cells by activating the p53 pathway of tumorigenic R248Q mutants.

8.2.1 MD Simulations of the Wild Type And Mutant Structures

The root mean square deviations (RMSD) of backbone atoms of the DBD and DNA duplex (for the p53-DNA complexes) were computed over the final ns of each trajectory. The results are shown in Figure 8-1 for the wild type at 300 K. In the rest of the cases the behavior was similar (data not shown). The RMSDs of DNA are significantly higher and are associated with larger fluctuations than those of the protein in all trajectories. The higher mobility of the DNA backbone relative to the protein backbone in both complexes at all temperatures can be attributed to the dynamics of the terminal residues of the double helix that are not bound to the DBD. The RMSD plots of DNA-bound and DNA-free proteins are generally similar. The mean RMSD of the DBD is slightly smaller in the p53-DNA complexes than in the apo-structures for both p53 variants. Similar observations were reported by Noskov *et al.*[32] for the same protein at 300 K. In general, the backbone RMSDs appear to be relatively stable to temperature changes over the range investigated.

8.2.2 Hydrogen Bonding & Water Distribution

Several reports of the crystal structure of the p53-DNA complexes have highlighted the central roles of water molecules and hydrogen bonds in stabilizing interactions between the two biomolecules. In some of these investigations, the failure of p53 to bind DNA has been correlated with the loss of one or two hydrogen bonds mediated by a single residue within the DBD. For example, the mutations of the hot-spot residue Arg-273[33] or Arg-249[2] into histidine or serine respectively, induces a sequence of hydrogen bond disruptions that ultimately lead to the loss of DNA binding. For these two mutations, the hydrogen bond network could be restored or compensated by means of an additional single mutation. Changing the 284-residue to arginine conferred DNA-binding ability to the R273H mutant. Similarly, substitution of the residue at position 268 by arginine partially restored the activity of the R249S mutant. The influence of hydration on p53 folding has been studied by Silva *et al.* and revealed that water interactions with both p53 and the DNA were essential for proper folding and enhanced stability of the complex [23]. The presence of a DNA molecule augmented the stability of the DBD within p53 [25]. Below we confirm these concepts by investigating the dynamical character of the hydrogen bond network. We also compare the different connections among the protein, DNA and water residues for both the wild type and the R248Q mutant at all temperature ranges investigated.

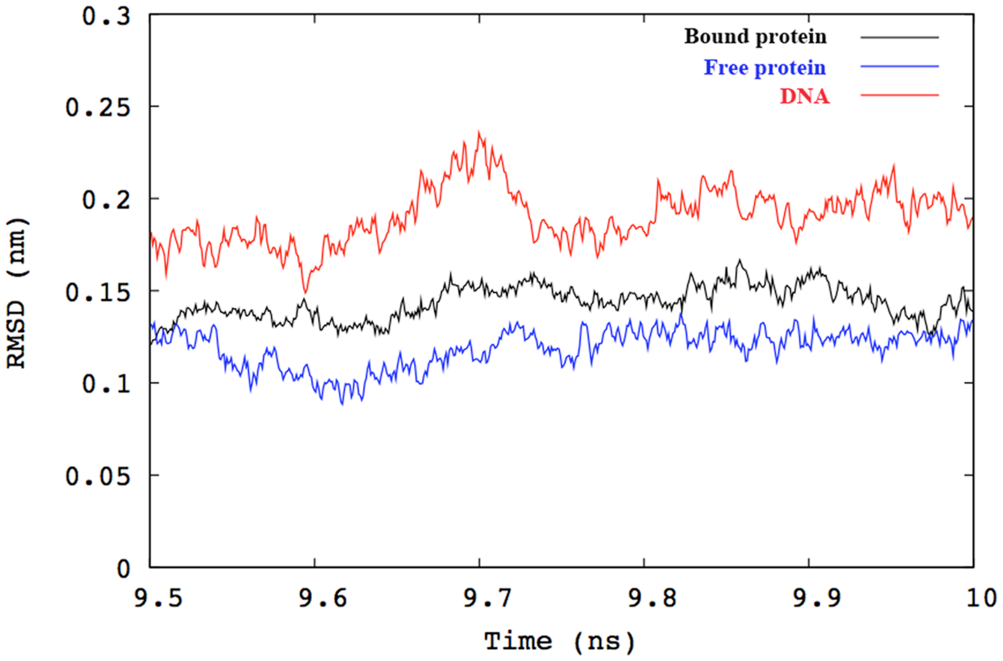


Figure 8-1: Plots of backbone RMSD for the DNA-bound and DNA-free WT p53 at 300 K over the final ns of 10 ns trajectories.

The black, blue and red lines correspond to RMSD values of the DBD bound to DNA, in absence of DNA, and of DNA only, respectively.

Figure 8-2 and Figure 8-3 describe the complicated hydrogen bond networks formed by interfacial atoms in the dominant structures extracted from clustering of the MD simulations. In Figure 2-A, several direct contacts can be identified between the DBD of WT p53 and DNA nucleotides at 300 K. Arg-248 from the loop L3 protrudes into the minor groove of the DNA molecule resulting in favorable electrostatic interactions between the positively charged guanidinium group of Arg-248 and the negatively charged DNA backbone. The minor groove adjacent to Arg-248 is compressed and its bases are buckled so that the side chain of Arg-248 makes three direct contacts with the DNA. Likewise, the side-chains of Cys-242, Lys-120 and Ser-116 directly interact with DNA. Within the protein structure, Cys-277 is hydrogen-bonded to the side chain of Lys-120. In addition to direct p53-DNA contacts, seven ordered water molecules are located at the interface. Among these water molecules are conserved crystallographic water molecules present in the original crystal structure, thus supporting their inclusion in the starting structures for MD simulations. For clarity, only water molecules participating in the hydrogen bond network and which act as linkers between the different interacting residues are depicted in the figures. Water molecules appear to have a stabilizing role on the direct p53-DNA contacts. W1 and W2 connect Arg-248 to DNA through three different hydrogen bonds. W3 mediates an interaction between the side chain of Asp-281 and the backbone of Ala-276 while at the same time connecting them to the guanine base of DG-303. W4 and W5 are involved in water-bridged hydrogen bonds linking Asn-239 to Cys-277 and Ala-276 to Cys-277, respectively. W6 and W7, on the other hand, are responsible for maintaining a hydrogen bond network through which Ser-121 interacts with the DNA molecule via two different hydrogen bonds. Among the residues identified in the vicinity of DNA, Lys-120 and Ser-121 have been suggested as key participants in DNA binding in a crystallographic analysis of DNA-bound and DNA-free forms of the WT DBD.[34] In addition, p53 DBD lacking residues 100-120 displayed reduced binding during antibody binding experiments.[35]

Raising the temperature to 305 K does not significantly alter the overall structure of the protein-DNA binding interface or its hydrogen bond network (see Figure 2-B). Arg-248 has retained one of the direct contacts and two water-bridged hydrogen bonds (W1, W2) with the DNA molecule. A direct contact between DNA and Ser-241, that was absent at 300 K, is present at 305 K alongside nucleotide-interactions with Asn-239, Cys-277 and Lys-120. Moreover, the water molecules W3, W4 and W5 mediate interactions between Asn-239, Cys-275 and DNA. At 310 K, interactions between Arg-248 and DNA amount to four hydrogen bonds and no water-bridged linkages are present, as shown in Figure 2-C. However, W1 is involved in connecting Arg-248 and Ser-241, which in turn interacts with DNA nucleotides through W2 and W3. Cys-275 is connected to DNA through its side chain and through W4. Cys-277 is involved in an extensive hydrogen bond network via backbone interactions with the side chain of Asn-239 mediated by W5, W6 and W7 and side-chain interactions with DNA and Lys-120. Moreover, Ser-121 and Ser-116 are connected to DNA through the two water molecules; W8 and W9. Finally, similar to the previous cases, Lys-120 maintains its direct connection to DNA and to Cys-277. These strong interactions and persistence of the native fold of p53 DBD confirmed the fact that the wild type is stable at all three temperatures [17]. Interactions with the DNA molecule are extremely favored by the protein. They enhance its stability and prevent it from misfolding or aggregating.[25]

Armed with the validation of our protocol in reproducing the native p53 conformation and constructing a fine grid of detailed hydrogen bonding interactions, we proceeded to investigate in detail the R248Q case. Switching to the mutated p53 structure has yielded interesting findings. Figure 3-A illustrates the hydrogen bond network at 300 K. Glu-248 is connected to DNA through a direct hydrogen bond and a water-mediated hydrogen bond. This water molecule, W1, also connects Glu-248 to Asn-247, which was absent in WT p53 at any of the three temperatures. Once more, water molecules play a major role in coordinating a number of hydrogen bonds at the DNA-p53 binding interface. For example, W2 connects Ser-241 to the DNA molecule. W3 and W5 connect Asn-239 to Cys-275. W4 and W6 connect the backbone of Ala-276 to the guanine residue DG-303 and its side chain to the cytosine residue DC-304, respectively. W8 mediates a superior interaction between the backbone of Lys-120 and guanine residue DG-318.

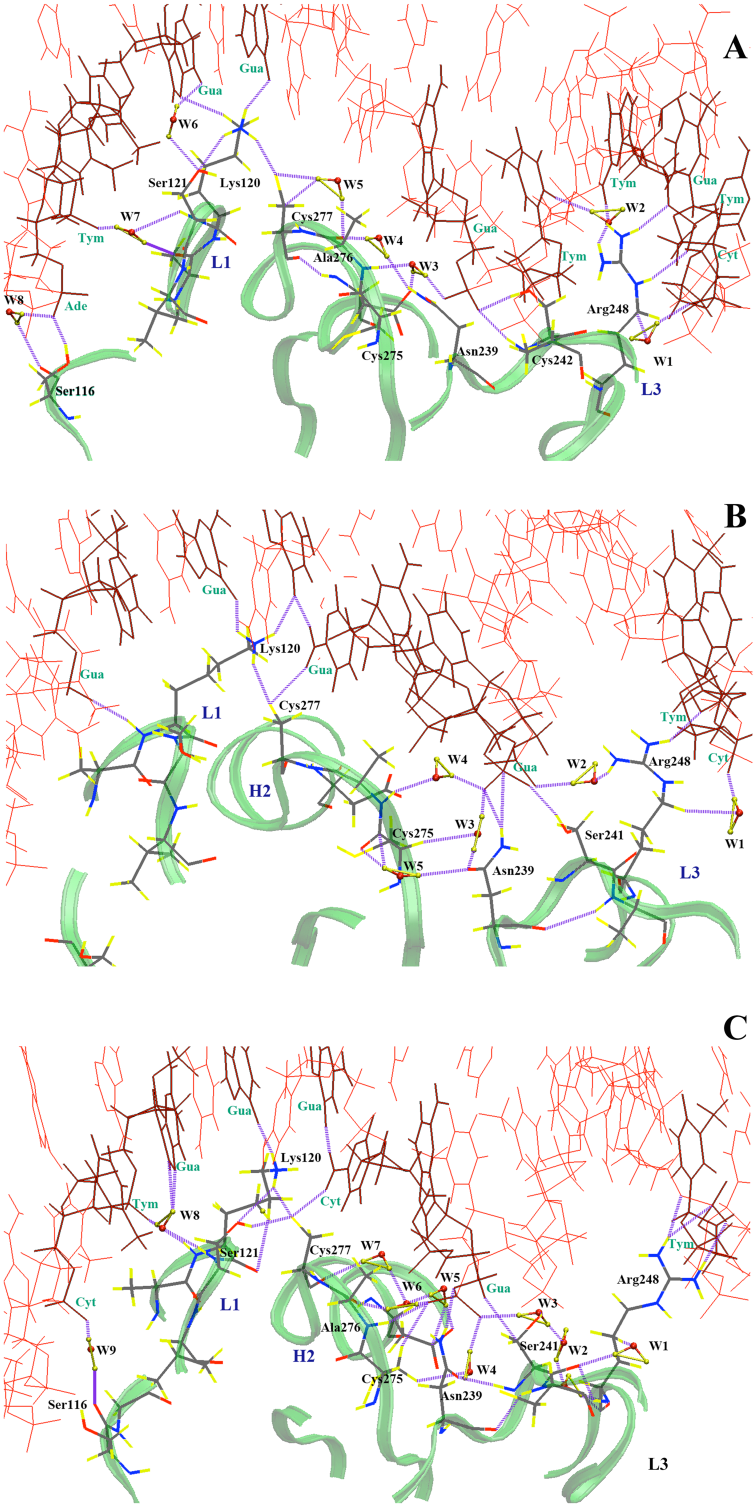


Figure 8-2: Hydrogen bond network for the wild type at three different temperatures. (A) 300 K, (B) 305 K and (C) 310 K.

