Structure-Based Drug Design and QSAR analysis of candidate drugs for the treatment of Cancer

by

Pelin Armutlu

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Industrial Engineering

Koç University

June 2008

Koc University
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Pelin Armutlu

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

____
Metin Türkay, Ph. D. (Advisor)

____
İbrahim Halil Kavaklı, Ph. D.

____
Ceyda Oğuz, Ph. D.

Date: ____________________________

ABSTRACT

In rational drug design known chemical interactions related to the target disease are exploited to discover novel drugs. By identifying the important reaction pathway and the key enzyme related to a specific disease is the first step in structure-based drug design, then; the active site of the enzyme is analyzed and molecules are designed to competitively bind to that active site.

In this study, structure-based rational drug design approach is employed to discover novel inhibitors for two different target proteins. The first target protein is Cytochrome P450 C17, the key enzyme in androgen synthesis and, the second target protein is XPF, a key member of the DNA excision repair system. In designing an inhibitor targeting Cytochrome P450 C17, we aim to prevent the progression of the prostate cancer, and targeting the XPF-ERCC1 pair we aim to design chemotherapeutic agent "aids" that prevent DNA repair and resistance development in cancerous cells against the chemotherapeutic agents, where as a result we were able to identify several promising candidate drugs.

Finally, a novel QSAR approach to predict the activity level of the inhibitors is introduced, since a priori analysis of the activity of inhibitors on the target protein by computational approaches can be useful in narrowing down drug candidates for further experimental tests to save from time and resources. The calculations utilizing the approach presented in this thesis resulted in better accuracies in classifying the activity of candidate drugs among other data mining tools presented in the literature.

ÖZET

Yapıya dayalı ilaç dizaynında, hedeflenen hastalikla ilişkisi bilinen kimyasal reaksiyonlarin ozelliklerinden faydalanılır. Hastalıgın gelişiminde rol oynayan kimyasal reaksiyonların ve bu reaksiyonlarda görev alan önemli enzimlerin saptanması yapıya dayalı ilaç dizaynının ilk adımıdır. Daha sonra, hedef olarak belirlenen enzimin aktif bolgesi belirlenir ve bu bolgeye yapışarak enzim çalışmasını engelleyecek moleküller geliştirilir.

Bu çalışmada, yeni ilaçlar geliştirilmesi hedeflenerek, yapıya dayalı ilaç dizaynı teknikleri iki farklı hedef proteine uygulandı. Cytochrome P450 C17, androjen sentezinde rol alan en önemli enzimlerden biridir ve bu enzimi hedefleyen bir ilaçla prostat kanserinin gelişimi engellenebilmektedir. Ikinci hedef protein olan XPF, DNA onarım sistemindeki en önemli proteinlerden biridir ve bu çalışmada XPF-ERCC1 protein komplexini hedef alarak DNA onarımını engelleyip kanserli hücrelerin kemoterapik ilaçlara karşı direnç kazanmasını önleyecek moleküller elde etmeyi amaçladık. Her iki çalışmada da umut vaad eden moleküller elde ettik.

Çalışmanın son kısmında sunulan yeni bir Sayısal Yapı-Aktivite İlişkisi analizi metoduyla, dizayn edilen ilaçların labaratuar deneyleri yapılmaksızın aktivite seviyelerini öngörebilmeyi amaçladık. Bu çalışmada sunulan method literatürde bulunan diğer methodlardan daha iyi sonuçlar elde etmeyi başardı.

ACKNOWLEDGEMENTS

There are a number of people I am deeply grateful without whom this thesis would not be completed.

I thank Dr. Metin Türkay and Dr. İ. Halil Kavaklı, for believing in me and for their patience, guidance and support during my graduate study and the thesis process. They created a candidate scientist from a curious college graduate. I have learned so much from them.

I thank Dr. Ceyda Oğuz, Dr. Fikri Karaesmen and Dr. Seda Kızılel for participating in my thesis committee, for their insight, and guidance.

I thank Muhittin Emre Özdemir for being a great Sensei and a friend.

I thank Ulaş Kemal Ayaz for keeping me as sane as possible.

I thank Fadime, Ali, Pınar and Can for having me as a friend and office mate with all my craziness.

I thank my dear friends Ferda, Figen, Neslihan, Selim, Seray, Sezer and Uğur for being my best friends, with them these two years were the best years of my life.

I thank Cengiz, Osman, Gözde, Nurcan, Emre, Ekin, Ekin, Bahar, Besray, Sefer, Özge, Seda, Natali, Hande, Onur for making the engineering building a nice place to live.

I thank Emir, Onur, Ati, Serhat, Mustafa, Sinan for meeting me every week in another world.

Finally, I thank my parents Aysen and Muharrem for their love, patience, and never ending support and my brother Emre for being the best roommate, the best cook, the best therapist, best stylist and being the better half of me.

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Historically, drug discovery was a study of trial and error testing of substances on animals or cultured cells and analyzing the effects, while in rational drug design known chemical interactions and responses related to a specific disease are exploited. Chemical substances are designed either to upregulate or to inhibit certain key reactions according to the desired treatment outcome [1]. In structure-based rational drug design, first the important reaction pathway and the key protein are identified, then 3-dimensional structure and the active site of the target protein is modeled, and finally molecules are designed to competitively bind to that active site. Computational tools are used to model the active site, to study interactions of the drug molecules with the target protein and, to predict the effectiveness of the drug molecules prior to in vitro or in vivo experiments [2].

Yet, another important consideration in designing drug molecules is the specificity of the drug molecule to the target protein, in other words the drug molecule should be selective and should not affect any other pathway in the organism.

Figure 1.1 shows the common steps of the rational drug design. The most critical input element for structure based drug design is the 3-dimensional structure of the target protein. The 3-dimensional confirmations of the proteins are derived by techniques such as x-ray crystallography and NMR spectroscopy. If the 3-dimentional structure of the target protein is not available, the structure of a closely related analog might be used directly or by homology modeling techniques a model of the target protein might be generated. Next, the features of the protein active site are identified and modeled [3].

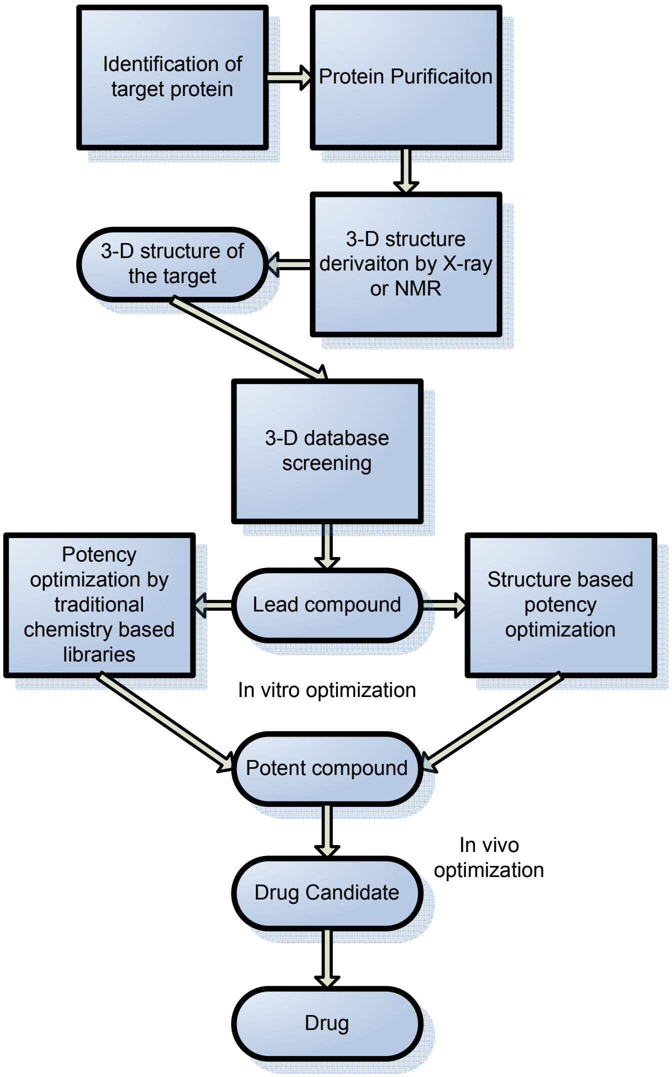

Identification of target protein
Protein Purificaiton
3-D structure derivaiton by X-ray or NMR
3-D structure of the target
3-D database screening
Potency optimization by traditional chemistry based libraries
Lead compound
Structure based potency optimization
In vitro optimization
Potent compound
In vivo optimization
Drug Candidate
Drug

Figure 1.1: Steps of rational drug design

Computational techniques such as docking, screening and molecular dynamics simulations are then utilized to design potential molecules that are structurally and chemically compatible with the target active site. Besides the experimental high throughput screening of libraries with robotics integrated systems, computational virtual screening methods are recently widely used by fast docking of vast number of molecules in databases into the binding site. Note that; high throughput screening has several disadvantages such as: high experimental costs, small number of readily available chemical substances for testing, and possible interactions of molecule with other binding sites on target protein [4].

Once a lead compound is discovered, first these initial leads should be confirmed experimentally, then structure based drug design methods are again used in potency optimization to increase affinity and selectivity by studying the molecular structure and the characteristics of the drug lead and the potential interaction schemes of the protein-drug complex to enhance the activity and the bioavailability of the lead compound.

In this study, structure based rational drug design approach is employed to discover novel inhibitors for two different target proteins through a cancer research perspective. The first target protein is Cytochrome P450 C17, the key enzyme in androgen synthesis. The second target protein is XPF, a key member of the DNA excision repair system. For the former, we aim to decrease androgen levels in the cells by inhibiting the activity of Cytochrome P450 C17 and therefore prevent the progression of the prostate cancer. In the latter case, blocking the interaction between XPF-ERCC1 hetero dimer to abolish excision repair is not a direct chemotherapeutic agent design approach, but we aim to design XPF-ERCC1 inhibitors as chemotherapeutic agent “aids” that prevent DNA repair and resistance development in cancerous cells against the chemotherapeutic agents.

Chapter 2 provides necessary background and literature review on the computational methods in structure-based drug design.

Chapter 3 is dedicated to XPF-ERCC1, first necessary biological background and the determination of the active site is presented, then the details of the employed computational methods with obtained results are discussed.

Chapter 4 is on designing Cytochrome P450 C17 inhibitor derivatives. The previous research on the subject is given, the strategies in designing inhibitor derivatives are explained, and the resulting inhibitors are presented.

A priori analysis of the activity of inhibitors on the target protein by computational approaches can be useful in narrowing down drug candidates for further experimental tests. A novel QSAR approach to predict the activity level of the inhibitors is discussed in Chapter 5.

Finally, this thesis is concluded with a summary of the performed study and future work in Chapter 6.

Chapter 2

OVERVIEW

2.1 Molecular Dynamics Simulation (MD)

Molecular dynamics simulation is the simulation of the motion of atoms and molecules, which interact for a period of time under laws of physics [5]. Since molecular systems contain a large number of particles, it is not possible to understand molecular interactions experimentally. Therefore; MD is used to link laboratory experiments with biophysical theories [6]. Recently, MD is generally used as a powerful tool to understand the relationship between molecular structure, function and dynamics [7]. Biological sciences benefit from MD simulations to study the affects of solvents, temperature and pressure on biological systems, especially proteins [8]. Before the development of dynamic models by MD simulations, proteins were studied as rigid structures and this strategy was problematic since internal motions and conformational changes are vital for their function [7].

There are three main applications of MD simulations to study biological systems. The first application involves the determination and refinement of structures with the data obtained from experiments, the second one uses MD to understand the characteristics of the systems in equilibrium and, the third one studies the molecular dynamics itself [7].

2.1.1 NAMD

In this study, among the several MD software packages NAMD (Nanoscale Molecular Dynamics) [9] is used with CHARMM [10] force field parameters. VMD [11] is used to prepare the input files for the MD simulations, and to visualize and analyze the simulation outputs.