In addition, the side chain of Ser-241 is hydrogen bonded to the side chain of Asn-239, and as observed for WT p53, Cys-277 preserves its hydrogen bond with Lys-120, which maintains direct contact with DNA. At 305 K (see Figure 3-B) an important modification takes place. Glu-248 loses its direct contact with the DNA molecule as the distance between the side chain of Glu-248 and the closest DNA residue is greater than 4 Å. This results in a large gap between the protein and DNA and leads to a distortion in the minor groove close to Glu-248. Despite this deviation, Glu-248 participates in two hydrogen bonds with DNA through two water molecules, W1 and W4. A fine hydrogen bond network connects guanine DG-324 through four water molecules (W2, W3, W4 and W5) to guanine DG-303 in the middle of the DNA. This final guanine residue is directly connected to Ser-241, Asn-239 and Ala-276. Surprisingly, Cys-276 maintains its interaction with Lys-120, which was connected to DNA through a water molecule, W9, unlike the previously mentioned cases where Lys-120 had a direct contact to the DNA. Finally, Ser-121 is connected to DNA through W10. A huge difference is found to occur at 310 K (see Figure 3-C). The separation between Glu-248 and DNA is more than 6 Å. No direct hydrogen bonds are established to connect Glu-248 to DNA. The DNA terminal near Glu-248 is completely distorted and separated from the protein. However, Glu-248 is connected to the center of the DNA duplex through a water-mediated hydrogen bond (W1) that also links Ser-241 to DNA. Cys-275 is connected to DNA through W3. Ala-276 is hydrogen bonded to Lys-120 while Ser-241 is attached to DNA via two water-mediated hydrogen bonds.

The aforementioned atomistic details of the DNA-contact geometry reveal a reasonable dependence on temperature. The L3 loop was directly linked to the minor groove of the bound DNA via Arg-248 at three different temperatures for the wild type, or via Glu-248 at 300 K for the R248Q mutant. During the six different simulations, the conformations of R/Q248 side chain were fully extended and contacted the DNA nucleotides either directly or indirectly through water molecules. It has also become apparent that the L3 loop plays a dual role in DNA binding. Besides contacting DNA through Arg-248, it is also an essential part of the DBD of p53 by aiding in the stabilization of the zinc-binding site and hence can affect other regions of the protein. Although the minor groove area is largely affected upon the mutation at physiological temperatures, the major groove contacts, *i.e.*, Lys-120, Ala-276 and Cys-277 maintain their interactions with DNA even after Arg-248 was mutated to glutamine. In addition to the well-documented stabilizing roles of water-mediated interactions in biological complexes, hydration can also have a destabilizing effect, as recently described by Silva *et al.*[25] The authors attributed the enhanced stability of cognate DNA-WT p53 complexes to the exclusion of unfavorable water-mediated interactions from the protein surface. Conversely, infiltration of water inside mutant complexes would be responsible for their destabilization and promote aggregation of p53 molecules.

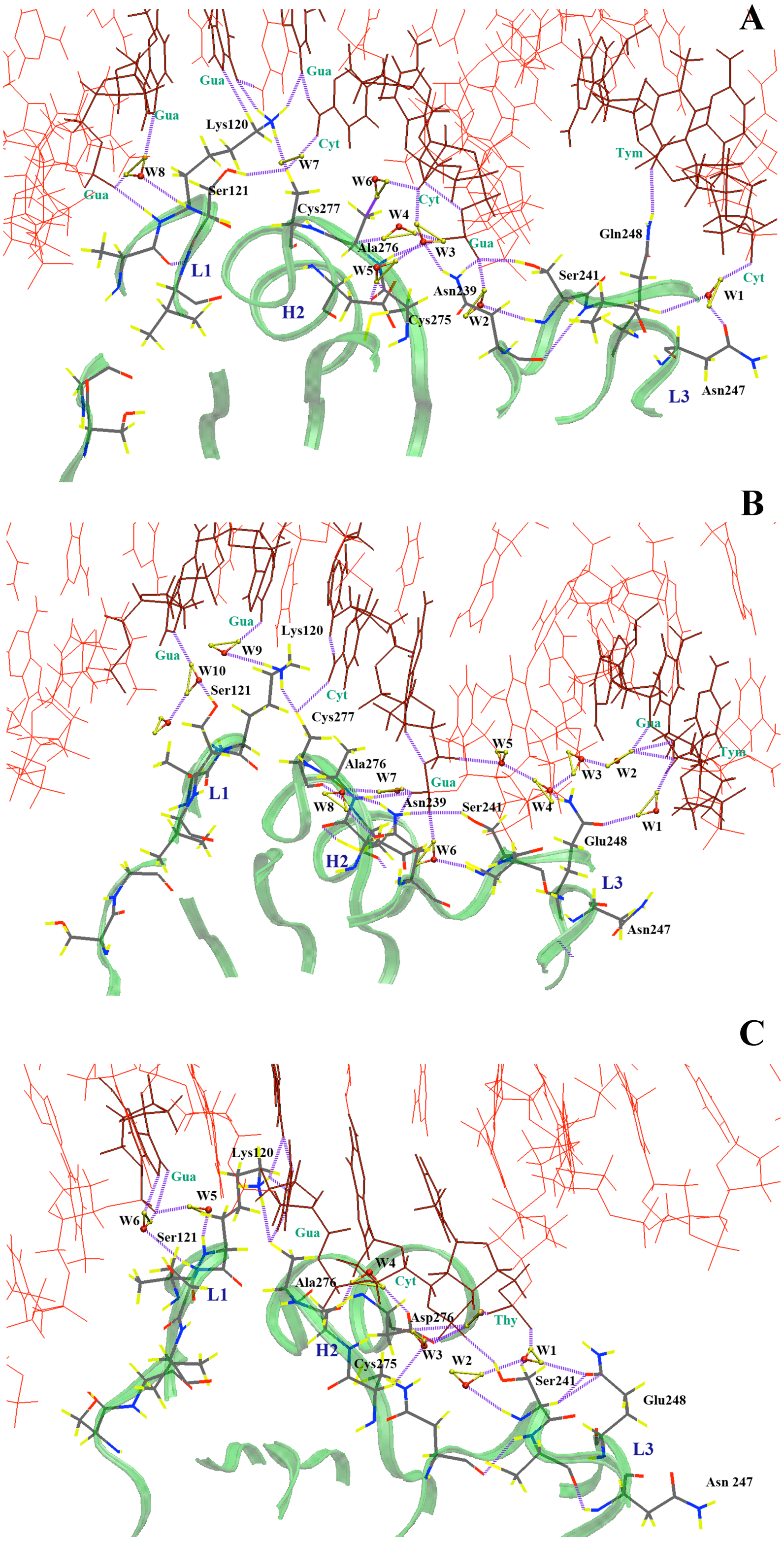


Figure 8-3: Hydrogen bond network for the Q248 mutant at three different temperatures. (A) 300 K, (B) 305 K and (C) 310 K.

The above reasoning suggests that the lower stability of the R248Q p53 complex at high temperatures is a result of structural changes in its hydration networks, as evidenced by the formation of an interfacial water-filled cavity at 305 K and the transformation of direct DNA contacts into water-mediated interactions at 305 and 310 K. Therefore, alterations in the hydrogen bond network provide an effective structural framework for understanding changes in DNA binding for the R248Q mutant p53 at physiological temperatures.

8.2.3 Binding Energy Analysis

MD simulations of the p53–DNA complexes and the hydrogen bonding analysis provided valuable insights into the dynamics of their interactions and the role of water at the interface of complexes. Our next step was to investigate the influence of temperature on the stability of the p53 variants. To this end, the thermodynamics of p53-DNA binding were evaluated using the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method, a well-established technique that takes into account the effects of solvation, ionic concentrations, entropy and molecular mechanics interactions. It has been previously employed in many similar studies[36],[37],[38] and has produced accurate free energy estimates at a reasonable computational cost. Its main advantages include the lack of adjustable parameters and the possibility of using a single MD simulation for the complete system to determine all energy values.

The binding energy calculations are listed in Table 8-1 for the two p53 structures at three different temperatures. It should be mentioned that binding energies are reported relative to the WT binding energy at 300 K, which was estimated as -12 kcal/mol. Our calculations indicate that binding to DNA is maintained by the WT protein both at 305 and 310 K. This is supported by experimental evidence that WT p53-complex has a melting temperature of 322 K, indicating that the complex is stable at 310 K.[17] While our results indicate that the binding affinity is enhanced at physiological temperature, *in vitro* measurements showed a decrease in the binding affinity of WT p53 at 310 K.[26],[39] The conflicting observations may be related to experimental conditions and techniques. It has been shown that the stability of p53 and DNA binding affinity is highly sensitive to ionic strength, DNA sequence and pressure.[40],[26] Nonetheless, the results agree on the qualitative aspects of binding, *i.e.*, WT p53 DBD can bind to the DNA at all three temperatures and also validates the MM-PBSA method as an adequate binding energy evaluation technique.[26] At 300 K, the binding energy of R248Q is decreased by ~3 kcal/mol compared to the WT at the same temperature, signifying its possible binding to DNA. When the temperature is raised to 305 K and 310 K, the binding energy of the mutant p53 increases by 12 and 15 kcal/mol, respectively, relative to the WT. These observations indicate that binding of R248Q to DNA becomes highly unfavorable with increasing temperature. Taken together with our observations from hydrogen bond analysis, changes in the binding energy of the mutant p53 may be interpreted as a significant weakening of DNA-binding at 305 and 310 K while WT p53 retains its binding characteristics at these same temperatures.

Type	T (K)	BE_T-BE_{WT300K} (kcal/mol) ± 1
WT	300	0
	305	3
	310	-12
R248Q	300	3
	305	12
	310	15

Note: BEWT300K = -12 kcal/mol

Table 8-1: Binding energy changes between DNA and the p53 core domain due to temperature alterations. All binding energies are relative to that of the WT at 300 K. Our calculations predict that the WT p53 maintains its DNA binding at all temperatures. On the other hand, while the Glu-248 mutant (R248Q) does not lose its DNA binding activity at 300 K, binding is highly unfavorable at 310 K.

To further identify the regions of the protein that cause the loss of DNA binding, we decomposed the binding energy into residue contributions. Table 8-2 lists the individual contributions of residues that amount to at least ±1 kcal/mol of the binding energies, computed at 300, 305 and 310 K for the WT and R248Q p53. Again, the reported binding energies are relative to the WT at 300 K. Comparing the WT to the mutant p53, as expected, the substitution of arginine to glutamine carries the largest penalty which is associated with a cost of ~ 8 kcal/mol at all temperatures. The residues Ser-241 and Asn-239, which are close to the mutation site, reduce the binding energy by ~ 4 kcal/mol. This loss of binding energy is balanced by gains at the 119, 120, 276 and 277 sites.

Comparing these findings to the hydrogen bond analysis mentioned earlier reveals an outstanding correspondence. The stability of the hydrogen bond network at the three different temperatures in the wild-type protein indicates an unremitting binding to DNA. On the other hand, the lack of strong hydrogen bonding in the mutant variant at higher temperatures, namely 305 K and 310 K, causes a parallel effect on the binding affinity to DNA. In general, our analysis reveals that temperature-sensitive residues are located in the three loops and in the C-terminal region. The substitution of arginine by glutamine at residue 248 leads to changes in binding far from the mutation site, particularly in loop L1. This is consistent with differences observed in the major groove contacts during hydrogen bond analysis, and with the proposed classification of the R248Q as a dual structural/contact mutant.[14],[29] In addition, the zinc ion contributed significantly to the overall binding energy between the protein and DNA in all simulations. These energies ranged from -7 kcal/mol for the WT to -8 kcal/mol for the mutant p53 at 300 K. These results are consistent with the findings of Butler et al. that zinc is crucial for proper DNA binding, and that the stability of the zinc ion within the R248Q mutant is quantitatively comparable to that of the WT protein.[30]

Residue	BE$_{WT}$-BE$_{WT300K}$ (kcal/mol)			BE$_{RQ}$-BE$_{WT300K}$ (kcal/mol)		
	Temperature (K)			Temperature (K)		
	300	305	310	300	305	310
119	0	-3	-5	-5	-1	-4
120	0	1	-4	-3	2	-1
122	0	1	0	2	1	2
174	0	0	1	0	0	1
180	0	-1	-1	0	0	-1
184	0	0	-1	0	-1	-1
239	0	2	3	0	-1	2
240	0	0	0	0	0	1
241	0	0	0	1	1	2
243	0	0	0	-2	-1	0
248	0	1	-2	8	9	8
273	0	0	0	1	3	1
275	0	0	-1	-2	-1	0
276	0	-1	-2	-1	-2	-1
277	0	-2	-1	-2	-1	-2
ZN+2	0	0	0	-1	-1	0
DNA300	0	0	0	1	0	0
DNA301	0	-1	0	0	1	0
DNA302	0	2	2	1	3	3
DNA303	0	2	0	0	-1	2
DNA304	0	0	-3	-2	-2	-2
DNA315	0	-1	-1	-1	-1	-2
DNA306	0	-1	1	0	0	-1
DNA317	0	-1	-3	2	1	2
DNA318	0	2	3	3	2	3
DNA319	0	0	0	0	1	0
DNA320	0	-2	0	-2	-1	-2
DNA324	0	1	1	3	2	2
DNA325	0	2	1	2	1	1
DNA326	0	1	0	0	0	1

Table 8-2: Binding energy decomposition per residue for WT and R248Q p53-DNA complexes at 300, 305 and 310 K. Binding energies are given relative to the energy of the DNA-bound WT p53 complex at 300 K. Residues 119, 120, 248 and 277 from p53 contributed the most to temperature-induced changes in binding energy. At least eight DNA residues involved in close contacts with the protein contributed significantly to binding.

8.3 Conclusion

In about half of human cancers, p53 is inactivated by mutations located primarily in it's DNA binding domain (DBD).[41,1] More than 1,300 distinctive carcinogenic single amino-acid alterations in the core domain of the protein have been reported. Of these, the R248 into Q mutant is one of the most frequently encountered in human cancers. The rescue of DNA-binding and subsequent restoration of activity in mutant p53 is a challenging strategy in developmental cancer therapy. The viability of the approach relies on designing small molecule drugs that reactivate p53 mutants upon binding which requires a thorough understanding of both the effects of p53 mutations and the molecular basis of the resulting inactivation.

Experiments have revealed that the mutation of R248, among others, possesses temperature-induced DNA-binding characteristics. In particular, R248-p53 mutant was defective for binding to DNA at 37°C although it was able to bind specifically to several p53 response elements at sub-physiological temperatures (25–33°C)[18,20,26].

In this work, MD simulations of the p53–DNA complexes and the hydrogen bonding analysis provided valuable insights into the structure and short-time dynamics of p53-DNA interactions and revealed the central role of water at the interface of the complexes. Hydrogen bond networks involving major groove contacts are retained in the R248Q hot-spot mutant at 300, 305 and 310 K. However, direct minor groove contacts involving the mutated residue were disrupted above 300 K. Accordingly, estimates of binding energy show that interactions between DNA and the R248Q mutant become increasingly unfavorable above 300 K. By decomposing the calculated energies into individual contributions, the mutated residue Q248 together with residues in loop L1 and the short loop preceding H2 were identified as key participants in DNA-binding. These findings highlight the critical nature of the R248-DNA interactions and suggest that targeting the mutated residue may bring about restoration of the p53 activity in contact mutations.

The protocol employed in the present work can be applied to other mutant structures in order to compile a new data set of p53 structural information. The resulting data set together with existing structural data can form the basis for developing a novel class of chemotherapeutics, targeting the most frequently occurring p53 mutations, with the intent of restoring their WT functionality and providing new clinical tools for the treatment of a broad range of cancers.

8.4 Bibliography

1. Sjoblom, T.; Jones, S.; Wood, L. D.; Parsons, D. W.; Lin, J.; Barber, T. D.; Mandelker, D.; Leary, R. J.; Ptak, J.; Silliman, N.; Szabo, S.; Buckhaults, P.; Farrell, C.; Meeh, P.; Markowitz, S. D.; Willis, J.; Dawson, D.; Willson, J. K.; Gazdar, A. F.; Hartigan, J.; Wu, L.; Liu, C.; Parmigiani, G.; Park, B. H.; Bachman, K. E.; Papadopoulos, N.; Vogelstein, B.; Kinzler, K. W.; Velculescu, V. E., The consensus coding sequences of human breast and colorectal cancers. *Science* **2006,** *314* (5797), 268-74.

2. Suad, O.; Rozenberg, H.; Brosh, R.; Diskin-Posner, Y.; Kessler, N.; Shimon, L. J.; Frolow, F.; Liran, A.; Rotter, V.; Shakked, Z., Structural basis of restoring sequence-specific DNA binding and transactivation to mutant p53 by suppressor mutations. *J Mol Biol* **2009,** *385* (1), 249-65.

3. http://www-p53.iarc.fr/Statistics.html.

4. Hainaut, P.; Hernandez, T.; Robinson, A.; Rodriguez-Tome, P.; Flores, T.; Hollstein, M.; Harris, C. C.; Montesano, R., IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. *Nucleic Acids Res* **1998,** *26* (1), 205-13.

5. Ory, K.; Legros, Y.; Auguin, C.; Soussi, T., Analysis of the most representative tumour-derived p53 mutants reveals that changes in protein conformation are not correlated with loss of transactivation or inhibition of cell proliferation. *EMBO J* **1994,** *13* (15), 3496-504.

6. Cho, Y.; Gorina, S.; Jeffrey, P. D.; Pavletich, N. P., Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **1994,** *265* (5170), 346-55.

7. Maslon, M. M.; Hupp, T. R., Drug discovery and mutant p53. *Trends Cell Biol* **2010,** *20* (9), 542-55.

8. Sharpless, N. E.; DePinho, R. A., Cancer biology: gone but not forgotten. *Nature* **2007,** *445* (7128), 606-7.

9. Selivanova, G., Mutant p53: the loaded gun. *Curr Opin Investig Drugs* **2001,** *2* (8), 1136-41.

10. Sugikawa, E.; Hosoi, T.; Yazaki, N.; Gamanuma, M.; Nakanishi, N.; Ohashi, M., Mutant p53 mediated induction of cell cycle arrest and apoptosis at G1 phase by 9-hydroxyellipticine. *Anticancer Res* **1999,** *19* (4B), 3099-108.

11. Foster, B. A.; Coffey, H. A.; Morin, M. J.; Rastinejad, F., Pharmacological rescue of mutant p53 conformation and function. *Science* **1999,** *286* (5449), 2507-10.

12. Demma, M. J.; Wong, S.; Maxwell, E.; Dasmahapatra, B., CP-31398 restores DNA-binding activity to mutant p53 in vitro but does not affect p53 homologs p63 and p73. *J Biol Chem* **2004,** *279* (44), 45887-96.

13. Friedler, A.; Hansson, L. O.; Veprintsev, D. B.; Freund, S. M.; Rippin, T. M.; Nikolova, P. V.; Proctor, M. R.; Rudiger, S.; Fersht, A. R., A peptide that binds and stabilizes p53 core domain: chaperone strategy for rescue of oncogenic mutants. *Proc Natl Acad Sci U S A* **2002,** *99* (2), 937-42.

14. Bykov, V. J.; Issaeva, N.; Shilov, A.; Hultcrantz, M.; Pugacheva, E.; Chumakov, P.; Bergman, J.; Wiman, K. G.; Selivanova, G., Restoration of the tumor suppressor function to mutant p53 by a low-molecular-weight compound. *Nat Med* **2002,** *8* (3), 282-8.

15. Lambert, J. M.; Gorzov, P.; Veprintsev, D. B.; Soderqvist, M.; Segerback, D.; Bergman, J.; Fersht, A. R.; Hainaut, P.; Wiman, K. G.; Bykov, V. J., PRIMA-1 reactivates mutant p53 by covalent binding to the core domain. *Cancer Cell* **2009,** *15* (5), 376-88.

16. Bykov, V. J.; Issaeva, N.; Zache, N.; Shilov, A.; Hultcrantz, M.; Bergman, J.; Selivanova, G.; Wiman, K. G., Reactivation of mutant p53 and induction of apoptosis in human tumor cells by maleimide analogs. *J Biol Chem* **2005,** *280* (34), 30384-91.

17. Bullock, A. N.; Henckel, J.; DeDecker, B. S.; Johnson, C. M.; Nikolova, P. V.; Proctor, M. R.; Lane, D. P.; Fersht, A. R., Thermodynamic stability of wild-type and mutant p53 core domain. *Proc Natl Acad Sci U S A* **1997,** *94* (26), 14338-42.

18. Bullock, A. N.; Henckel, J.; Fersht, A. R., Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene* **2000,** *19* (10), 1245-56.

19. Tan, Y.; Luo, R., Structural and functional implications of p53 missense cancer mutations. *PMC Biophys* **2009,** *2* (1), 5.

20. Zhang, W.; Guo, X. Y.; Hu, G. Y.; Liu, W. B.; Shay, J. W.; Deisseroth, A. B., A temperature-sensitive mutant of human p53. *EMBO J* **1994,** *13* (11), 2535-44.

21. Ishimaru, D.; Andrade, L. R.; Teixeira, L. S.; Quesado, P. A.; Maiolino, L. M.; Lopez, P. M.; Cordeiro, Y.; Costa, L. T.; Heckl, W. M.; Weissmuller, G.; Foguel, D.; Silva, J. L., Fibrillar aggregates of the tumor suppressor p53 core domain. *Biochemistry* **2003,** *42* (30), 9022-7.

22. Ishimaru, D.; Maia, L. F.; Maiolino, L. M.; Quesado, P. A.; Lopez, P. C.; Almeida, F. C.; Valente, A. P.; Silva, J. L., Conversion of wild-type p53 core domain into a conformation that mimics a hot-spot mutant. *J Mol Biol* **2003,** *333* (2), 443-51.

23. Silva, J. L.; Vieira, T. C.; Gomes, M. P.; Bom, A. P.; Lima, L. M.; Freitas, M. S.; Ishimaru, D.; Cordeiro, Y.; Foguel, D., Ligand binding and hydration in protein misfolding: insights from studies of prion and p53 tumor suppressor proteins. *Acc Chem Res* **2010,** *43* (2), 271-9.

24. Bom, A. P.; Freitas, M. S.; Moreira, F. S.; Ferraz, D.; Sanches, D.; Gomes, A. M.; Valente, A. P.; Cordeiro, Y.; Silva, J. L., The p53 core domain is a molten globule at low pH: functional implications of a partially unfolded structure. *J Biol Chem* **2010,** *285* (4), 2857-66.

25. Ishimaru, D.; Ano Bom, A. P.; Lima, L. M.; Quesado, P. A.; Oyama, M. F.; de Moura Gallo, C. V.; Cordeiro, Y.; Silva, J. L., Cognate DNA stabilizes the tumor suppressor p53 and prevents misfolding and aggregation. *Biochemistry* **2009,** *48* (26), 6126-35.

26. Friedlander, P.; Legros, Y.; Soussi, T.; Prives, C., Regulation of mutant p53 temperature-sensitive DNA binding. *J Biol Chem* **1996,** *271* (41), 25468-78.

27. North, S.; Pluquet, O.; Maurici, D.; El-Ghissassi, F.; Hainaut, P., Restoration of wild-type conformation and activity of a temperature-sensitive mutant of p53 (p53(V272M)) by the cytoprotective aminothiol WR1065 in the esophageal cancer cell line TE-1. *Mol Carcinog* **2002,** *33* (3), 181-8.

28. Shiraishi, K.; Kato, S.; Han, S. Y.; Liu, W.; Otsuka, K.; Sakayori, M.; Ishida, T.; Takeda, M.; Kanamaru, R.; Ohuchi, N.; Ishioka, C., Isolation of temperature-sensitive p53 mutations from a comprehensive missense mutation library. *J Biol Chem* **2004,** *279* (1), 348-55.

29. Wong, K. B.; DeDecker, B. S.; Freund, S. M.; Proctor, M. R.; Bycroft, M.; Fersht, A. R., Hot-spot mutants of p53 core domain evince characteristic local structural changes. *Proc Natl Acad Sci U S A* **1999,** *96* (15), 8438-42.

30. Butler, J. S.; Loh, S. N., Structure, function, and aggregation of the zinc-free form of the p53 DNA binding domain. *Biochemistry* **2003,** *42* (8), 2396-403.

31. Benoit, V.; de Moraes, E.; Dar, N. A.; Taranchon, E.; Bours, V.; Hautefeuille, A.; Taniere, P.; Chariot, A.; Scoazec, J. Y.; de Moura Gallo, C. V.; Merville, M. P.; Hainaut, P., Transcriptional activation of cyclooxygenase-2 by tumor suppressor p53 requires nuclear factor-kappaB. *Oncogene* **2006,** *25* (42), 5708-18.

32. Noskov, S. Y.; Wright, J. D.; Lim, C., Long-range effects of mutating R248 to Q/W in the p53 core domain. *J Phys Chem B* **2001,** *106*, 13047-57.