NAMD aims to simulate dynamic systems by mimicking a realistic cell environment. Newtonian equations of motion are solved for each atom in the system to determine atomic trajectories, the control of pressure and temperature are enforced by statistical mechanics and, partial mesh Ewald is used for the evaluation of the electrostatic forces.

For a predetermined time-step length (*Δt*) and simulation duration, the final coordinates of the atoms in the system are calculated iteratively. At each iteration, the coordinates and the velocities of the atoms from the previous step are used with the cumulative potential energies of atoms in order to calculate the new coordinates. Newtonian equations of motion are applied for the corresponding calculations.

$$F_i = m_i a_i = m_i \frac{d^2 r_i}{dt^2} = -\frac{\partial U}{\partial r_i} \tag{2.1}$$

At time = t_0 initial coordinates of the system are known from the structure file and the initial velocities are assigned by calculating the expected value of the kinetic energy of the system at an equilibrium for a known T(temperature), where v_i is assumed to come from a Gaussian distribution with mean 0 and standard deviation ($k_B\ T/\ m_i$).

$$\hat{E}_{kin} = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 = \frac{1}{2}(N)k_B T \tag{2.2}$$

$$\hat{v}_i^2 = \frac{k_B T}{m_i} \tag{2.3}$$

Then, at each iteration the system is advanced for a time-step (*Δt*), the forces and the velocities are recalculated by Newtonian equations of motion (eq. 2.1), where the total potential energy U is given as a sum of bonded and non-bonded interaction potentials (eq. 2.4).

$$U = U_{bondstretching} + U_{bending} + U_{torsional} + U_{Lennard-Jones} + U_{electrostatic} \tag{2.4}$$

The bonded interactions include bond stretching, bending and torsional interactions and, the corresponding potentials are calculated as follows:

$$U_{bondstretcing} = \sum_{bonds} K_b (b - b_0)^2 \tag{2.5}$$

$$U_{bending} = \sum_{angles} K_\theta (\theta - \theta_0)^2 \tag{2.6}$$

$$U_{torsional} = \sum_{\substack{dihedral \\ angles}} K_\phi [1 + \cos(n\phi - \delta)] \tag{2.7}$$

The non-bonded interaction potentials are the Lennard-Jones potentials (eq 2.8), which approximate Van der Waal’s forces, and electrostatic interactions potentials (eq 2.9).

$$U_{Lennard-Jones} = \sum_{\substack{non-bonded \\ pairs}} \frac{A}{r^{12}} + \frac{B}{r^6} \tag{2.8}$$

$$U_{electrostatic} = \sum_{\substack{non-bonded \\ pairs}} \frac{q_1 q_2}{\varepsilon r} \tag{2.9}$$

2.2 Protein-Small molecule Docking

Molecular docking is a method to predict the preferred orientation of a ligand with a receptor in a stable bound complex form. Binding affinity of the complex can be calculated by scoring functions to predict the strength of the association [12]. Recently, docking methods are widely used to predict the binding position and the affinity of small drug molecules to protein targets to have an understanding on their activity (agonism or antagonism). In other words, the aim of the molecular docking is to simulate the molecular recognition process computationally. Therefore, molecular docking is an important tool in structure based drug design mainly for hit identification and lead optimization processes [13]. Hit identification is performed by the quick screening of large databases to find potential drugs in silico, by combining molecular docking with scoring functions to optimize binding free energies of the complex. In lead optimization, molecular docking is used to identify the orientation (docking pose) of the drug molecule in the binding pocket and interacting parts of the drug molecule with the receptor protein. Therefore, more potent and selective inhibitors can be designed.

The molecular docking problem can be viewed as an optimization problem predicting the "best-fit" orientation of the protein-ligand complex [14], where the objective function is the minimization of the free energy of the system. If both receptor protein and small molecule ligand are flexible in the model, the "best-fit" orientation of the complex is formed by the conformational adjustment of two flexible pairs, and the final confirmation is called an "induced-fit" [15].

Molecular docking algorithms composed of two parts: a search algorithm and a scoring function. Several search algorithms (simulated annealing, fragment building and genetic algorithms) are developed to find the "best-fit" orientation among the all possible confirmations of protein-ligand complex (search space). Then, scoring functions are utilized to approximate the strength of the non-covalent interactions (binding affinity) between receptor and ligand [16], where binding affinity scores are used to determine the most promising molecules in drug design. There are three types of scoring functions, first of which is force-field scoring function, where binding affinity is approximated by the strength of non-bonded interactions by computing total van der Waals and electrostatic interactions between protein and ligand. The second is empirical scoring function [17] that calculates the number of hydrogen bonds, hydrophobic contacts, hydrophilic contacts and number of immobile rotatable bonds, and the third one is the knowledge based scoring function [18], which is based on statistical observations of intermolecular contacts.

2.2.1 Autodock

Autodock, which is a flexible ligand docking tool, is used for virtual screening and detailed docking analysis. Autodock predicts the optimal confirmations of the receptor-ligand complex and report binding affinity scores by assuming a structure model with a rigid receptor (protein) and a flexible ligand (drug molecule).

Among several different search algorithms (Monte Carlo Simulated Annealing, Genetic Algorithm) Lamarckian-Genetic Algorithm is used due to its high degrees of freedom. Finally, a force-field type scoring function is used in order to calculate binding affinity scores of the protein-ligand complex. For the calculations AutoDock uses AutoGrid component to generate atomic affinity grid maps for each type of atom in the ligand, where every atom is assigned a non-bended interaction potential with the protein and electrostatic potentials. The AutoDock scoring function is based on an empirically-derived linear free energy model (eq 2.10) including terms for van der Waals energy, hydrogen bond energy, Coulombic energy, change in desolvation free energy and the loss of torsional degrees of freedom upon binding.

$$FreeEnergy = 0.1485 vdW + 0.1146 elec + 0.0656 Hbond + .3113 tors + 0.1711 desolv \quad (2.10)$$

As a result of the docking simulation, AutoDock computes intermolecular energy, internal energy and torsional energy as outputs, the first two forms the ‘docking energy’, while the first and the third combine together to give ‘binding energy’.

2.2.2 Autodock Docking Setup

To run a docking algorithm AUTODOCK requires four types of input files: the PDBQS file for the protein, the PDBQ file for the ligand, the GPF file to create the active site grid parameter files and the DPF file containing the docking parameters. These files may be prepared by ADT (AUTODOCK tools) user interface as well as they may be prepared by command terminal scripting for large scale docking purposes.

To prepare the PDBQS file for the macromolecule first the polar hydrogen atoms should be added to the PDB structure, but since the PDB structure file derived from the MD simulation already contains the polar hydrogens, only the non-polar hydrogens are merged. Then, the Kollman charges are added and the final structure is stored in the PDBQS file.

Next, the PDBQ file is prepared for the ligand molecule. Polar hydrogens are added and non-polar hydrogens are merged as in the preparation of the PDBQS file for the protein. Then, Kollman charges are added if the ligand file is a peptide, otherwise Gasteiger charges are added. Finally, by calculating the angle between consecutive C atoms, planar and non-planer C atoms are marked as well as the torsional freedom of the bonds. All of the above information is then stored in the PDBQ file for the flexible ligand docking.

The GPF files contain the parameters of the grid maps that will be created to define the active site for the docking. AUTODOCK models the active site as a box and the grid maps are modeled accordingly. The GPF file contains the grid center coordinates of this box, the grid size as number of points, the spacing between two grid points and, the number of the grid maps that will be created. The number of the required grid maps depends on the types of atoms that are present in the ligand molecule. Since this study involves the large scale virtual screening of a compound set including millions of molecules, grid maps are prepared for each type of atom that may be present in drug-like molecules.

In this study, two separate GPF files are prepared: first file is to create the grid maps for the virtual screening, the second is to create the grid maps for high-grid detailed docking. The differences between these grid map parameters make up the major difference between the virtual screening and high-grid detailed docking. For virtual screening, to save from computational time, the corresponding GPF file is designed to create rather low resolution maps. The grid box including the entire active site is defined with 0.375 Å spacing in these low resolution maps where as for detailed docking purposes the spacing parameter is decreased to 0.150 Å.

The DPF file contains the setup for the run parameters of the Lamarckian-Genetic Algorithm such as: population size, number of generations, number of runs, crossover rate, mutation rate and number of evaluations. Here, also two separate DPF files are prepared: one is for virtual screening and the other is for detailed docking. In the DPF file, the run parameters for virtual screening are defined as follows: population size is 50, number of generations is 2.7 x 10^4, crossover rate is 0.8, mutation rate is 0.02, number of runs is 10, and number of evaluations is 1 x 10^6 and the run parameters for detailed docking are defined as: population size is 250, number of generations is 5 x 10^4 , crossover rate is 0.8, mutation rate is 0.02, number of runs is 256, and number of evaluations is 1.5 x 10^7 .

In this study, for molecular docking the structure files of the macromolecules are derived from results of the MD simulation and the ligand structure files are obtained from a commercial compound database, AMBINTER. Note that, for both XPF and P450 C17 docking, a molecular weight filter is applied beforehand on the ligand set to save from the computational time that may be spent on the docking of molecules too large to fit the defined active site. For P450 C17, molecules heavier than 350 g/mol are not used in virtual screening being too large for the active site and for XPF this threshold is set to be 390 g/mol.

2.3 Protein-Protein Docking

Protein-protein docking is defined as the prediction of molecular structure of the protein complexes without laboratory experiments. Proteins that do not change their structure during the complex formation can be docked with high success rates, where as several methods are still under development for the protein complexes that change internal confirmation substantially [19]. Protein-protein docking mainly aims to predict whether proteins actually bind in vivo or not, the structure in the complex form and the strength of the interaction [20].

There are two types of protein-protein docking methods widely used in literature: rigid-body docking and flexible docking. In rigid body docking bond angles, bond lengths and torsion angles are held constant during docking, and this type of docking is weak in predicting structures where, substantial conformation changes occur during protein-protein binding [21, 22]. Therefore, for these structures flexible docking methods are developed allowing a set of intelligent conformational changes during the simulation.

2.3.1 Rosetta

Rosetta predicts protein-protein complex structures from the coordinates of the unbound protein monomers. The prediction algorithm first utilizes a rigid-body Monte Carlo search, and then by using Monte Carlo minimization, simultaneous optimization for backbone displacement and side-chain conformations is performed. During the main algorithm, several independent simulations are performed up to a total number of 105, and the resulting “decoys” are ranked by using an energy function including components such as van der Waals interactions, implicit solvation model, and a hydrogen bonding model [22, 23]. Then, clusters are formed by using the top ranking “decoys”, and the final best predictions are reported.

2.3.2 HEX

Hex docking algorithm is based on a 3D protein model in which, each molecule is represented by parametric functions of surface shape, electrostatic charge and potential distributions. This representation allows each property to be described by a vector of coefficients, where surface shape representation uses a 3D surface skin model [23] of the protein topology and the electrostatic model is based on the classical electrostatic theory. Finally, the rigid body docking scores are reported as a function of the six degrees of freedom with avoiding overlapping pairs by additional expressions. Then, scaling factors are introduced to the derived scoring function and by minimizing the total weighted score best protein confirmations are reported.

Chapter 3

XPF-ERCC1

3.1 Introduction

3.1.1 Nucleotide excision repair

Nucleotide excision repair is the primary repair system for bulky base lesions, resulting from covalent binding between adjacent nucleotides, in *E-coli* and the only repair system in humans [24-26]. The damage is removed in three consecutive steps: damage recognition, removal of the nucleotide oligomer after dual incision, the resynthesis of the removed DNA and ligation [27-29]. In both systems damage recognition is an ATP independent process, where an unstable DNA protein complex is formed, and for the damage removal an ATP-dependent multisubunit excision nuclease is used [30]. As nucleotide excision repair removes the DNA damage, the length of the oligomer removed is 12-13 base pairs long in prokaryotes, and it is 24-32 base pairs long in humans [31, 32].

3.1.2 Excision Repair in Humans

In humans, the excision repair is carried out by an excision nuclease, which is composed of 6 subunits: XPA, RPA, XPC, TFIIH, XPG, XPF-ERCC1 [31]. XPA, RPA and XPC are DNA damage recognition proteins, TFIIH unwinds the double helix on the damaged site, XPG incises the DNA backbone $3^{'}$ to the damage, where as XPF-ERCC1 incises 5’ [33].

XPA, which is a 31-kDa protein, binds double stranded DNA with moderate preference for damaged regions [34]. RPA, which binds to single-stranded DNA, is shown to have an affinity for damaged DNA [35]. XPC is also a single-stranded DNA binding protein, which shows preference for damaged and bent DNA [36]. The three damage recognition enzymes are suggested to act cooperatively and bind to the damaged DNA in a random order [30]. The two helicase subunits (ERCC2, ERCC3) of TFIIH protein unwind ~20 nucleotides long DNA helix about the lesion site by hydrolyzing ATP [37]. Then, XPG is summoned to the damage site by XPC, which benefits from the hydrolysis of ATP by TFIIH for this action. XPC leaves the complex afterwards [38]. Finally, XPF-ERCC1 binds to the complex for 5' incision. It is shown that the nuclease activity of the pair is only detected when both proteins are present [39]. As a result, a 24-32 nucleotides-long oligomer is removed, including the damage site [32]. Next, the protein complex disassociates from the damage site except RPA, and polymerase (RFC/PCNA and Pol δ/ε) completes the gap and DNA ligase 1 completes the reactions [40]. The schematic representation of human excision repair system is given in Figure 3.1.

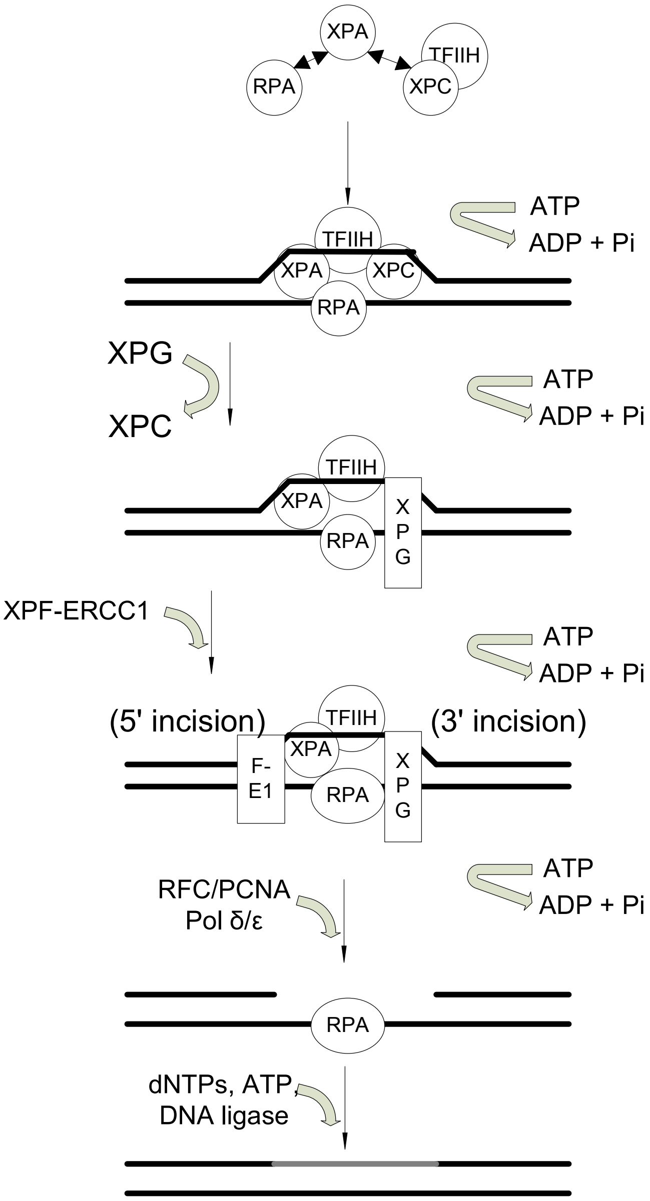


Figure 3.1: The schematic representation of human excision repair system

3.1.3 XPF-ERCC1

Besides its key role in nucleotide excision repair, making a single-strand incision adjacent to the damaged DNA, XPF-ERCC1 pair is shown to have several other repair functions in the cell. ERCC1-XPF is suggested to have a role in the repair of DNA double-strand breaks (DSB) by single-strand annealing (SSA) [41], and recently it is also argued to have a key role in ICL (interstrand crosslink) repair.

In cancer research interstrand crosslink (ICL) repair is an important and exciting area since unrepaired ICLs are highly cytotoxic. Therefore, ICL-inducing agents such as mitomycin C (MMC), nitrogen mustards (HN2s) and cisplatin are designed as promising chemotherapeutic agents [42]. However, development of resistance against the ICL-inducing agents by the cancerous cells is still a problem [43]. The importance of XPF-ERCC1 in ICL (Interstrand crosslink) repair is argued after XPF-ERCC1 mutant cell lines show 90-fold sensitivity to mitomycin C (an ICL- inducing chemotherapeutic agent) [44, 45].

Hence, we aim to design XPF-ERCC1 inhibitors as chemotherapeutic agent aids that prevent DNA repair and resistance development in cancerous cells.

3.2 The target site on protein

In this study, the strategy used for the inhibitor design for XPF-ERCC1 pair is preventing the dimerization of XPF-ERCC1 heterodimer. Experimental data suggests that the deletion of Phe-293 eliminates binding to XPF, where Phe-293 is naturally located in a pocket surrounded by XPF residues 837-905. The most important residues in this pocket reported to be Leu-841, Met-856, Val-859 and Ile-862 [46, 47]. Therefore, inhibitors are designed to competitively inhibit the dimerization of XPF-ERCC1 by binding to the 837-905 pocket on XPF. Several computational methods are also applied to further study the protein-protein interface and hot spot residues for more efficient design.

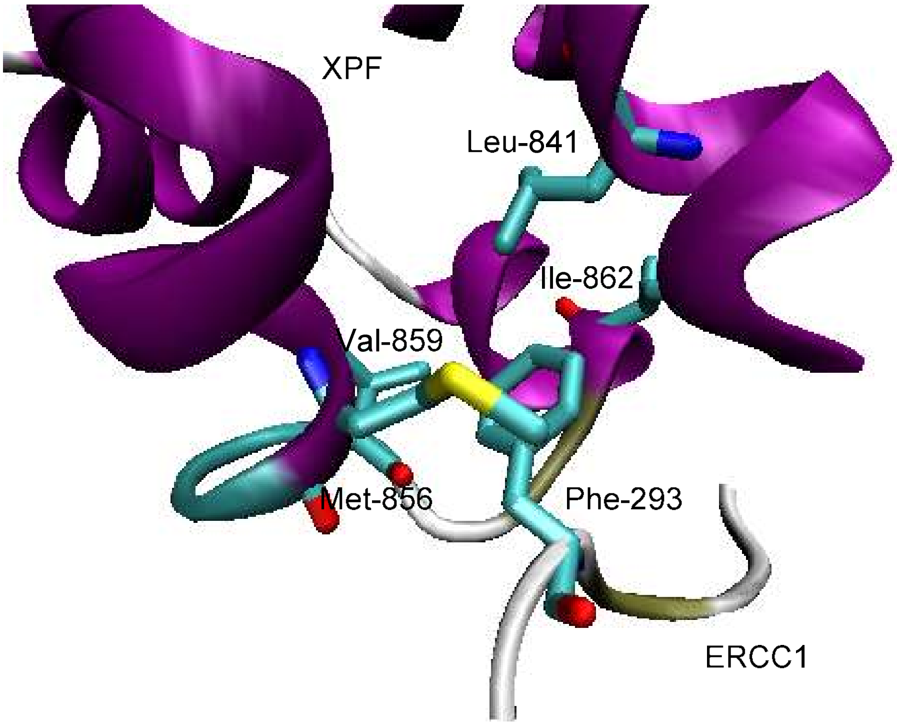


Figure 3.2: The representation of the binding site on XPF (Leu-841 Met-856 Val-859 Ile-862) interacting with the key interface residue Phe-293 on ERCC1

PDBsum server is used to visualize interface contact residues (Figure 3.3). (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl) Number of interface residues on XPF and ERCC1 are 20 and 21 respectively, where the surface area of interaction on XPF is1090 **Å**^{2}and on ERCC1 is 1137 **Å**^{2}, and on the interface there are a total number of 5 hydrogen bonds and 111 non bonded interactions.

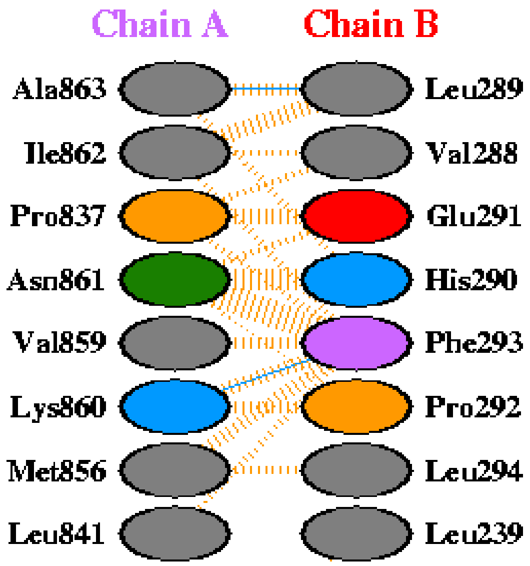


Figure 3.3: Schematic representation of contact residues around Phe-293, where blue lines represent hydrogen bonding and dashed lines represent non-bonded interactions.

The hotspot residues on the protein interface were also calculated by HotSprint server (http://prism.ccbb.ku.edu.tr/hotsprint/). HotSprint uses pScore+ASA for prediction, where pScore is the propensity scaled conservation score and ASA is accessible surface area. HotSprint reported Pro-837 Phe-840 Leu-841 Met-844 Met-856 Ile-862 Leu-865 and Phe-889 as hot spot residues on XPF, and Arg-234 Leu-239 Ser-259 and Phe-293 on ERCC1, which agrees with the experimental data previously reported [46, 47]. Also residue free energies are calculated by FastContact2.0 server, which reports the minimum free energy contacts as an indicator of the most important residues responsible for binding (Table 3.1). (http://structure.pitt.edu/servers/fastcontact/)

Min receptor-ligand residue free energy contacts				
Energy (kcal/mol)	ERCC1 residue		XPF residue	
-1.711	**293**	PHE	**856**	MET
-1.660	289	LEU	**862**	ILE
-1.502	238	CYS	840	PHE
-1.297	291	GLU	837	PRO
-1.194	260	LEU	890	ILE
-1.117	231	PHE	889	PHE
-1.089	**293**	PHE	**860**	LYS
-1.026	261	GLU	890	ILE
-1.020	260	LEU	889	PHE
-0.932	**293**	PHE	**841**	LEU
-0.859	**293**	PHE	**862**	ILE
-0.822	264	ILE	863	ALA
-0.824	237	GLU	843	LYS
-0.823	**293**	PHE	861	ASN
-0.810	238	CYS	844	MET
-0.807	294	LEU	**856**	MET
-0.800	235	VAL	889	PHE
-0.294	258	GLY	894	PHE
-0.757	289	LEU	863	ALA
-0.734	**293**	PHE	**859**	VAL

Table 3.1: Min receptor-ligand residue free energy contacts

Finally, computational alanine scanning [48] was applied to the XPF-ERCC1 pair to determine important interactions in the protein-protein interface (http://www.robetta.org/). Experimental “alanine scanning mutagenesis” is widely used as a powerful tool to measure the effect of deleting the side chain of an amino acid in the protein-protein interface. Here, only computational alanine scanning is used, which calculates affects of alanine mutations on the binding free energy of the protein complex, based on free energy calculations [48]. The free energy calculations involve a linear combination of Lennard-Jones potential [49], an implicit solvation model [50], hydrogen bonding potentials from protein structures [51] and some statistical terms to approximate rotamer probabilities and unfolded reference state energies [49, 52].

amino acid	chain	ΔΔG
840	A	3.07
862	A	1.61
889	A	2.74
894	A	2.61
231	B	2.48
234	B	1.26
260	B	1.13
264	B	1.64
293	***B***	***5.29***
294	B	1.05

Table 3.2: The highest ΔΔG score residues

In Table 3.2 only the residues with high binding free energy difference are reported, as they have the potential to be hotspot residues. The highest ΔΔG score belongs to the residue 293, where ΔΔG is the predicted change in binding free energy upon alanine mutation.