33. Wright, J. D.; Lim, C., Mechanism of DNA-binding loss upon single-point mutation in p53. *J Biosci* **2007,** *32* (5), 827-39.

34. Wang, Y.; Rosengarth, A.; Luecke, H., Structure of the human p53 core domain in the absence of DNA. *Acta Crystallogr D Biol Crystallogr* **2007,** *63* (Pt 3), 276-81.

35. Xirodimas, D. P.; Lane, D. P., Molecular evolution of the thermosensitive PAb1620 epitope of human p53 by DNA shuffling. *J Biol Chem* **1999,** *274* (39), 28042-9.

36. Fogolari, F.; Moroni, E.; Wojciechowski, M.; Baginski, M.; Ragona, L.; Molinari, H., MM/PBSA analysis of molecular dynamics simulations of bovine beta-lactoglobulin: free energy gradients in conformational transitions? *Proteins* **2005,** *59* (1), 91-103.

37. Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M., Validation and use of the MM-PBSA approach for drug discovery. *J Med Chem* **2005,** *48* (12), 4040-8.

38. Barakat, K.; Mane, J.; Friesen, D.; Tuszynski, J., Ensemble-based virtual screening reveals dual-inhibitors for the p53-MDM2/MDMX interactions. *J Mol Graph Model* **2010,** *28* (6), 555-68.

39. Hainaut, P.; Butcher, S.; Milner, J., Temperature sensitivity for conformation is an intrinsic property of wild-type p53. *Br J Cancer* **1995,** *71* (2), 227-31.

40. Butler, J. S.; Loh, S. N., Folding and misfolding mechanisms of the p53 DNA binding domain at physiological temperature. *Protein Sci* **2006,** *15* (11), 2457-65.

41. Feki, A.; Irminger-Finger, I., Mutational spectrum of p53 mutations in primary breast and ovarian tumors. *Crit Rev Oncol Hematol* **2004,** *52* (2), 103-16.

Chapter 9: Summary And Future Work

This dissertation discussed the implementation of an improved virtual screening protocol and its application to discover inhibitors for a number of important cancer-related molecular targets. Two of these targets are DNA repair proteins that are related to the “drug resistance” phenomena. These are Excision Repair Cross-Complementation Group 1 (ERCC1), and DNA polymerase beta (pol β). The third target is p53, a guardian of the genome that is inactivated in more than half of all human cancers. The thesis also discussed the possibility of activating an otherwise inactive protein (R248Q p53 mutant).

The thesis started by a detailed introduction to the realm of virtual screening as a background for the coming chapters. Applications of these methods were then presented for: identifying inhibitors for the ERCC1-XPA interaction; identifying inhibitors for DNA polymerase beta; identifying dual inhibitors for the p53-MDM2/4 interactions; and probing the atomistic alternations that took place in the interaction of p53 to DNA due to the R248Q p53 mutation at different temperatures. Thus, this thesis investigated four interesting problems. Each problem has its own impact on the field of cancer research. Here, I will summarize the results we found and indicate the directions that should be followed in the future for every case.

9.1 ERCC1-XPA Inhibitors

Nucleotide excision repair (NER) is the major DNA repair mechanism that removes cisplatin-induced DNA damage, and that resistance to platinum-based therapy correlates with high expression of ERCC1, an essential element of the NER machinery.[1] Accordingly, a novel strategy to reverse resistance and potentiate the efficacy of cisplatin is to regulate the NER pathway, through targeting the interactions of ERCC1 with other proteins involved in NER. One solution is to inhibit the ERCC1-XPA protein-protein interactions. XPA plays a vital role in DNA lesion recognition and attraction of many other NER repair proteins. Its interaction with ERCC1 is necessary for a functional NER pathway.

This study utilized the RCS technique (section 3.3.3) to screen two compound databases for inhibitors of the ERCC1-XPA interaction and construct a pharmacophore model demonstrating the crucial features necessary for their inhibition. The databases included the National Cancer Institute Diversity Set (NCIDS) and DrugBank compounds (section 3.2.1). The study utilized a minimized model of the XPA binding site within ERCC1 to employ flexible residue docking as implemented in AutoDock 4.0. This was then followed by RCS docking, where MD simulations and RMSD conformational clustering were used to generate a set of forty-four representative conformations of the binding site within ERCC1. AutoDock was then used to screen against a set of seven target conformations, composed of the six most dominant cluster-representative structures along with an equilibrated folded conformation for the binding site produced by employing principal component analysis on the ERCC1 trajectory. Top hits were rescored by docking them to the whole set of cluster-representative structures and ranked by their weighted average binding energy (section 3.4.3). The non-redundant hits from these screens were then used to identify a dynamic binding-site pharmacophore that target the ERCC1-XPA interaction. The pharmacophore model was then compared to docking results for the weak inhibitor of NER, UCN-01 (7-hydroxystaurosporine) (section 4.1.4).

Comparing the methodology that was used here to the workflow discussed in the background material, one can make three observations. First, the virtual screening methodology depended mainly on docking scoring to rank the compounds. Second, the clustering analysis that was used to extract dominant conformations of the target were not iterative, it used a cut off RMSD value that is commonly employed in the literature. Finally, no post-docking refinements were performed on the final set of compounds. These shortcomings were properly adjusted in a subsequent study.[2]The new study screened CN chemical library (~100,000 compounds) (see section 3.2.1) and exactly followed the screening protocol described in this thesis. The hit rate of the new study was higher than that of the one described here, indicating the importance of utilizing more accurate scoring, performing iterative clustering and refining the docked structures using MD simulations.

Recently, we carried out a second VS round against the ERCC1-XPA interaction and following the procedure outlined in chapter 3 using the French national chemical library (section 3.2.1). A Promising hit was discovered and validated on a UV radiation sensitivity cell-based assay (compound 12 in Figure 9-1).[3] The validated hit is effective in sensitizing colon cancer cells to UV radiation, which induces the same type of damage as cisplatin and its lesions are removed by ENR. The compound is termed NER inhibitor 01 (NERI01) and its binding mode is shown in Figure.

Furthermore, an additional screening exercise was carried out targeting a different interaction related to ERCC1. This was the ERCC1-XPF interaction (Figure 9-3).[4]The full virtual screening methodology described in chapter 3, was used to screen the CN chemical library, NCI set and DrugBank compounds for inhibitors of this interaction. A number of promising hits were experimentally validated and were very effective in disrupting the NER pathway and potentiating cisplatin efficacy (Figure 9-4).

Future directions of this problem include identifying more novel inhibitors of the two different interactions, namely ERCC1-XPA and ERCC1-XPF, optimizing the discovered lead structures for better drug-like properties and advancing them through pre-clinical and clinical drug trails.

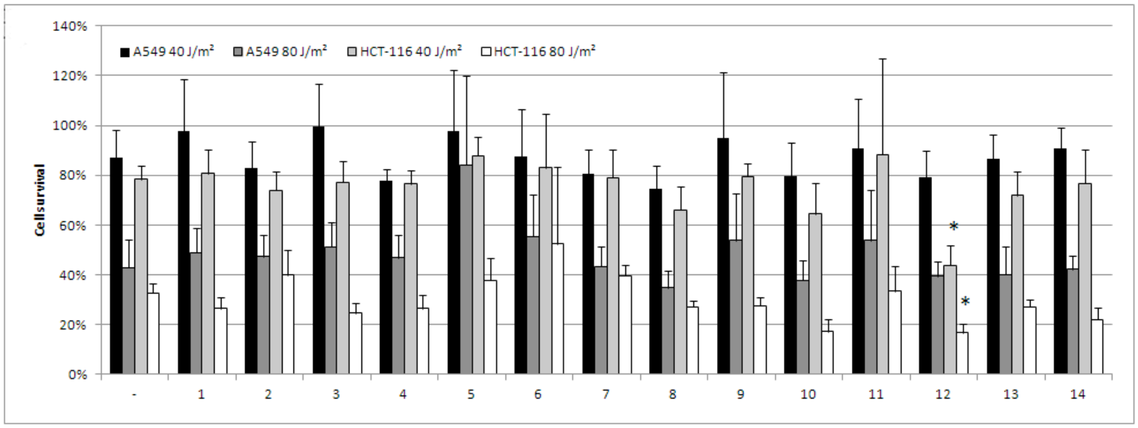


Figure 9-1. Sensitivity of cancer cells to UVC irradiation alone or in combination with potential inhibitors of the interaction between ERCC1 and XPA. IC50 values (J/m²). Compound 12 showed promising effect on cancer cells and was termed NERI01.

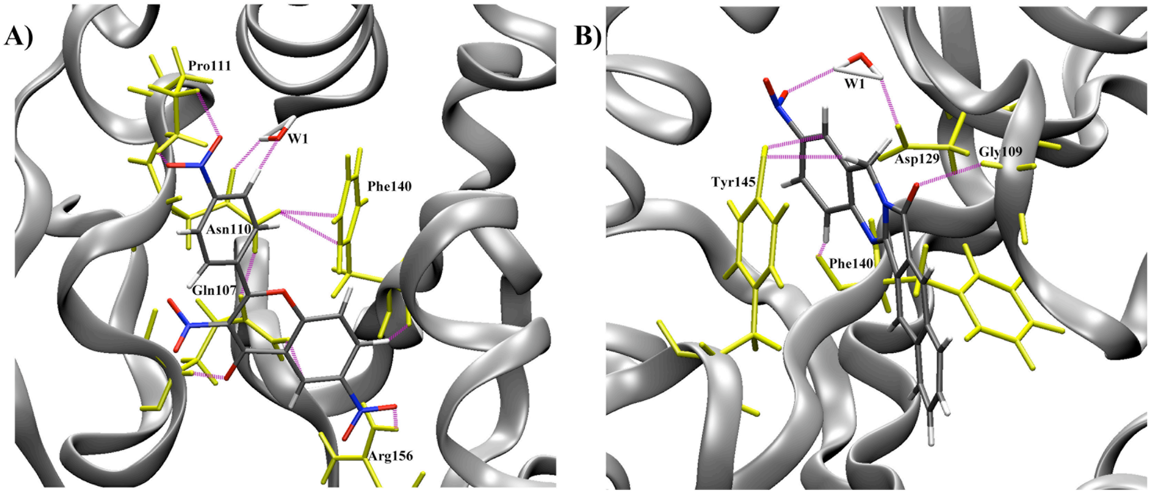


Figure 9-2. Binding modes and hydrogen bonding of the two selected hits (compounds 12 and 14 in Figure 9-1). Binding mode of NERI01 (A) and of 14 (B).

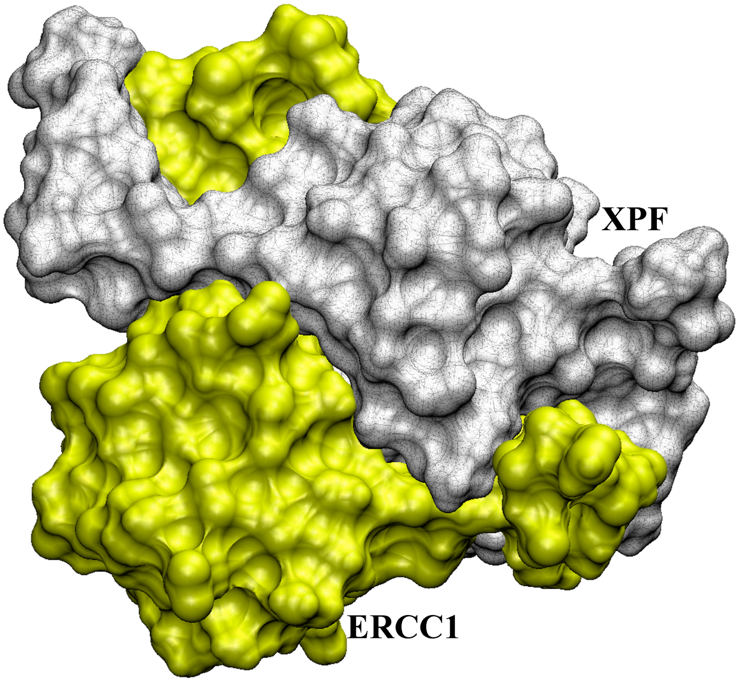


Figure 9-3. The ERCC1-XPF complex.