Hence, according to the experimental and the computational data, the inhibitor target site is determined on XPF to be the pocket surrounded by residues Leu-841, Met-856, Val-859, Lys-860, interacting the key interface residue Phe-293 in the original structure.

3.3 MD Simulation of XPF-ERCC1

3.3.1 Simulation Setup

Coordinate of the initial structure is obtained from Protein Data Bank (PDB code 2A1J)[47]. The reported structure is a truncated form of the full length XPF-ERCC1 where, XPF lacks the N-terminal helicase-like domain in the reported structure. Experimental studies show that this truncated form is still able to perform the structure-specific endonuclease activity with similar specificity to that of full-length XPF-ERCC1 [47]. The truncated XPF structure that is reported in this pdb file contains the target active site and necessary interface residues to bind the ERCC1 pair. First, MD simulations were performed on the XPF-ERCC1 pair to analyze the stability of the pair in specific temperature and pressure values mimicking the cell environment in the presence of water as the solvate. Then ERCC1 is deleted from the structure and XPF is simulated alone to study the stability of XPF without its pair and to predict the natural conformation of XPF backbone and side chains in the cellular environment. Predicting the natural conformation of the side chains about the active is especially important since; this relaxed structure given by the MD simulation will be used later in the virtual screening and the docking computations.

The input files to perform a simulation in NAMD are prepared by using several packages included in VMD. The *psfgen* package was used to create the protein structure file (PSF), which replaces the atom and residue names ones recognized by NAMD. VMD Solvate plug-in is used to solvated structure in a water-box with minimum 10 Å distance from any atom of protein to the boundary and, the system is neutralized by adding of Na, or Cl ions with VMD Autoionize plug-in.

The total MD simulation study is composed of the main steps: the equilibration and the simulation. During the equilibration first, 10000 steps minimization is performed only on side chain atoms by fixing the coordinates of backbone residues, then the whole system is run in 10000 steps simulation without pressure control. Next, to set the system temperature to the standard temperature of the cell, the simulation temperature is increased by 10 K increments up to 310 K. After each temperature increase 10ps simulation is performed by applying the corresponding Langevin temperature set.

After the equilibration and minimization steps the main simulation is performed with constant temperature and pressure control. The calculations are done in parallel at 4 X Intel(R) Xeon(R) CPU 2.33GHz with 4096 KB ram by NAMD 2.6 and CHARMM27 force field parameters. The time-step of the simulation is set to be 2 fs and, the bonded interactions, the van der Waals interactions (12 Å cut-off), and the long-range electrostatic interactions with partical-mesh Ewald (PME) is included in the calculations to define the forces acting on the system. The damping coefficient is set to be 5 ps^{-1} using Langevin dynamics to handle pressure control and, 1 atm constant pressure is set with 100 fs decay period and 50 fs damping time. With this simulation setup a 3 ns simulation is run.

3.3.2 MD Results

To analyze the MD simulations, VMD is utilized. Corresponding RMSD (root mean square deviation) versus time graphs are drawn according to the frames derived from the simulation at every 1000 time steps which has a step size of 2 ps.

On Figure 3.4 and 3.5, RMSD is plotted through the simulation. For every frame, RMSD is calculated by aligning the structure of the protein at that frame to the initial structure and calculating the distance between the residues one by one.

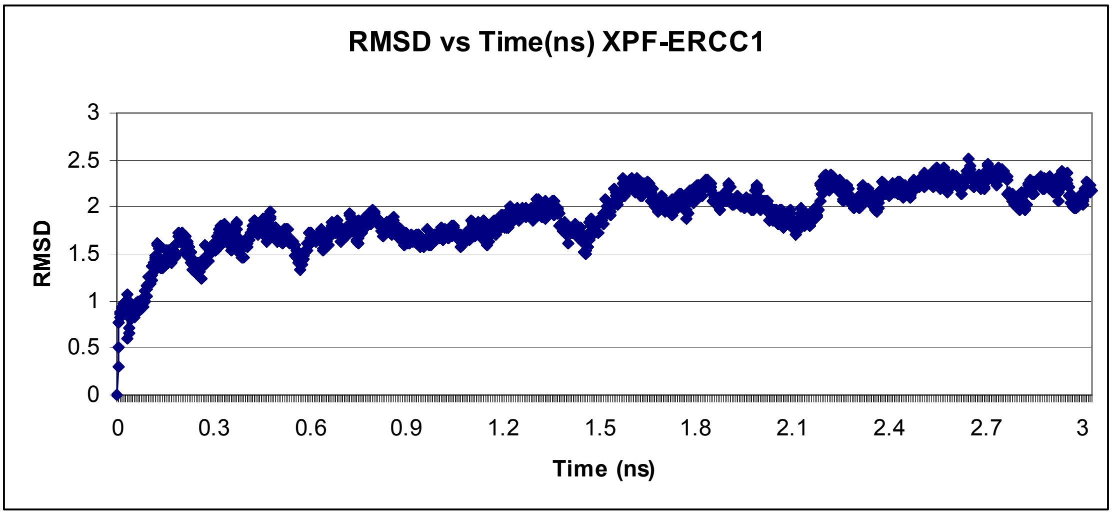


Figure 3.4: RMSD vs Time graph of MD simulation of XPF-ERCC1

Figure 3.4 shows the RMSD graph of the XPF-ERCC1 complex. The RMSD calculations include every residue in the system. Graph includes both equilibration and main simulation steps. The equilibration takes place in the first 0.35 ns time period. The first flat region between 0 ns and 0.1 ns corresponds to the minimization period where the backbone of the protein is fixed, and only hydrogen atoms are allowed to move and position themselves in a minimum energy confirmation. The second flat region between 0.1 ns and 0.3 ns corresponds to the period where, the fixed-backbone system is heated with 10 K increments slowly up to natural temperature of the cell environment. Finally, between 0.3 ns and 0.35 ns, the backbone is released slowly and the final coordinates were stored to be used in the main simulation. The XPF-ERCC1 pair shows expected solvation, heating and relaxation behavior in the equilibration steps where, no deformation is observed in the XPF-ERCC1 complex.

In the main simulation, RMSD continues to increase in the first 2 ns and did not show significant deviations for the rest of the 3 ns period, and the average RMSD of the XPF-ERCC1 in the last nanosecond is still observed to be below 2, which indicates the stabilization of the protein conformation for given temperature, pressure and solvate parameters.

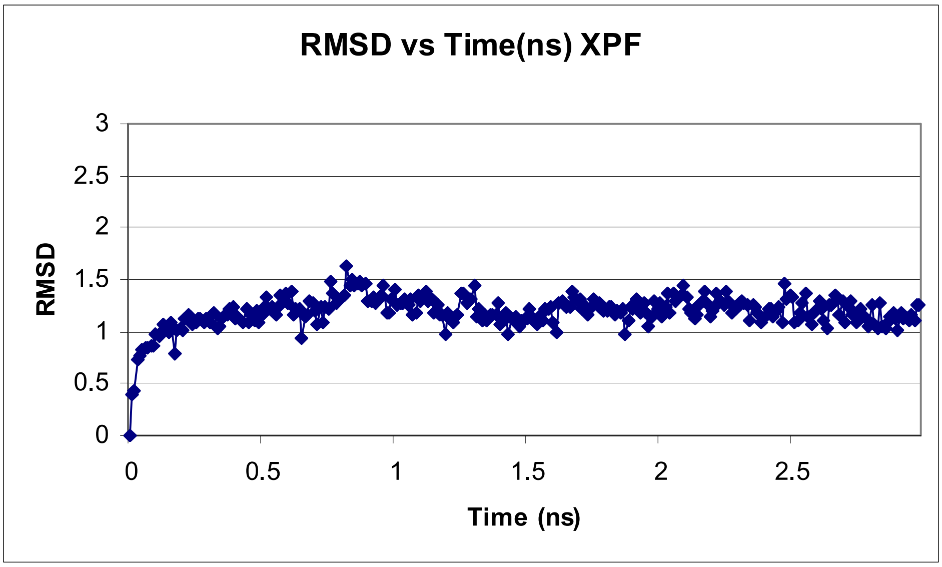


Figure 3.5: RMSD vs Time graph of MD simulation of XPF

Figure 3.5 shows the RMSD versus time graph of XPF without ERCC1. Even if during the simulation every residue of XPF is included in the conformation, the RMSD calculations are performed excluding the last 5 residues, since these residues are significantly mobile without the ERCC1, do not affect the total stability of the protein and have no interaction with the active site. Therefore including these residues might mislead the calculations increasing the total RMSD. Corresponding RMSD (root mean square deviation) versus time graphs are drawn according to the frames derived from the simulation at every 5000 time steps which has a step size of 2 ps.

XPF shows a very stable pattern during the equilibration and main simulation steps, the RMSD stabilizes about 0.7 ns after the start of the simulation and does not exceed 1.5, which indicates very stable conformation for XPF without its pair and, this behavior is significant for inhibitor design purposes. If there were instability and deformations in the structure of XPF without its pair, it would not be possible to identify an active site and build a reliable rigid biding pocket model for the docking studies to find inhibitor hits.

Next, both systems are further analyzed based on the RMSD values for every residue individually to identify the residues with relatively higher and lower mobility. There are several higher RMSD regions, first of which are the regions where loops are located. Since loops do not have hydrogen bonding patterns, they are expected to be less stable in the whole structure. The second region with higher-RMSD residues is located in the XPF-ERCC1 complex system: it is the α-helix at the N-terminus of ERCC1 (Table 3.3 and Figure 3.6). Even if the helical content of this region is very stable, the α-helix itself is mobile due to the lack of interaction with other XPF and ERCC1 domains in this truncated form. This higher RMSD region also explains the reason why the total RMSD of the XPF-ERCC1 system is higher than the total RMSD of XPF alone.