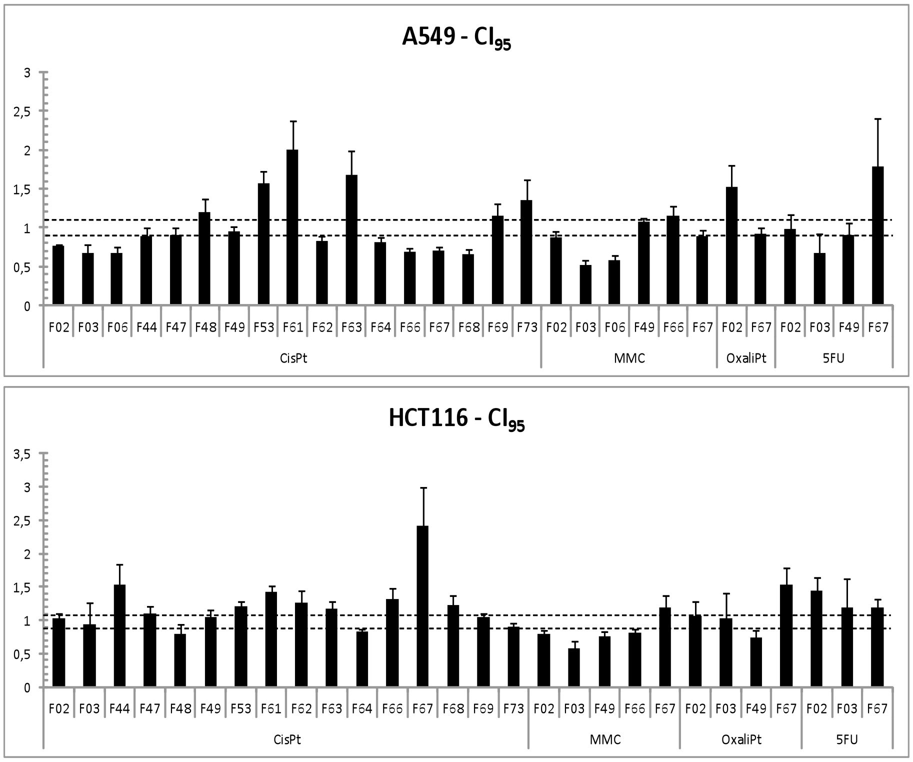


Figure 9-4. Combination index 95 of inhibitors and different anti-cancer drugs in A549 and HCT116 cells. Results are mean values from at least seven experiments with various ratios of compounds and error bars are standard error of means. Dotted horizontal lines indicate limits for synergy (<0.9), additivity (0.9 < CI95 < 1.1) and antagonism (>1.1).

9.2 DNA Pol Beta Inhibitors

Base Excision Repair (BER) is the major cellular pathway that is responsible for the recovery of single strand breaks (SSB) and removal of damaged bases such as oxidized-reduced, alkylated and deaminated bases.[5] However, it also constitutes a prevailing way that is usually adopted by cancer cells to reduce the efficacy of and to promote resistance against a growing list of DNA damaging agents including bleomycin,[6] monofunctional alkylating agents,[7] cisplatin[8] and other platinum-based compounds. Therefore, regulating this pathway has been proposed in the cancer research community as a way to reduce resistance to these DNA damaging agents. One way to do that is to inhibit its major DNA polymerase, pol β.

Chapters 5 and 6 focused on this problem. Chapter 5 provided a comprehensive literature review of existing inhibitors of pol β. While chapter 6 discussed the application of virtual screening to find inhibitors for the lyase active site of pol β.

Following the same virtual screening procedure discussed earlier, this study focused the search space on the binding site of PA (a well-validated pol β-inhibitor) using it as a positive control. The aim was to discover more potent drug candidates through filtering a library of ~12,500 structures. The molecules we tested included the NCI diversity set, the DrugBank set of small-molecules and more than 9,000 fragment structures with drug-like properties extracted from ZINC database. The top 300 hits that showed strong affinity for pol β have been rescored using a more robust scoring function, the MM-PBSA method.

Similar to ERCC-XPA case and following the NER study, a future direction of this exercise is to validate the identified hits experimentally. Once active compounds are identified, we will start developing derivative structures for the identified hits and optimize them for better drug-like properties.

9.3 P53-MDM2/4 inhibitors

P53 play vital roles in cell cycle, apoptosis, DNA repair and senescence[9,10,11,12] As such, tumor cells have developed numerous ways to disable its function. In the about 50% of human cancers, although p53 retains its wild type structure, its activity is eradicated by its main cellular inhibitors, murine double minute 2/4 (MDM2/4).[13,14,15]MDM2 and MDM4 are two structurally related proteins that regulate p53 activity.[15] the two proteins are over expressed in many types of cancer, reducing the activity of p53 and allowing cancer cells to survive and polifirate.[16] The last decade has witnessed the identification of an increasing number of non-peptide, small-molecule MDM2 inhibitors with promising binding affinities.[17] However, these compounds are more highly selective for MDM2 than for its homolog MDM4, meaning that they have no effect on MDM4. Therefore, developing new compounds that are MDM4-specific or optimized for dual-inhibition of MDM2 and MDM4 is a necessary step to achieve full activation of p53 in tumor cells.

This study screened the NCI diversity set, the DrugBank set of small-molecules and more than 3,168 derivative structures extracted from the known MDM2- inhibitors against twenty-eight different MDM2 models that represent the apo- and holo-structure's collective conformational dynamics. The top 300 hits that showed strong affinity for MDM2 have been used in a second round of screening against the p53 binding-site within MDM4 depicts the basic strategy that was followed in this work. This procedure identified dual-inhibitors that are predicted to disrupt the MDM2/MDM4- p53 interaction and allow for the full activation of the p53 pathway.

The future directions of this study would follow two independent but complementary paths. First is to experimentally validate the predicted hits and focus the search space on the compounds that show biological activity. These compounds will be used to enrich the screened ligand collection by creating derivatives of their structure and use similarity search methods to locate similar compounds in available databases. Second, is to follow the same path that was used in the previous NER exercise and screen large compounds libraries such as the CN chemical library or Zinc database.

9.4 Toward R248Q p53 activation

This study was a continuation to the research problem discussed in the previous chapter. As p53 is the most mutated protein in human cancers,[18] and mutations of p53 alone account for more than half of invasive types of cancer,[19] a simple idea to cure these types of cancer is to reactivate the mutated p53 variants. Therefore, this study investigated the key structural changes that were induced on the wild type p53 due to one of its most frequent mutations, R248Q. Upon understanding of these changes, one can suggest a mechanism by which p53 returns to its active state. This study was inspired by two facts. First, there are a few successful stories that managed to activate mutant p53 and cure certain types of cancer. Second, the R248 mutant along with some other hot spot mutant variants has a very interesting property. They possess temperature-dependence on their binding to the DNA. That is, at temperature below the physiological range they act similarly to the wild type protein, however, at physiological temperatures they lose their DNA binding capacity. Consequently, the current study analyzed the association of p53 both in its mutated and wild type states to the DNA at different temperature ranges.

The study identified those changes that took place in the hydrogen bond network between the DNA and the core domain of p53. There was a direct correlation between the loss of hydrogen bonds at the R248 site at physiological temperatures and the reduction in the binding energy between p53 and the DNA molecule. This hydrogen bonding was preserved at low temperatures but was lost at higher temperatures. This loss induced local structural changes around the mutated residue. An activator of p53 should bind close to the mutated R248 and try to restore the hydrogen bonding interactions that were lost due to mutation.

A future direction for this study is to design a set of small molecules that can establish a few hydrogen bonds between the DNA and the mutated protein structure. These structures will be docked to the cavity between the core domain and the DNA close to the R248 location and designed so that the hydrogen bonding would be maintained at physiological temperatures.

This concludes the work described in this thesis and I hope it would benefit humanity by using it to improve cancer therapy.

9.5 Bibliography

1. (a) Metzger, R.; Leichman, C. G.; Danenberg, K. D.; Danenberg, P. V.; Lenz, H. J.; Hayashi, K.; Groshen, S.; Salonga, D.; Cohen, H.; Laine, L.; Crookes, P.; Silberman, H.; Baranda, J.; Konda, B.; Leichman, L., ERCC1 mRNA levels complement thymidylate synthase mRNA levels in predicting response and survival for gastric cancer patients receiving combination cisplatin and fluorouracil chemotherapy. *J Clin Oncol* **1998,** *16* (1), 309-16; (b) Handra-Luca, A.; Hernandez, J.; Mountzios, G.; Taranchon, E.; Lacau-St-Guily, J.; Soria, J. C.; Fouret, P., Excision repair cross complementation group 1 immunohistochemical expression predicts objective response and cancer-specific survival in patients treated by Cisplatin-based induction chemotherapy for locally advanced head and neck squamous cell carcinoma. *Clin Cancer Res* **2007,** *13* (13), 3855-9; (c) Bellmunt, J.; Paz-Ares, L.; Cuello, M.; Cecere, F. L.; Albiol, S.; Guillem, V.; Gallardo, E.; Carles, J.; Mendez, P.; de la Cruz, J. J.; Taron, M.; Rosell, R.; Baselga, J., Gene expression of ERCC1 as a novel prognostic marker in advanced bladder cancer patients receiving cisplatin-based chemotherapy. *Ann Oncol* **2007,** *18* (3), 522-8; (d) Jun, H. J.; Ahn, M. J.; Kim, H. S.; Yi, S. Y.; Han, J.; Lee, S. K.; Ahn, Y. C.; Jeong, H. S.; Son, Y. I.; Baek, J. H.; Park, K., ERCC1 expression as a predictive marker of squamous cell carcinoma of the head and neck treated with cisplatin-based concurrent chemoradiation. *Br J Cancer* **2008,** *99* (1), 167-72.
2. K. Barakat; L. Jordheim; C. Dumontet; Tuszynski., J., Virtual screening and biological evaluation of inhibitors targeting the XPA-ERCC1 interaction. *Molecular Cancer Therapeutics* **Submitted Jan 2012**.
3. Barakat, K. H.; Jordheim, L. P.; Dumonte, C.; Tuszynski, J., Virtual screening and biological evaluation of inhibitors targeting the XPA-ERCC1 interaction. *Submitted to PLoS ONE* **2012**.
4. L. Jordheim; K. Barakat; C. Dumontet; Tuszynski., J., Discovering the first inhibitor of the XPF-ERCC1 interaction. **In progress**
5. Xu, G.; Herzig, M.; Rotrekl, V.; Walter, C. A., Base excision repair, aging and health span. *Mech Ageing Dev* **2008,** *129* (7-8), 366-82.
6. Parsons, J. L.; Dianova, II; Dianov, G. L., APE1 is the major 3'-phosphoglycolate activity in human cell extracts. *Nucleic Acids Res* **2004,** *32* (12), 3531-6.
7. Liu, L.; Nakatsuru, Y.; Gerson, S. L., Base excision repair as a therapeutic target in colon cancer. *Clin Cancer Res* **2002,** *8* (9), 2985-91.
8. Hoffmann, J. S.; Pillaire, M. J.; Garcia-Estefania, D.; Lapalu, S.; Villani, G., In vitro bypass replication of the cisplatin-d(GpG) lesion by calf thymus DNA polymerase beta and human immunodeficiency virus type I reverse transcriptase is highly mutagenic. *J Biol Chem* **1996,** *271* (26), 15386-92.
9. Teodoro, J. G.; Evans, S. K.; Green, M. R., Inhibition of tumor angiogenesis by p53: a new role for the guardian of the genome. *J Mol Med* **2007,** *85* (11), 1175-86.
10. Fridman, J. S.; Lowe, S. W., Control of apoptosis by p53. *Oncogene* **2003,** *22* (56), 9030-40.
11. Vousden, K. H.; Lu, X., Live or let die: the cell's response to p53. *Nat Rev Cancer* **2002,** *2* (8), 594-604.
12. Bourdon, J. C.; Laurenzi, V. D.; Melino, G.; Lane, D., p53: 25 years of research and more questions to answer. *Cell Death Differ* **2003,** *10* (4), 397-9.
13. Kubbutat, M. H.; Jones, S. N.; Vousden, K. H., Regulation of p53 stability by Mdm2. *Nature* **1997,** *387* (6630), 299-303.
14. Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. P., Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **1996,** *274* (5289), 948-53.
15. Shvarts, A.; Steegenga, W. T.; Riteco, N.; van Laar, T.; Dekker, P.; Bazuine, M.; van Ham, R. C.; van der Houven van Oordt, W.; Hateboer, G.; van der Eb, A. J.; Jochemsen, A. G., MDMX: a novel p53-binding protein with some functional properties of MDM2. *EMBO J* **1996,** *15* (19), 5349-57.
16. Toledo, F.; Wahl, G. M., Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer* **2006,** *6* (12), 909-23.
17. Patel, S.; Player, M. R., Small-molecule inhibitors of the p53-HDM2 interaction for the treatment of cancer. *Expert Opin Investig Drugs* **2008,** *17* (12), 1865-82.
18. Sjoblom, T.; Jones, S.; Wood, L. D.; Parsons, D. W.; Lin, J.; Barber, T. D.; Mandelker, D.; Leary, R. J.; Ptak, J.; Silliman, N.; Szabo, S.; Buckhaults, P.; Farrell, C.; Meeh, P.; Markowitz, S. D.; Willis, J.;

Dawson, D.; Willson, J. K.; Gazdar, A. F.; Hartigan, J.; Wu, L.; Liu, C.; Parmigiani, G.; Park, B. H.; Bachman, K. E.; Papadopoulos, N.; Vogelstein, B.; Kinzler, K. W.; Velculescu, V. E., The consensus coding sequences of human breast and colorectal cancers. *Science* **2006,** *314* (5797), 268-74.