Residue	RMSD	Residue	RMSD	Residue	RMSD	Residue	RMSD
837	0.97244	873	0.857074	***228***	***3.856205***	264	0.925348
838	1.474624	874	0.87658	***229***	***4.006086***	265	1.232805
839	1.737138	875	1.552143	***230***	***3.24657***	266	1.091482
840	1.218462	876	1.247716	231	1.22363	267	0.945425
841	0.793521	877	1.057269	232	1.044539	268	2.61
842	1.743226	878	1.2983	233	1.053423	269	2.766519
843	1.95926	879	1.579303	***234***	***4.374065***	270	1.185044
844	1.10833	880	1.030455	235	0.901483	271	0.95055
845	0.94556	881	1.275519	236	0.931843	272	1.30021
846	0.893742	882	1.439879	237	1.672693	273	1.978313
847	1.470824	883	0.729136	238	1.027719	274	2.269644
848	1.990774	884	1.223785	239	0.759713	275	2.198047
849	1.481766	885	1.763694	240	0.892163	276	2.343105
850	2.603884	886	0.808898	241	1.335624	277	1.848631
851	1.395181	887	1.068154	242	1.228141	278	2.017806
852	1.159083	888	1.333195	243	2.512165	279	2.093927
853	2.590813	889	0.891589	244	2.022148	***280***	***3.100834***
854	1.119893	890	0.83187	245	1.331273	281	1.274277
855	0.72842	891	1.096792	246	1.505071	282	0.857927
856	1.462596	892	0.890156	247	1.964621	283	2.327046
857	2.013875	893	1.109738	248	1.154864	284	2.222821
858	1.527267	894	2.276545	249	1.194729	285	0.817675
859	1.054584	895	4.65894	250	0.920954	286	1.081234
860	***3.485066***	***896***	***5.909884***	251	1.189667	287	2.455806
861	1.032537	897	6.30284	252	1.072738	288	0.9545
862	0.856387	***898***	***10.17056***	253	0.989902	289	0.931309
863	0.755044	***218***	***3.621054***	254	0.915747	290	3.61926
864	1.689157	***219***	***3.954778***	255	0.957447	291	1.550737
865	1.469914	***220***	***3.288988***	256	1.20575	292	1.336126
866	1.154291	***221***	***3.63042***	257	1.069331	293	1.275643
867	0.808042	***222***	***3.597269***	258	0.872195	294	1.612018
868	1.012571	***223***	***3.928873***	259	1.465525	***295***	***6.038751***
869	1.494511	***224***	***3.966277***	260	1.628131	***296***	***6.320977***
870	1.125678	***225***	***4.255321***	261	1.312564		
871	1.116508	***226***	***4.707144***	262	1.757515		
872	2.495375	***227***	***3.733469***	263	0.994862		

Table 3.3: Residual RMSD values for XPF-ERCC1 pair

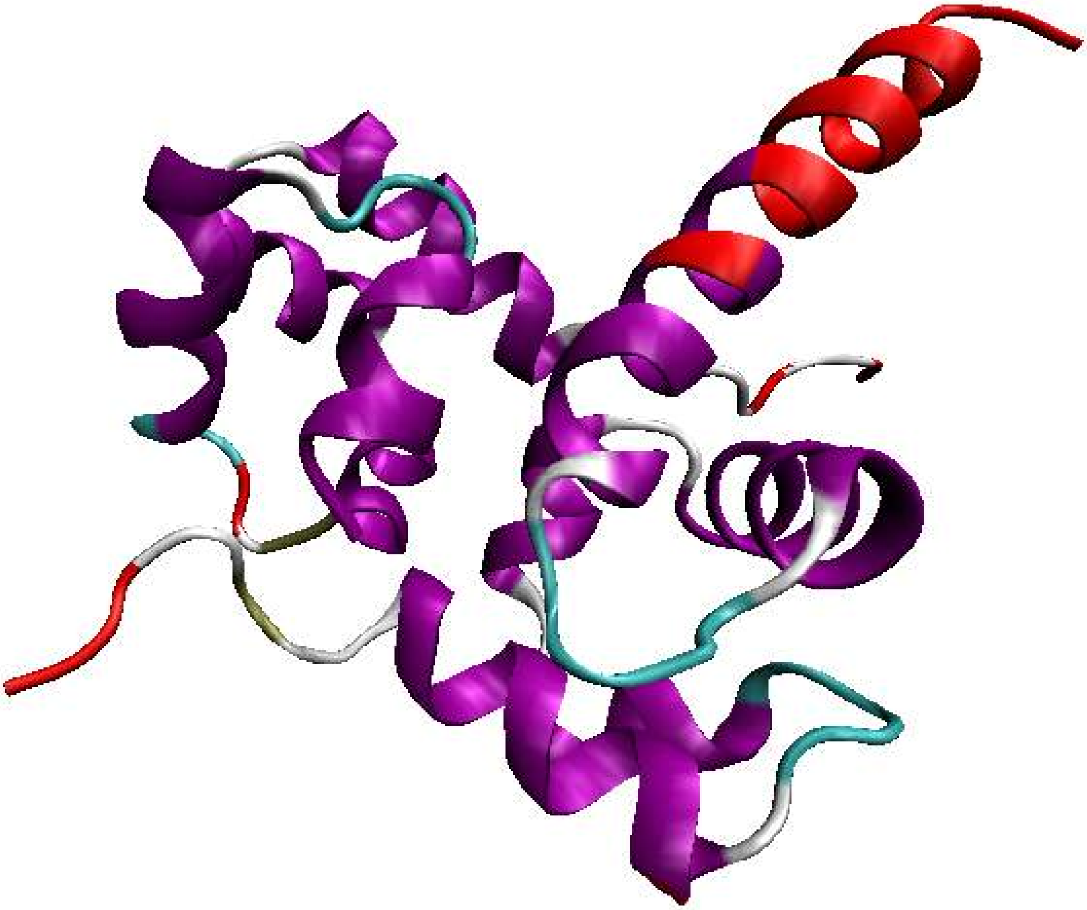


Figure 3.6: XPF-ERCC1. Mobile residues with RMSD larger than 3.

As highlighted at Figure 3.7 and Table 3.4, the higher mobility regions on XPF are located on loop regions only. The C-terminus is a high RMSD region with a loop structure, and the RMSD for residues 893 and 894 higher in the XPF simulation when their residual RMSDs are compared to the XPF-ERCC1 dimer simulation, but still the α-helix starting with residue 892 is notably rigid. The N-terminus of XPF, which immediately starts with a α-helix, is also significantly stable. Not only is the helical content of the N-terminus, but also the conformation of that helix within the domain significantly stable. This result is significant since the binding pocket includes the N-terminus helix. When the binding pocket including the significant residues 837, 841, 856, 859, 860, 861 and 862 is analyzed individually, even if the binding pocket is very rigid in its tertiary structure, the residues 837 860 and 862 in fact are higher RMSD residues (Figure 3.8). Therefore, the final structure given by the MD simulation is vital for the molecular docking studies and since Autodock works with a rigid binding pocket – flexible ligand docking model, the docking should be performed with several random conformations derived from the last nanosecond of the MD simulation.

Residue	RMSD	Residue	RMSD
837	***3.281105***	868	1.551328
838	1.786524	869	1.810141
839	1.896316	870	1.362788
840	1.660336	871	1.421635
841	1.762379	872	2.615097
842	1.868747	873	1.485832
843	2.23727	874	0.941229
844	1.365907	875	1.344696
845	1.491749	876	1.516751
846	1.865683	877	1.13054
847	1.636178	878	1.246322
848	1.86607	879	2.322775
849	1.649136	880	1.186368
850	2.308873	881	1.243518
851	2.562495	882	1.31843
852	1.501754	883	0.744464
853	2.2014	884	2.007512
854	1.902679	885	2.381284
855	1.540793	886	0.97967
856	1.466405	887	1.402827
857	2.423665	888	1.629852
858	1.612274	889	1.51638
859	2.0251	890	1.773753
860	***3.552824***	891	1.671642
861	1.875936	892	1.992681
862	***3.00747***	***893***	***3.619576***
863	2.397817	***894***	***5.751698***
864	2.196797	***895***	***6.759938***
865	1.842531	***896***	***5.803368***
866	1.320497	***897***	***5.855957***
867	1.435907	***898***	***6.779865***

Table 3.4: Residual RMSD values for XPF

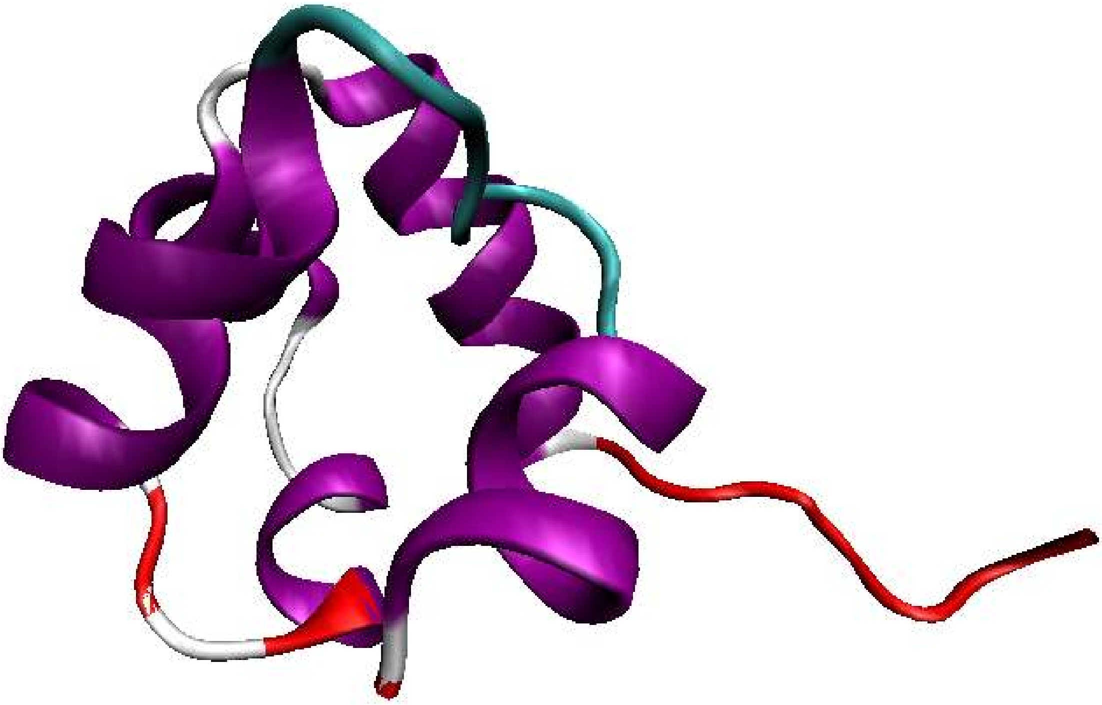


Figure 3.7: XPF. Mobile residues with RMSD larger than 3.

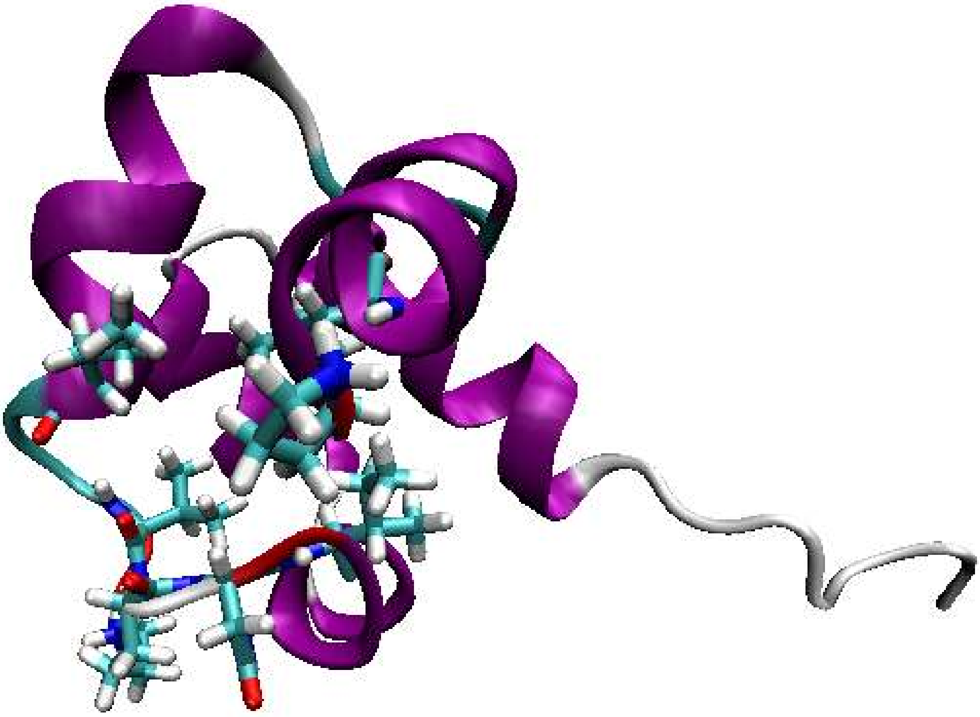
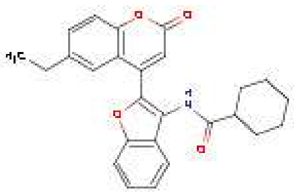
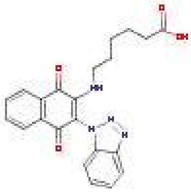
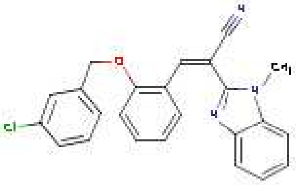
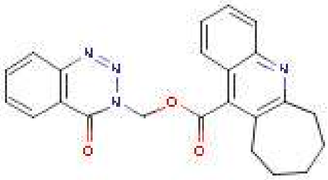


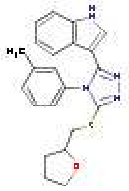
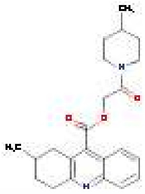
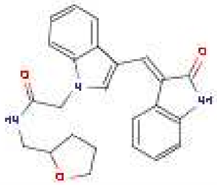
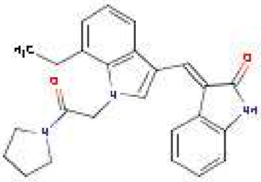
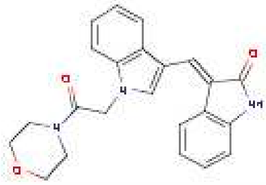
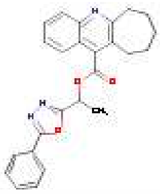
Figure 3.8: Significant residues on the binding pocket with higher RMSD residues highlighted with red color.

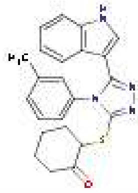
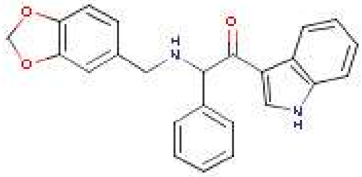
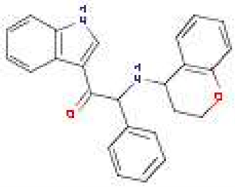
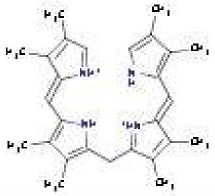
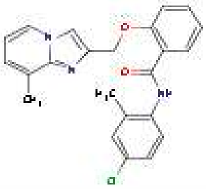
3.4 Virtual Screening and Detailed Docking of XPF

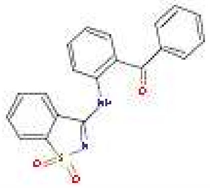
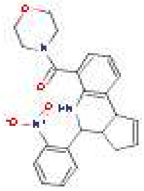
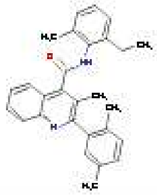
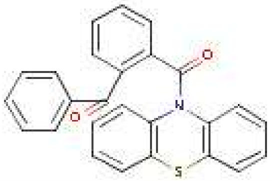
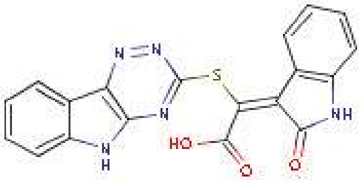
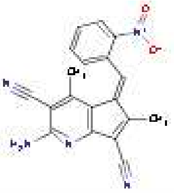
For the virtual screening purposes the AMBINTER molecule structure library of more than 1,000,000 was analyzed. As mentioned before, first these molecules were filtered to remove large molecules that would have difficulty to pass the membrane and would not fit the active site due to the size constraints and therefore would have small chance of inhibition. For XPF, the threshold was chosen to be 390 g/mol where, molecules with molecular weight larger than 390 g/mol were removed from the virtual library.

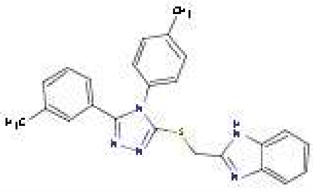
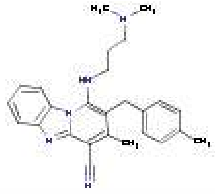
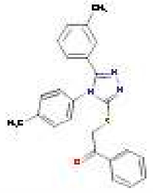
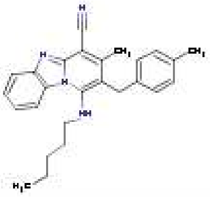
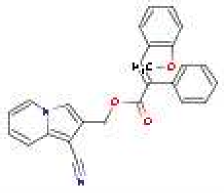
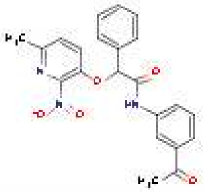
At the end of virtual screening, 270 compounds, which have binding energy score better than -10 kcal/mol were chosen for further detailed docking study. Table 3.5 shows the structures and, docking and binding energy scores of the best 28 scoring molecules at the end of detailed docking.

Molecule ID	Binding Score	Docking Score	Molecular Structure
13610_xpf	-9.28	-12.14	
119929_xpf	-9.47	-13	
111589_xpf	-10.13	-12.49	
131367_xpf	-10.46	-12.24	

158697_xpf	-9.53	-12.2	
153221_xpf	-10.01	-12.17	
116331_xpf	-9.87	-12.58	
150190_xpf	-9.92	-12.49	
123082_xpf	-9.98	-12.13	
116462_xpf	-10.49	-12.58	

15985_xpf	-9.73	-12.37	
117482_xpf	-10.57	-12.3	
134640_xpf	-10.33	-12.27	
141227_xpf	-10.01	-12.12	
126367_xpf	-9.64	-12.01	

10004_xpf	-10.01	-12.11	
146262_xpf	-10.13	-12.5	
163266_xpf	-10.21	-12.4	
141869_xpf	-10.12	-12.1	
150692_xpf	-10.47	-12.36	
10669_xpf	-10.74	-12.35	

155154_xpf	-9.57	-12.04	
146551_xpf	-9.62	-12.35	
146577_xpf	-9.16	-12.08	
132601_xpf	-9.98	-12.76	
119110_xpf	-9.67	-12.7	
117715_xpf	-9.72	-12.41	

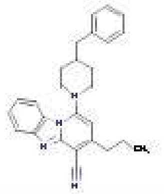
131596_xpf	-10.34	-12.37	

Table 3.5: Energies for Top Scoring Molecules in detailed Docking

Compounds shown on Table 3.5 are strong candidates for inhibition of XPF. However, further studies should be conducted to analyze the docking position of the inhibitors in the active site and the solubility behavior of the molecules to predict whether the molecules may easily be taken up by the cell or not.

In pharmacokinetics, logP, the distribution coefficient has a strong effect on the ADME (absorption, distribution, metabolism and excretion) properties of a drug molecule. Since the drug must pass through lipid bilayers during cellular transportation, for an efficient transport, the compound must be hydrophobic enough to partition in the bilayer membrane, but it should not be so hydrophobic due to the fact that when the compound is once in the bilayer it may accumulate there [53]. In pharmacodynamics, hydrophobicity is known as the main driving force of the protein-ligand binding[54, 55]. But, hydrophobic drugs are found to be more toxic since most of the time they are extensively metabolized, less selective in binding and retained longer in the body. Therefore, logP, the hydrophobicity measure, should be neither too hydrophobic nor too hydrophilic.

To evaluate drug likeness several methodologies were developed each of which refer to hydrophobicity (logP) as an important measure. Lipinski's Rule of Five suggests a partition coefficient roughly less than 5, where as Ghose et al gave a range from -0.4 to +5.6. [56, 57] Table 3.6 shows the predicted logP values for the selected 28 molecules with five different prediction algorithms and average logP values. The molecules within the drug-likeness range suggested by Ghose et al are highlighted.

	logP prediction algorithm					
Compounds	**miLogP**	**ALOGP**	**MLOGP**	**KOWWIN**	**XLOGP3**	**Average logP**
10004	-0.29	3.07	3.52	2.55	3.61	**2.96(+-1.31)}**
10669	3.72	4.1	2.72	2.19	3.85	**2.99(+-0.94)}**
13610	4.06	6.16	4.13	5.44	5.58	**5.15(+-0.69)}**
15985	3.79	4.55	4.65	4.27	3.81	**4.04(+-0.67)}**
111589	5.24	5.33	3.97	6.4	5.08	**5.40(+-0.77)}**
116331	0.12	2.43	2.35	2.96	2.63	**2.34(+-0.95)}**
116462	1.89	5.07	4.77	5.05	4.94	**4.66(+-1.15)}**
117482	3.48	3.74	2.84	3.95	3.59	**3.73(+-0.53)}**
117715	0.58	3.58	3.31	1.12	3.67	**2.57(+-1.21)}**
119110	1.74	5.47	3.24	6.04	5.6	**4.72(+-1.47)}**
119929	-1.32	1.9	1.5	-2.99	0.55	**0.90(+-1.90)}**
123082	0.14	2.4	1.79	2.61	2.23	**2.12(+-0.86)}**
126367	2.68	5.55	3.91	5.43	5.45	**4.63(+-0.98)}**
131367	4.34	5.04	3.6	5.49	4.96	**4.34(+-0.71)}**
131596	6.55	6.98	4.85	8.32	6.58	6.73(+-0.89)}
132601	7.15	7.3	4.65	8.62	7.67	6.89(+-1.03)}
134640	3.02	3.86	3.16	4.27	3.91	**3.99(+-0.65)}**
141227	7.89	8.5	3.25	8.68	5.97	6.11(+-2.07)}
141869	3.71	6.09	5.02	5.16	5.99	**5.40(+-1.14)}**
146262	-0.16	3.43	3.49	1.34	3.65	**2.68(+-1.44)}**
146551	4.88	5.47	3.94	7.47	5.08	**5.54(+-0.96)}**
146577	2.87	5.97	5.65	6.6	5.25	**5.56(+-1.18)}**
150190	1.03	4.34	3.2	5.76	3.62	**3.56(+-1.30)}**
150692	0.15	2.22	2.64	2.17	3	**2.07(+-0.86)}**
153221	1.05	4.29	3.2	5.32	4.33	**3.84(+-1.18)}**
155154	5.53	5.78	5.68	6.53	5.15	5.67(+-0.63)}
158697	4.08	4.5	4.59	4.78	3.54	**4.16(+-0.69)}**
163266	3.43	7.14	4.9	7.33	6.03	6.08(+-1.25)}

Table 3.6. Predicted logP values of XPF inhibitors.

3.5 The common motifs on the inhibitors

The inhibitors given by the virtual screening in fact show several common motifs.

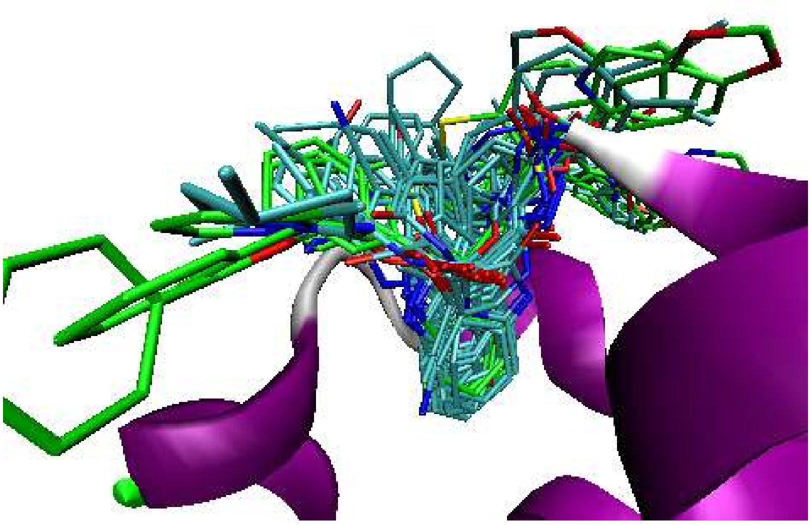


Figure 3.9. XPF inhibitors in the binding pocket

1) The ring structures in the center of the binding pocket.