19. Suad, O.; Rozenberg, H.; Brosh, R.; Diskin-Posner, Y.; Kessler, N.; Shimon, L. J.; Frolow, F.; Liran, A.; Rotter, V.; Shakked, Z., Structural basis of restoring sequence-specific DNA binding and transactivation to mutant p53 by suppressor mutations. *J Mol Biol* **2009,** *385* (1), 249-65.

Appendix A: Principal Component Analysis (PCA)

The covariance matrix of atomic coordinates is defined by:

$$\sigma_{ij} = \left\langle M_{ij}^{\frac{1}{2}}\left(r_i - \langle r_i \rangle\right) M_{ij}^{\frac{1}{2}}\left(r_j - \langle r_j \rangle\right) \right\rangle \qquad \text{Equ. A.1}$$

Where *M* is a diagonal matrix that has its diagonals are equal to the atomic masses of the N atoms or are equal to ones.

$$M = \begin{bmatrix} m_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & m_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & m_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & m_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & m_{N-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & m_N & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & m_N & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & m_N \end{bmatrix} \qquad \text{Equ. A.2}$$

σ is a symmetric matrix and can be diagonalized with an orthogonal transformation matrix *R*.

$$R^T \sigma R = diag(\lambda_1, \lambda_2, \lambda_3, ., ., ., ., ., \lambda_{3N}) \text{ where } \lambda_1 \geq \lambda_2 \geq .... \geq \lambda_{3N} \qquad \text{Equ. A.3}$$

The columns of *R* are the eigenvectors of σ, while λ_i are the eigenvalues of the covariance matrix. The principle components of the system ($p_i(t)$) can be calculated by projecting the actual MD trajectory onto the eigenvectors of the covariance matrix using the transformation:

$$\mathrm{p}(t) = R^T M^{\frac{1}{2}}\left(x(t) - \langle x \rangle\right) \qquad \text{Equ. A.4}$$

The MD trajectory can also be filtered along one or more principle component. For example, for filtering on one principle component:

$$\mathrm{X}^{filtered} = \langle \mathrm{X} \rangle + M^{-\frac{1}{2}} R_{*i} p_i(t) \qquad \text{Equ. A.5}$$

Since for a typical symmetric matrix *A*, one can calculate its square root as follows:

$$A^{\frac{1}{2}} = Rdiag\left(\lambda_1^{\frac{1}{2}},\lambda_2^{\frac{1}{2}},\lambda_3^{\frac{1}{2}},.....,\lambda_{3N}^{\frac{1}{2}}\right)R^T \qquad \text{Equ. A.6}$$

Now, define the difference between two matrices *A* and *B* as:

$$d(A,B) = \sqrt{\mathrm{tr}\left(\left(A^{\frac{1}{2}} - B^{\frac{1}{2}}\right)^2\right)} = \sqrt{\mathrm{tr}\left(A + B - 2A^{\frac{1}{2}}B^{\frac{1}{2}}\right)} \qquad \text{Equ. A.7}$$

Where tr is the trace of the matrix. This gives:

$$d(A,B) = \left(\sum_{i=1}^{N}\left(\lambda_i^A + \lambda_i^B\right) - 2\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\lambda_i^A\lambda_i^B}\left(R_i^A \cdot R_j^B\right)\right)^{\frac{1}{2}} \qquad \text{Equ. A.8}$$

Now, define the overlap between the two matrices as:

$$\text{Normalized overlap} = s(A,B) = 1 - \frac{d(A,B)}{\sqrt{\mathrm{tr}A + \mathrm{tr}B}} \qquad \text{Equ. A.9}$$

The overlap is one if and only if the two matrices are identical and is 0 if they are orthogonal.

Appendix B: Linearized Poisson-Boltzmann Equation (LPBE)

Poisson equation states that:

$$-\nabla \cdot \varepsilon(r)\nabla\psi(r) = 4\pi e^2 \rho(r)$$

Equ. B.1

where e is the elementary charge, ρ is the charge distribution, ψ is the potential to be solved, ε is a position dependant dielectric constant ($\varepsilon = 2$ inside the protein and $\varepsilon = 80$ outside), k is Boltzmann constant and T is the temperature.

For ionic solutions, the Poisson Boltzmann equation is used:

$$-\nabla \cdot \varepsilon(r)\nabla\psi(r) = 4\pi e^2 \rho(r) - \sum_{i}^{n} q_i c_i e^{\frac{-q_i\psi(r)-\phi_i(r)}{kT}}$$

Equ. B.2

where q is the ionic charge, c is the ionic concentration, ϕ is steric interaction with the fixed solute, and n is the number of ions within the solution.

A typical MD simulation comprises two types of ions with equal concentration but opposite charges (e.g Na$^+$ and Cl$^-$), meaning that:

$$-q\psi(r) - \phi(r) << kT \quad \text{and} \quad q = q_1 = q_2; c = c_1 = c_2$$

Equ. B.3

and leading to the approximation:

$$\sum_{i=1}^{2} qce^{\frac{-q\psi(r)-\phi(r)}{kT}} = qc\sinh\left(\frac{-q\psi(r)-\phi(r)}{kT}\right) \approx qc\left(\frac{-q\psi(r)-\phi(r)}{kT}\right)$$

Equ. B.4

Substituting into Equ. B.2 leads to the linearized Poisson-Boltzmann equation (LPBE):

$$\nabla \cdot \left[\varepsilon(r)\nabla\psi(r)\right] = -4\pi e^2 \rho(r) + qc\,\frac{q\psi(r) + V(r)}{kT}$$

Equ. B.6

The equation is commonly solved numerically using a multigrid finite difference approach.

Appendix C: Methods For Chapter 4

Chapter 2 introduced in details most of the methods described below. However, here I will describe the specific parameters that were used and the precise workflow that was applied to the ERCC1-XPA case.

C.1. Molecular Dynamics Simulations

The central domain of ERCC1 (residues 99-214), both free and bound to an 11-residue fragment of XPA (residues 67-77) was taken from PDB entry 2JNW.[1] Molecular Dynamics (MD) simulations were carried out using the NAMD program,[2] at a mean temperature of 300K and physiological pH (pH 7) using the all-hydrogen AMBER99SB force field.[3] Protonation states of all ionizable residues were calculated using the program PDB2PQR. Following parameterization, the ERCC1 protein alone or in complex with the XPA peptide was immersed in the center of a TIP3P water cube after adding hydrogen atoms to the initial protein structure. The cube dimensions were chosen to provide at least a 20-Å buffer of 16596 (15323) water molecules around the systems. To neutralize and prepare the XPA-bound or (free systems) under a physiological ionic concentration, 32 (29) chloride and 30 (27) sodium ions were respectively added by replacing water molecules having the highest electrostatic energies on their oxygen atoms. The fully solvated protein was then minimized and subsequently heated to the simulation temperature with heavy restraints placed on all backbone atoms. Following heating, the system was equilibrated using periodic boundary conditions for 100 ps and energy restraints reduced to zero in successive steps of the MD simulation. The simulations were then continued for 50 ns during which atomic coordinates were saved to the trajectory every 2 ps. The root-mean-square deviation (RMSD) and B-factors for the protein backbone were then computed over the last 10 ns of the MD simulation using the PTRAJ utility within AMBER10. Hydrogen bond analysis was performed by computing the average distance between donor and acceptor atoms. A hydrogen bond was defined by a heavy donor – heavy acceptor distance greater than 3.4 Å, a light donor-heavy acceptor distance greater than 2.5 Å, and a deviation of less than 60° from linearity.

C.2. Principal Component Analysis

Following the procedure described in section 3.3.4, PCA was performed on the entire MD trajectory. First, the trajectory was RMSD fitted to a reference structure, in order to remove all rotations and translations. The covariance matrix was then calculated from their Cartesian co-ordinates as in EQ. 3-1. The eigenvectors of the covariance matrix constituted the essential vectors of the motion. Convergence of sampling was predicted using the normalized overlap method (EQ. 3-2). That is, the MD trajectory was divided into three parts and the normalized overlap between each pair was calculated to determine the completeness of sampling (see results).

C.3. Extracting Representative Structures

To generate a reduced set of representative models of the ERCC1 binding site, we performed root-mean-square difference (RMSD) conformational clustering with the average-linkage algorithm as implemented in the PTRAJ utility of AMBER10 using a critical distance of 1.3 Å. For the apo-ERCC1 simulation, structures were extracted at 2 ps intervals over the entire 50 ns simulation. All C_{α} atoms were RMSD fitted to the minimized initial structure in order to remove overall rotation and translation. RMSD-clustering was performed on the 22 residues that line the XPA binding site, namely those numbered: 106 - 112, 129, 140-146, 148, 149, 152, 153, 156, 172, and 174. These residues were clustered into groups of similar conformations using the atom-positional RMSD of the entire amino acid, including side chains and hydrogen atoms, as the similarity criterion. The cutoff was chosen after evaluation of the dependence of cluster populations against the total number of clusters using a range of 0.9-1.4 Å. Forty-four clusters were obtained and the six most dominant clusters represented approximately 48%, 8%, 6%, 5.5%, 4% and 3.8% of the whole ensemble, respectively. The centroid of each cluster, the structure having the smallest RMSD to all members of the cluster, was chosen as the cluster representative structure and was used as rigid template for docking experiment.

C.4. Equilibrated ERCC1 Model

A detailed representation of the conformational dynamics can be obtained by projecting the trajectory onto the planes spanned by the most dominant eigenvectors of the covariance matrix. The higher the occupancy of a conformational state in this projection, the lower the free energy of that state.[4,5] Therefore, by observing the regions at which many conformations cluster, one can predict the minimal energy conformations visited by an MD trajectory and estimate a representative conformation for these structures. The entire MD trajectory was projected onto the planes spanned by the first and second, the first and third and the second and third principal components (see results). The conformations residing within the global minimum region were used to predict an equilibrated binding site template. The equilibrated model was compared to the most dominant cluster representative structure and has been appended to the set of conformations used in VS experiments.

C.5. Energy Evaluation of the ERCC1/XPA Interaction

The trajectory of the ERCC1/XPA MD simulation was analyzed using the MMPBSA utility of AMBER10 to calculate the individual binding energies between residues within the XPA peptide and the ERCC binding site. The binding energy was further divided into individual residue contributions to recognize the key residues in the interaction between the two proteins. Following the identification of the significant residues, we carried out alanine scanning on these residues by performing MD simulation on substituted models and calculating the resultant binding energies. Consequently, residues essential for the interaction between ERCC1 and XPA were determined for subsequent flexible docking.

C.5.1. Binding Free Energy

Binding free energies between the ERCC1 and XPA peptides were calculated using the MM–PBSA method as implemented in AMBER10.[48] Following the equations described in section 3.7, binding energy analysis and energy decomposition for each snapshot was calculated using the SANDER module of AMBER10. The binding free energy between the XPA67-77 and the ERCC199-214 binding site can be approximated by:

$$\Delta G^{o} = \Delta G_{gas}^{ERCC1\bullet XPA} + \Delta G_{solv}^{ERCC1\bullet XPA} - \{\Delta G_{solv}^{ERCC1} + \Delta G_{solv}^{XPA}\} \qquad \text{EQ. C-1}$$

C.6. Selection of Ligand Database

For this study, the VCC was comprised of two main libraries, namely, the National Cancer Institute Diversity Set (NCDIS) and DrugBank-small-molecules. The two databases are described in chapter 3 (section 3.2.1). In addition, UCN-01, a compound that has been previously demonstrated to weakly inhibit the ERCC1-XPA interaction (see above), was used as a comparison during the screening experiments and for subsequent pharmacophore validation.[6]

C.7. Ligand Screening

Virtual screening on the ERCC1 binding site was performed using AutoDock, version 4.0. Hydrogen atoms were added to ERCC1 and ligands and partial atomic charges were then assigned using the Gasteiger-Marsili method.[7] Atomic solvation parameters were assigned to the protein atoms using the AUTODOCK 4.0 utility ADDSOL. A docking grid map with 70x70x70 points and grid point spacing of 0.375 Å was then centered on the XPA binding site within the ERCC1 receptor using AUTOGRID4.0 program. Rotatable bonds of each ligand were then automatically assigned using AUTOTORS utility of

AUTODOCK4.0. Docking was performed using the Lamarckian Genetic Algorithm (LGA) method with an initial population of 150 random individuals; a maximum number of energy evaluations; 100 trials; 27,000 maximum generations; a mutation rate of 0.02; a crossover rate of 0.80 and the requirement that only one individual can survive into the next generation. A total of eight independent virtual screening runs were performed against the full set of docked ligands. The first used the minimized holo crystal structure of the ERCC1-XPA binding site with key residues, determined from previous MD experiments, set as flexible during the docking experiment. The other seven experiments utilized the six representative conformations of the dominant clusters produced from the clustering analysis along with the equilibrated model of the ERCC1 binding site as determined using principal component projections.