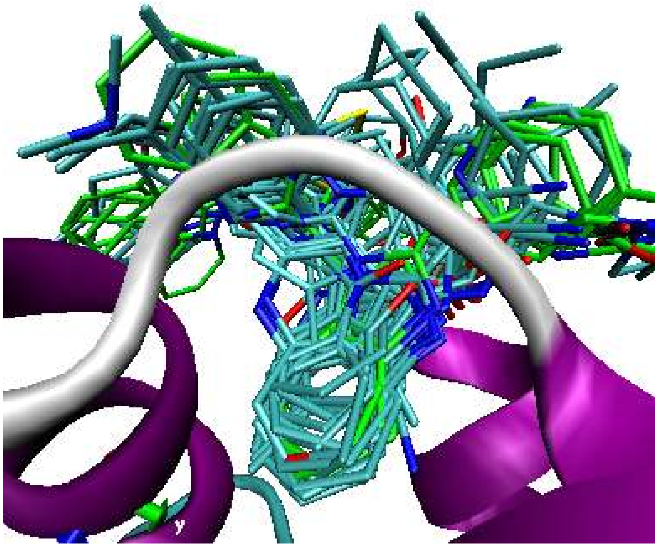


Figure 3.10. XPF inhibitors in the binding pocket

2) The confirmation of several O and N atoms on the molecules.

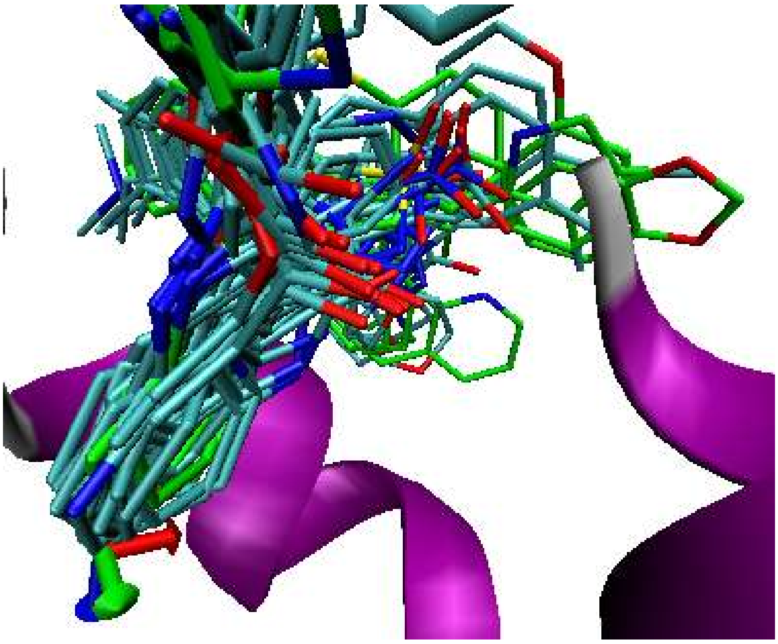


Figure 3.11. XPF inhibitors in the binding pocket

3.6 Peptide design as a competitive inhibitor to prevent XPF-ERCC1 binding

To prevent protein-protein interactions, rational design of small molecule inhibitors is a great challenge due to the existence of large interfaces with many intermolecular contacts [58]. Studies on human growth hormone-receptor complex [59] and erythropoietin receptor complex [60] showed that proteins may interact with small surface binding epitopes. These findings brought the possibility to design inhibitors to block protein-protein binding by mimicking the small binding epitopes. Recently, several studies are conducted to generalize the applicability of designing peptide drugs in developing new therapeutic strategies for many diseases and understanding protein-protein interactions [61].

In this study, to inhibit XPF-ERCC1 complex formation a smilar strategy is followed to block Phe-293 (ERCC1) by mimicing the loop region on the protein-protein interface of XPF that originally interacting with Phe-293.

3.6.1 Designing a peptide drug

Phe-293 is suggested to be the key residue in XPF-ERCC1 binding as explained in the Introduction section of this thesis and the interacting residues on XPF are reported to be Leu-841, Met-856, Val-859, Lys-860, Ile-862 Asn-861 by both experimental and computational methods. Thus, mimicking the loop region of XPF containing the above residues (841 - 862) to design a small peptide drug to block ERCC1 Phe-293 is the main approach in this study.

7 to 11 amino acid-long small peptides are designed considering the possibility that larger peptides possess several disadvantages: difficulty to deliver the target tissue, difficulty to pass through the cell membrane, possibility of aggregate formation and accumulation and, the loss of specificity. By using VMD sequence viewer tool several peptides (7, 8, 9, 11 residue long) are cut and saved directly from their original conformation on XPF. The cut peptides then docked computationally with ERCC1 to predict specificity and their stability in cell environment are studied by MD simulations. For docking studies Rosetta is utilized and for the MD simulation NAMD is used to build a 3-ns simulation with periodic boundary conditions.

Finally, according to the deformation scale of the peptides during the MD simulation and the active site specify prediction after docking studies several mutations are made to prevent deformation and enhance specificity. Hydrophobic residues on the tails of the peptides are replaced with hydrophilic ones to improve stability and residues with appropriate side chain charges are replaced neutral residues to improve specificity.

3.6.2 Detailed analysis of the designed peptides

3.6.2.1 Peptide-7 (856-862)

Figure 3.12 shows the docking scores versus RMSD graph of the top 1000 scoring conformations. The data points on the graph indicate that the positions of the top 10 scoring decoys are not necessarily the minimum RMSD positions. Thus, the designed Peptide-7 may perform poor selectivity and binding. Figure 3.13 also shows the visualization of the top 10 scoring conformations and only 2 of the 10 top scoring conformations are actually predicted to bind to the desired position (marked with red color in Figure 3.13).

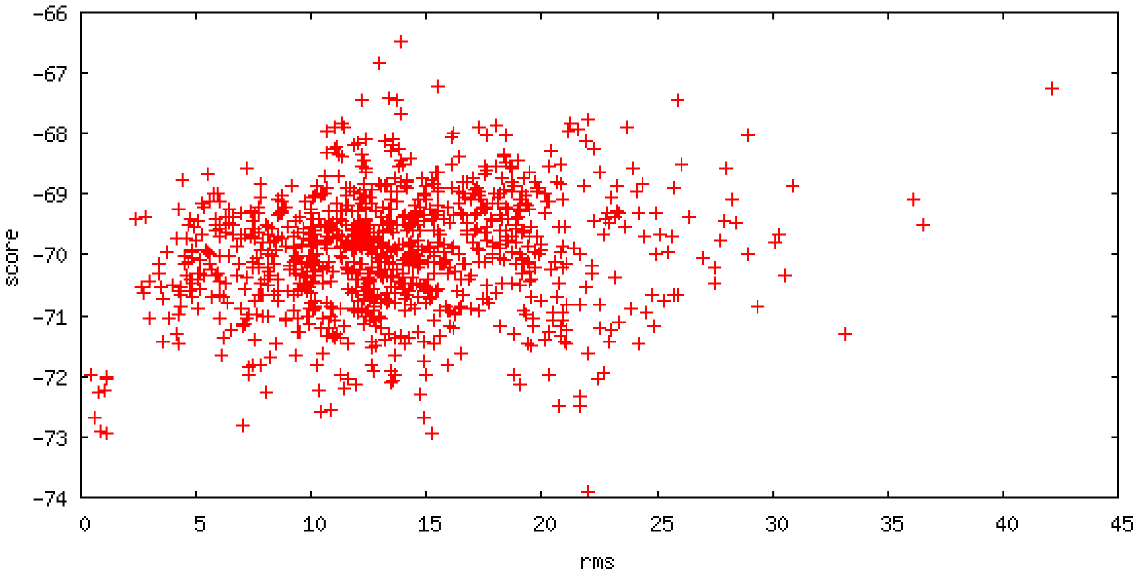


Figure 3.12 : Docking score vs. RMSD graph of Peptide-7

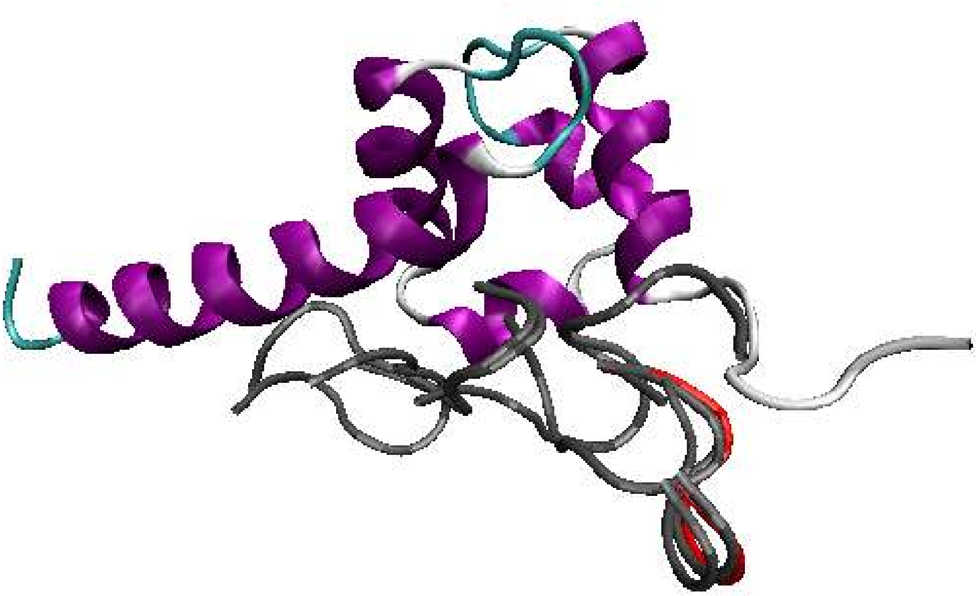


Figure 3.13: Positions of the Top 10 peptides with the reference position

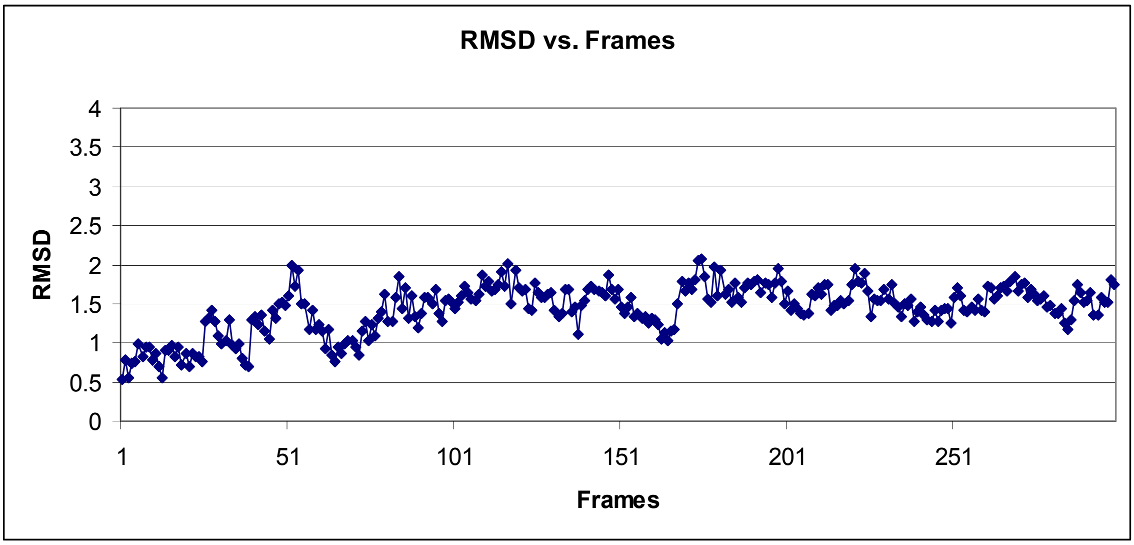


Figure 3.14: RMSD vs Frames graph for Peptide-7

Residue	Average RMSD
856	3.364
857	2.085
858	2.200
859	1.255
860	2.431
861	1.737
862	2.839

Table 3.7: Average RMSD of each residue through the simulation

The RMSD graph of Peptide-7 versus simulation time is shown in Figure 3.14 The total RMSD of the peptide is below 2 Å through the simulation and, this is the indicator of significant stability of the peptide in cellular environment. Table 3.7 shows the average RMSD values of each residue in the peptide and this allows us to study the stability of the amino acids individually. For a 7-residues-long small peptide the reported average RMSD values are significantly low suggesting a very stable structure with small deviations on the tails.