C.8. Clustering of Docked Poses

Following the procedure described in section 3.4.2, all docking results were iteratively clustered in order to identify the optimal number of clusters and their best clustering pattern (see results).

C.9. RCS and Receptor Flexibility

As described in chapter 3 (section 3.3), the relaxed complex scheme (RCS) is a hybrid technique that combines the rewards of docking algorithms with dynamic structural information provided by molecular dynamics (MD) simulations, therefore explicitly accounting for the flexibility of both the receptor and the docked ligands. To apply the RCS approach to the ERCC10-XPa case, we performed 7 independent screening runs against rigid templates of the binding site within ERCC1. This set of conformations comprised of the central member cluster structures of the 6 dominant clusters that constitute about 75% of the whole MD trajectory along with the equilibrated model produced from PCA. Docking results were sorted by the lowest binding energy of the most populated cluster using the proposed ligand clustering technique. We only considered a compound among the top hits if the most populated cluster includes at least 25% of all docked conformations. The top 50 hits from each system were combined to produce an irredundant set of promising compounds. To validate and refine the virtual screening results, re-docking experiments were performed on the combined hits into the rest of the 44 clustering representative structures, to accounts for 100% of the ensemble of the apo MD trajectory. Following the same docking procedure and parameter set described in the previous sections, docking poses were ranked using their weighted average binding energies (EQ. 3-4) and were used for further analysis.

C.10. Electrostatic Surface Calculations

Electrostatic potentials for ERCC1 were calculated using the APBS program[8] and mapped onto a reduced molecular surface with the VMD visualization program.[9] ERCC1 was treated as a low dielectric medium surrounded by a high dielectric solvent (for water). The ionic strength was set to 0.1 M. The low-dielectric region of the protein was defined as the region inaccessible to contact by a 1.4 Å sphere rolling over the molecular surface, defined by atomic co-ordinates of the MD structure and vdW radii taken from the all-hydrogen AMBER99SB force field. The electrostatic potential calculations employed a 200x200x200 grid with a spacing of 0.5 Å.

C.11. ERCC1-XPA Binding Site Pharmacophore

To produce a final pharmacophore (section 2.2.2) for the XPA ERCC1 binding interaction, we used Accelrys Discovery Studio 2.1 (Accelrys Inc., 2008) to construct models for the top 30 poses collected from each of the top seven screening simulations. Pharmacophore generation involved using the catalyst/HipHop program to generate feature-based 3D pharmacophore alignments.[10] This was accomplished by examining each separate pose for the presence of certain chemical features, followed by the determination of a three-dimensional configuration of the chemical features. Catalyst provides a predefined dictionary of chemical features found to be important in drug-enzyme/receptor interactions. These are hydrogen bond donors, hydrogen bond acceptors, hydrophobic group, ring aromatic and positive/negative ionizable groups. For the pharmacophore modeling runs, common features selected for the run were ring aromatic, hydrogen bond donor and acceptor, hydrophobic group and ionizable groups. Since we do not have access to any activity data for any of the compounds being screened, we adopted a strategy, where HipHop assumes that differences in activities will be related to the differences in other relevant factors like conformational energies, but not due to the absence of any important features required for binding. Merging and overlaying of each of the resulting pharmacophores was then accomplished using the clique detection algorithm combined with the Kabsch alignment approach.[11] However, due to the vast structural dissimilarity of each of the ligands obtained from the docking stage, chemical features associated with each residue found within the ERCC1 binding site were tabulated and a pattern of interaction was then determined manually.

C.12. Bibliography

1. Tsodikov, O. V.; Ivanov, D.; Orelli, B.; Staresincic, L.; Shoshani, I.; Oberman, R.; Scharer, O. D.; Wagner, G.; Ellenberger, T., Structural basis for the recruitment of ERCC1-XPF to nucleotide excision repair complexes by XPA. *EMBO J* **2007,** *26* (22), 4768-76.

2. Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K., NAMD2: Greater Scalability for Parallel Molecular Dynamics. *Journal of Computational Physics* **1999,** *151* (1), 283-312.

3. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-25.

4. Grubmuller, H., Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **1995,** *52* (3), 2893-2906.

5. Kosztin, I.; Barz, B.; Janosi, L., Calculating potentials of mean force and diffusion coefficients from nonequilibrium processes without Jarzynski's equality. *J Chem Phys* **2006,** *124* (6), 64106.

6. Jiang, H.; Yang, L. Y., Cell cycle checkpoint abrogator UCN-01 inhibits DNA repair: association with attenuation of the interaction of XPA and ERCC1 nucleotide excision repair proteins. *Cancer Res* **1999,** *59* (18), 4529-34.

7. Gasteiger J; Marsili M, Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* **1980,** *36*, 3219.

8. Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A., Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **2001,** *98* (18), 10037-41.

9. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1995,** *14* (1), 33-38.

10. PW, S., Automated Chemical Hypothesis Generation and Database Searching with Catalyst. In: Müller K, editor. Perspectives in Drug Discovery and Design. . *ESCOM Science Publishers B.* **1995,** *3*, 1-20.

11. Kabsch, W., A solution for the best rotation to relate to sets of vectors. *Acta Crystallogr* **1976,** *A32*, 922-923.

Appendix D: Methods For Chapter 6

D.1. Molecular Dynamics Simulations

The 8-kDa domain of polβ (residues 1-87) was taken from the PDB entry 1DK3[1]. MD simulations were carried out using the NAMD program,[2] at a mean temperature of 310K and physiological pH (pH 7) using the all-hydrogen AMBER99SB force field.[3] Protonation states of all ionizable residues were calculated using the program PDB2PQR.[4] Following parameterization, the protein was immersed in the center of a TIP3P water cube after adding hydrogen atoms to the initial protein structure. The cube dimensions were chosen to provide at least a 20Å-wide buffer of 17605 water molecules around the systems. To neutralize and prepare the simulated system under a physiological ionic concentration, 41 chloride and 32 sodium ions were respectively added by replacing water molecules having the highest electrostatic energies on their oxygen atoms. The number of counter ions for each case was calculated by first estimating the amount of ions that is needed to set up the solvated system under normal physiological conditions (pH 7), followed by adding the number of chloride ions required to bring its net charge to zero. The fully solvated protein was then minimized and subsequently heated to the simulation temperature with heavy restraints placed on all backbone atoms. Following heating, the system was equilibrated using periodic boundary conditions for 100 ps and energy restraints reduced to zero in successive steps of the MD simulation. The simulations were then continued for 92 ns during which atomic coordinates were saved to the trajectory every 2 ps. The total simulation time was determined by visualizing the quality of sampling as predicted by PCA (see below). The RMSD (data not shown) and B-factors (see Figure 3) for the protein backbone were then computed over the last 10 ns of the MD simulation using the PTRAJ utility within AMBER10.[5] Hydrogen bond analyses were performed by computing the average distance between donor and acceptor atoms. A hydrogen bond was defined by a heavy donor – heavy acceptor distance < 3.4 Å, a light donor-heavy acceptor distance < 2.5 Å, and a deviation of less than $\pm 60^{o}$ from linearity.

Following the MD protocol mentioned above we prepared 300 MD simulations for each top hit that resulted from the ensemble-based screening (see Results). Parameters for ligands were assigned using the generalized AMBER force field[6] and partial charges were calculated with the AM1-BCC method[7] using Antechamber in the AMBER 10 package. Following parameterization, the protein/ligand complexes were subjected to MD simulations for a production phase of 2 ns. Snapshots were extracted every 2 ps and used for the MM-PBSA binding energy analysis.

D.2. Clustering Analysis

To generate a reduced set of representative polβ models, we performed RMSD conformational clustering with the average-linkage algorithm as implemented in the PTRAJ utility of AMBER10 using cluster counts ranging from 5 to 150 clusters. Structures were extracted at 2 ps intervals over the entire simulation times. All C_α-atoms were RMSD fitted to the minimized initial structure in order to remove overall rotation and translation. RMSD-clustering was performed on the 21 residues contained in the PA (DNA) binding site (residues numbered: 30, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 43, 63, 64, 65, 66, 67, 68, 69 and 70). These residues were clustered into groups of similar conformations using the atom-positional RMSD of the entire amino acid, including side chains and hydrogen atoms, as the similarity criterion. The optimal numbers of clusters were chosen after evaluation of the two clustering metrics, described in section 3.3.5, for different cluster counts (see Results). A total of 45 clusters were extracted from the trajectory. The centroid of each cluster, the structure having the smallest RMSD to all members of the cluster, was chosen as the cluster representative structure and the most dominant structures were used as rigid templates for the ensemble-based docking experiments (see Results).

D.3. Principal Component Analysis

PCA was performed according to the same concepts and procedure described in chapter 3 (see section 3.3.4)

D.4. Selection of Ligand Database

The National Cancer Institute Diversity Set (NCDIS),[8] DrugBank-small-molecules,[9] and a set of 9,135 fragment structures were used as our test libraries of compounds. For more information about the NCI and DrugBank compound libraries see section 3.2.1. In addition, we included a set of 9,135 clean-fragments compounds, downloaded from the ZINC database. These fragments have a Tanimoto coefficient of 70% (see section 2.3.1), molecular weight lower than 250 Da, xlogP lower than 2.5, number of rotatable bonds less than 5 and only a single stereoisomer and protonation state for each compound.

D.5. Ligand Screening

Virtual screening on the PA binding site (which coincides with the DNA binding site)[10] within polβ was performed using AutoDock, version 4.0.[11] Hydrogen atoms were added to the protein and ligands and partial atomic charges were then assigned using the Gasteiger-Marsili method.[12] Atomic solvation parameters were assigned to the atoms of the protein using the AutoDock 4.0 utility ADDSOL. Docking grid maps with 118x98x104 points and grid point spacing of 0.26 Å was then centered on the PA binding site within polβ using AutoGrid4.0 program.[11] Rotatable bonds of each ligand were then automatically assigned using the AUTOTORS utility of AutoDock 4.0. Docking was performed using the Lamarckian Genetic Algorithm (LGA) method with an initial population of 400 random individuals; a maximum number of 10,000,000 energy evaluations; 100 trials; 50,000 maximum generations; a mutation rate of 0.02; a crossover rate of 0.80 and the requirement that only one individual can survive into the next generation. A total of eleven independent virtual screening runs were performed against the full set of docked ligands with all residues of the receptors set rigid during docking experiments. This set of polβ models comprises one structure that represents the minimized NMR conformation of polβ and ten conformations that represent ~85% of the MD trajectory (see Results).

D.6. Clustering of Docked Poses

Clustering of docked poses followed the same procedure described in section 3.4.2.

D.7. Rescoring Of Top Hits Using MM-PBSA

Binding free energies were calculated using the molecular mechanics Poisson-Boltzmann surface area (MM–PBSA) method[13] as implemented in AMBER10 and described in section 3.7.

D.8. Bibliography

1. Maciejewski, M. W.; Liu, D.; Prasad, R.; Wilson, S. H.; Mullen, G. P., Backbone dynamics and refined solution structure of the N-terminal domain of DNA polymerase beta. Correlation with DNA binding and dRP lyase activity. *J Mol Biol* **2000,** *296* (1), 229-53.
2. Laxmikant Kalé; Robert Skeel; Milind Bhandarkar; Robert Brunner; Attila Gursoy; Neal Krawetz; James Phillips; Aritomo Shinozaki; Krishnan Varadarajan, NAMD2: Greater Scalability for Parallel Molecular Dynamics. *Journal of Computational Physics* **1999,** *151*, 283-312.
3. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-25.
4. Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W522-5.
5. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J Comput Chem* **2005,** *26* (16), 1668-88.
6. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-74.
7. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem* **2002,** *23* (16), 1623-41.
8. http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html *(Last checked May,20, 2010)*.
9. http://zinc.docking.org/vendor0/dbsm/index.html *(Last checked May 20, 2010)*.
10. Pelletier, H.; Sawaya, M. R.; Kumar, A.; Wilson, S. H.; Kraut, J., Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* **1994,** *264* (5167), 1891-903.
11. Garrett MM; David SG; Robert SH; Ruth H; William EH; Richard KB; Arthur JS, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem.* **1999,** *19*, 1639.
12. Gasteiger J; Marsili M, Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* **1980,** *36*, 3219.
13. Kollman PA; Massova I; Reyes C; Kuhn B; Huo S; Chong L; Lee M; Lee T; Duan Y; Wang W; Donini O; Cieplak P; Srinivasan J; Case DA; Cheatham TE, Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum model,. *Acc. Chem. Res.* **2000,** *33*, 889.

Appendix E: Methods For Chapter 7

E.1. MD Simulations

The amino-terminal domain of MDM2 (residues 25-109) bound to a 13-residue transactivation domain peptide of p53 (residues 17-29) was taken from PDB entry 1YCR.[1] MD simulations were carried out using the NAMD program at a mean temperature of 310K and physiological pH (pH 7) using the all-hydrogen AMBER99SB force field. Protonation states of all ionizable residues were calculated using the program PDB2PQR. Following parameterization, the MDM2 protein alone (subsequent to removing the p53-peptide from the p53-MDM2 crystal structure) or in complex with the p53 peptide was immersed in the center of TIP3P water cube after adding hydrogen atoms to the initial protein structure. The cube dimensions were chosen to provide at least a 20Å buffer of 12724 (12653) water molecules around the systems. To neutralize and prepare the p53-bound or (free) systems under a physiological ionic concentration, 30 (28) chloride and 23 (23) sodium ions were respectively added by replacing water molecules having the highest electrostatic energies on their oxygen atoms. The number of counter ions for each case was calculated by first estimating the amount of ions that is needed to set up the system in normal physiological conditions (pH 7), followed by adding the number of chloride ions required to bring its charge to zero. The fully solvated protein was then minimized and subsequently heated to the simulation temperature with heavy restraints placed on all backbone atoms. Following heating, the system was equilibrated using periodic boundary conditions for 100 ps and energy restraints reduced to zero in successive steps of the MD simulation. The simulations were then continued for 55 (78) ns during which atomic coordinates were saved to the trajectory every 2 ps. The total simulation time was determined by visualizing the quality of sampling as predicted by PCA (see below). The RMSD and B-factors for the protein backbone were then computed over the last 10 ns of the MD simulation using the PTRAJ utility within AMBER10. Hydrogen bond analyses were performed by computing the average distance between donor and acceptor atoms. A hydrogen bond was defined by a heavy donor – heavy acceptor distance ≤ 3.4 Å, a light donor-heavy acceptor distance ≤ 2.5 Å, and a deviation of less than $\pm 60^o$ from linearity.

Following the same MD protocol mentioned above we prepared two equilibrated models for the PMI/MDM2 (PDB entry: 3EQS) and PMI/MDMX (PDB entry 3EQY) complexes.[2] Parameters for ligands were assigned using the generalized AMBER force field and partial charges were calculated with the AM1-BCC method using Antechamber in the AMBER 10 package. Following parameterization, the protein/ligand complexes were subjected to the same MD protocol we used before (see above) for a production phase of 2 ns. Snapshots were extracted every 2 ps and used for the MM/PBSA binding energy analysis.

E.2. Extracting Representative Structures

Following the procedure described in chapter 3 (section 3.3.5), clustering analysis was performed on the trajectories of the free and bound MDM2 MD simulations. To generate a reduced set of representative MDM2 models, we carried out RMSD conformational clustering with the average-linkage algorithm as implemented in the PTRAJ utility of AMBER10 using cluster counts ranging from 5 to 100 clusters. For the two MDM2 simulations, structures were extracted at 2 ps intervals over the entire simulation times. All C_α-atoms were RMSD fitted to the minimized initial structure in order to remove overall rotation and translation. RMSD-clustering was performed on the 18 residues that line the p53 binding cleft within MDM2, namely those numbered: 25, 26, 50, 51, 54, 58, 61, 62, 67, 72, 73, 93, 94, 96, 97, 99, 100, 104. These residues were clustered into groups of similar conformations using the atom-positional RMSD of the entire amino acid, including side chains and hydrogen atoms, as the similarity criterion. The optimal numbers of clusters for the two systems were chosen after evaluation of the two clustering metrics, described above, for different cluster counts (see results). Sixty clusters were obtained for the apo-MDM2, while thirty clusters were extracted for the holo-MDM2. The centroid of each cluster was chosen as the cluster representative structure and the most dominant structures were used as rigid templates for the ensemble-based docking experiments (see results).

E.3. Principal Component Analysis

PCA analysis as described in chapter 3 (section 3.3.4) reduced the original Cartesian coordinates' space of the MDM2 simulations into a reduced set of independent variables comprising that represent its essential dynamics. The same analysis was used to reassure the quality of sampling of the MD simulations (see results).

E.4. Selection of Ligand Database

The National Cancer Institute Diversity Set (NCDIS), DrugBank-small-molecules and a set of 3,168 derivative structures for Nutlin-3, MI-219 and TDP665759 were used as our test libraries of compounds. The description of the NCIDS and DrugBank-small molecule libraries are presented in chapter 2 (section 3.2.1). We also appended the set of derivative structures of MDM2-inhibitors to the docked compounds for two reasons. First, we wanted to build upon the intensive efforts that have been previously done and incorporate variations in the original structures for these inhibitors in order to improve their performance in binding to MDM2 and MDMX. Moreover, since we search for dual-MDM2/MDM4 inhibitors, we expect that due to the structural similarity of the p53-binding sites within the two proteins, an

MDM4-inhibitor should be a derivative structure of the known MDM2-inhibitors. Based on this assumption, we created a library of 3,168 derivative structures similar to Nutlin-3, MI-219, and TDP665759 by searching the PubChem database and then extracting the results using a similarity that is greater than or equal to 90%.[3] Compounds similar to the query structure are measured using the Tanimoto score (see section 2.3.1). A Tanimoto score of 100% represents an "exact match" to the provided chemical structure query, while a value of 0% means return all chemical structures deposited in the PubChem database. The threshold of >=90% is chosen for efficiency of search since similarity links in PubChem are pre-computed at this value. Also, at this threshold, the compounds that are returned by the search would not be very close from the original query structure and yet provide reasonable number of chemical structures for this work. Therefore, the full set of ligands used in this study comprised 6,617 different compounds.

E.5. Ligand Screening

Virtual screening on the p53 binding sites within MDM2 and MDM4 was performed using AutoDock, version 4.0. Hydrogen atoms were added to MDM2, MDM4 and ligands and partial atomic charges were then assigned using the Gasteiger-Marsili method. Atomic solvation parameters were assigned to the atoms of the protein using the AutoDock 4.0 utility ADDSOL. Docking grid maps with $126 \times 108 \times 126$ points and grid point spacing of 0.21 Å was then centered on the p53 binding site within the MDM2 and MDMX receptors using AUTOGRID4.0 program. Rotatable bonds of each ligand were then automatically assigned using AUTOTORS utility of AutoDock. Docking was performed using the Lamarckian Genetic Algorithm (LGA) method with an initial population of 400 random individuals; a maximum number of 10×10^6 energy evaluations; 100 trials; 50,000 maximum generations; a mutation rate of 0.02; a crossover rate of 0.80 and the requirement that only one individual can survive into the next generation. A total of twenty-eight independent virtual screening runs were performed against the full set of docked ligands with all residues of the receptors set rigid during docking experiments. This set of MDM2 models comprises one structure that represents the minimized holo-crystal conformation of MDM2, twenty-two conformations that represent ~80% of the apo-MDM2 trajectory and five models that constitute ~ 75% of the holo-MDM2 trajectory (see results). We also performed a VS run on an equilibrated model for MDMX using the top 300.

E.6. MDM2-Hits Resulted From The Ensemble-Base Screening

Following the procedure described in section 3.4.2, all docking results from the MDM2 screening were iteratively clustered in order to identify the optimal number of clusters and their best clustering pattern (see results). The primary ranking of docking results employed the same criteria that were followed for the ERCC1 case. That is, a ligand was considered as a top hit if the largest cluster of docking has at least 25% of the population of all docking solutions.

E.7. Rescoring Of Top Hits Using MM-PBSA

All binding free energy analysis was calculated using the molecular mechanics Poisson-Boltzmann surface area (MM–PBSA) method as implemented in AMBER10. The equations and description of the method is illustrated in chapter 3 (section 3.7).

E.8. Bibliography

1. Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. P., Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **1996,** *274* (5289), 948-53.
2. Pazgier, M.; Liu, M.; Zou, G.; Yuan, W.; Li, C.; Li, J.; Monbo, J.; Zella, D.; Tarasov, S. G.; Lu, W., Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX. *Proc Natl Acad Sci U S A* **2009,** *106* (12), 4665-70.
3. http://pubchem.ncbi.nlm.nih.gov/ (Last checked May 1, 2011).

Appendix F: Methods For Chapter 8

F.1. Generation Of The Mutated Structure

Starting from Cho's 1TSR[1] wild type p53 crystal structure we used the software DeepView (Swiss PDB Viewer)[2]to create different mutant configurations. The program suggested different orientations for the side chain of the mutated residue. These side chain rotations are usually restricted to a number of pre-defined experimental conformations stored in rotamer libraries. After exploring the rotational degrees of freedom for each generated structure we selected the most favorable configuration that had less satiric clashes and more hydrogen bonds with the surrounding residues. The most favorable mutant generated model was then handed to subsequent molecular dynamics simulations for its heating and equilibration at the targeted temperatures.

F.2. Molecular Dynamics Simulations

The wild-type and mutant structures both with (holo) and without (apo) the DNA were subjected to different molecular dynamics simulations over a temperature range of 25-37^{0}C employing the software NAMD[3] at physiological pH (pH 7) using the all-hydrogen AMBER99SB force field.[4] Protonation states of all ionizable residues were calculated using the program PDB2PQR.[5] The three cysteine residues along with the histidine residue that are coordinating the Zn^{+2} ion were deprotonated. Following parameterization, and keeping co-crystallized water molecules in their locations in the initial 1TSR structure, the newly generated systems were immersed in the center of TIP3P water cube after adding hydrogen atoms to the original protein, DNA and water structures. The cube dimensions were chosen to provide at least a 20Å buffer of water molecules around the systems. To neutralize and prepare the protein-DNA complexes with a physiological ionic concentration, chloride and sodium ions were respectively added by replacing water molecules having the highest electrostatic energies on their oxygen atoms. The number of counter ions for each case was calculated by first estimating the number of ions that would be needed to set up the system at normal physiological conditions (pH 7), followed by adding the number of chloride (sodium) ions required to bring the net charge to zero. The fully solvated systems were then minimized and subsequently heated to the simulation temperatures with heavy restraints placed on all backbone atoms. Following heating, the systems were equilibrated using periodic boundary conditions for 100 ps and energy restraints were reduced to zero in successive steps of the MD simulations. The simulations were then continued for ~10 ns and for the last 1 ns, atomic coordinates were saved to the trajectory every 0.1 ps for subsequent collective dynamics analysis and binding energy calculations.

F.3. Clustering Analysis Protocol

Clustering analysis followed the procedure described in section 3.3.5.

F.4. Hydrogen Bonding Analysis

Hydrogen bond analyses were performed using the visualization software VMD[6] on the dominant structures extracted from clustering analysis. A hydrogen bond was defined by a cutoff distance of 3.5Å between a donor and acceptor atom and an angular deviation less than 50^0 from linearity.

F.5. Binding Energy Analysis And Energy Decomposition

Binding energy and it’s decompositions into residue contributions were calculated using the same procedure described in section 3.7 with the DNA molecule was considered as the ligand.

F.6. Bibliography

1. Cho, Y.; Gorina, S.; Jeffrey, P. D.; Pavletich, N. P., Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **1994,** *265* (5170), 346-55.
2. Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C., SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* **2003,** *31* (13), 3381-5.
3. Grindon, C.; Harris, S.; Evans, T.; Novik, K.; Coveney, P.; Laughton, C., Large-scale molecular dynamics simulation of DNA: implementation and validation of the AMBER98 force field in LAMMPS. *Philos Transact A Math Phys Eng Sci* **2004,** *362* (1820), 1373-86.
4. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-25.
5. Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W522-5.
6. http://www.ks.uiuc.edu/Research/vmd/.