3.6.2.2 Peptide 8 (856-863)

Figure 3.15 shows the docking scores versus RMSD graph of the top 1000 scoring conformations. The data points on the graph indicate that the positions of the top 10 scoring decoys are strictly the minimum RMSD positions. Thus, the designed Peptide-8 is expected to perform significant selectivity and binding. Figure 3.16 also shows the visualization of the top 10 scoring conformations and all of the 10 top scoring conformations are actually predicted to bind to the desired position (marked with red color in Figure 3.16).

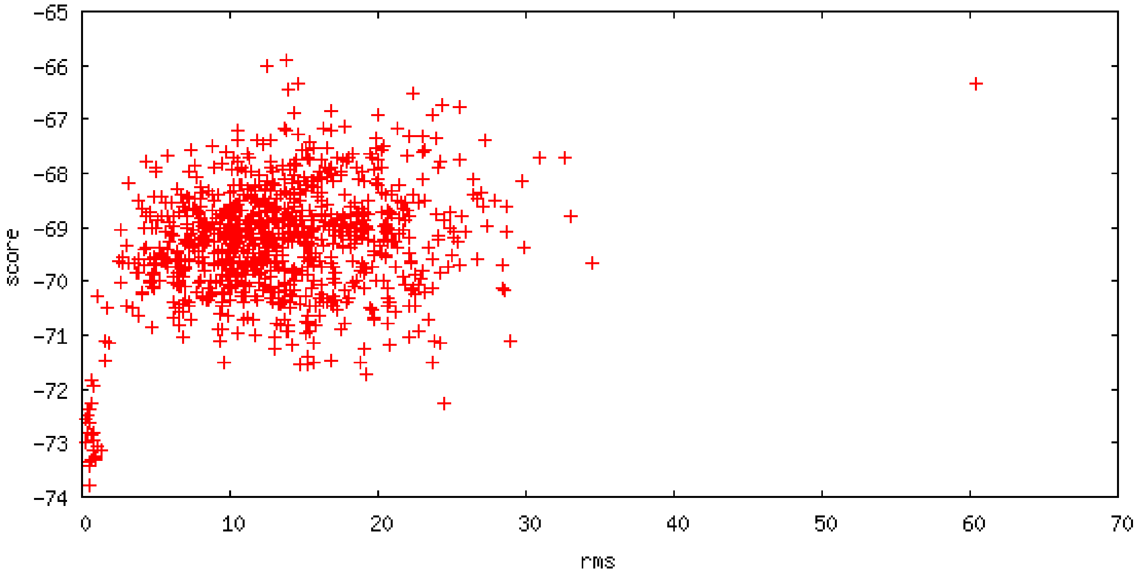
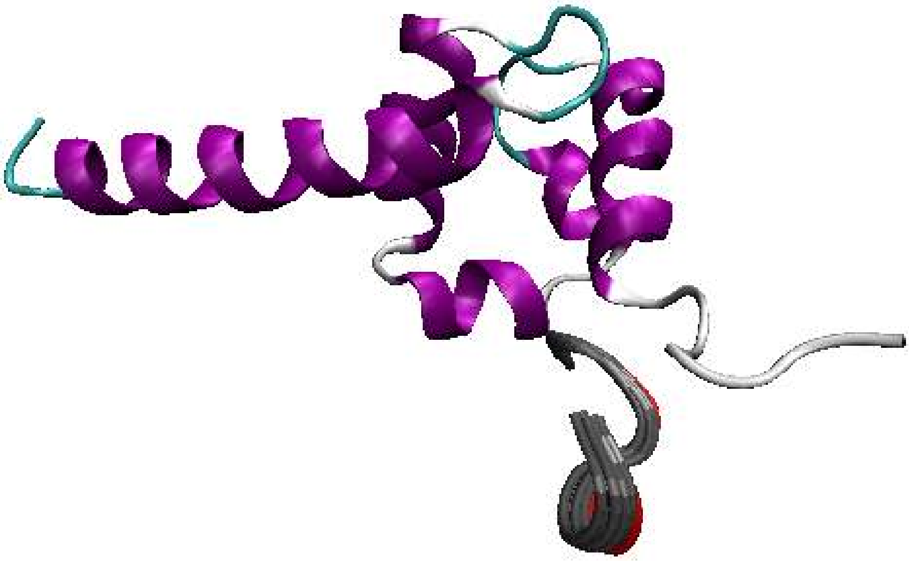


Figure 3.15: Docking score vs. RMSD graph of Peptide-8



3.16: Positions of the Top 10 peptides with the reference position

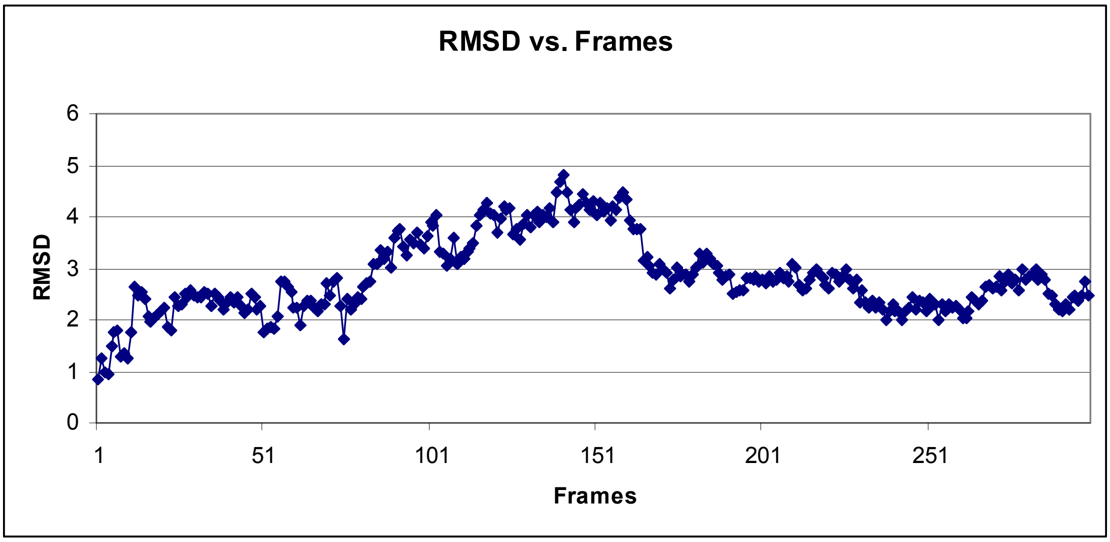
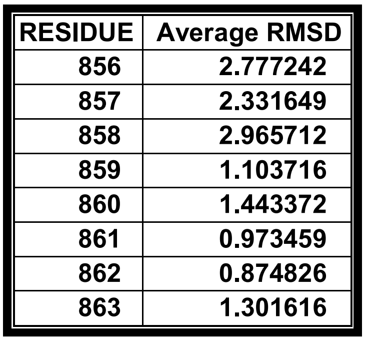


Figure 3.17: RMSD vs Frames graph for Peptide-8

RESIDUE	Average RMSD
856	6.195
857	4.088
858	5.702
859	3.253
860	3.083
861	2.399
862	2.706
863	4.105

Table 3.8: Average RMSD of each residue through the simulation

The RMSD graph of Peptide-8 versus simulation time is shown in Figure 3.17. The total RMSD of the peptide in the last nanosecond of the simulation is below 3 Å but, during the second nanosecond of the simulation the peptide shows a strange behavior where, several deformations occur in the tail sections, but later the peptide stabilizes in the last nanosecond of the simulation. The last nanosecond of the simulation with RMSD below 3 Å is an indicator of stability of the peptide in cellular environment. Table 3.8 shows the average RMSD values of each residue in the peptide and this allows us to study the stability of the amino acids individually. For the 8-residues-long small peptide the reported average RMSD values are below 4 Å for four residues and higher than 4 Å for the rest of the residues in the tail section. The results of the molecular dynamics simulation suggest a moderately stable structure with deviations on the tails. The main reason for the increase in the average RMSD is the high RMSD region during the second nanosecond of the simulation. We calculate the residual RMSD values after the simulation reached the equilibrium (after frame 170), and the average RMSD scores are presented in Table 3.9. It can be argued that when the simulation reached the equilibrium, Peptide-8 shows significant stability with all of the residues have average RMSD scores below 3 Å.



RESIDUE	Average RMSD
856	2.777242
857	2.331649
858	2.965712
859	1.103716
860	1.443372
861	0.973459
862	0.874826
863	1.301616

Table 3.9: Average RMSD of each residue after the system reached the equilibrium

3.6.2.3 Peptide 9 (855-863)

Figure 3.18 shows the docking scores versus RMSD graph of the top 1000 scoring conformations. The data points on the graph indicate that the positions of the top 10 scoring decoys are strictly the minimum RMSD positions. Therefore, the designed Peptide-9 is expected to perform significant selectivity and binding. Figure 3.19 shows the visualization of the top 10 scoring conformations (marked with gray color in Figure 3.19) and all of the 10 top scoring conformations are actually predicted to bind to the desired position (marked with red color in Figure 3.19).

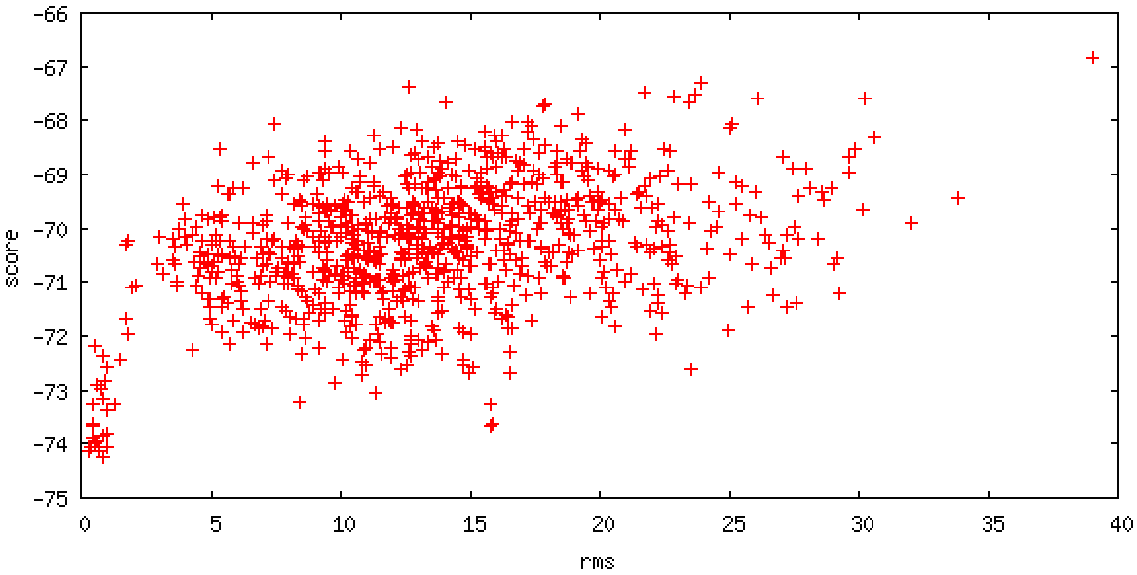


Figure 3.18: Docking score vs. RMSD graph of Peptide-9

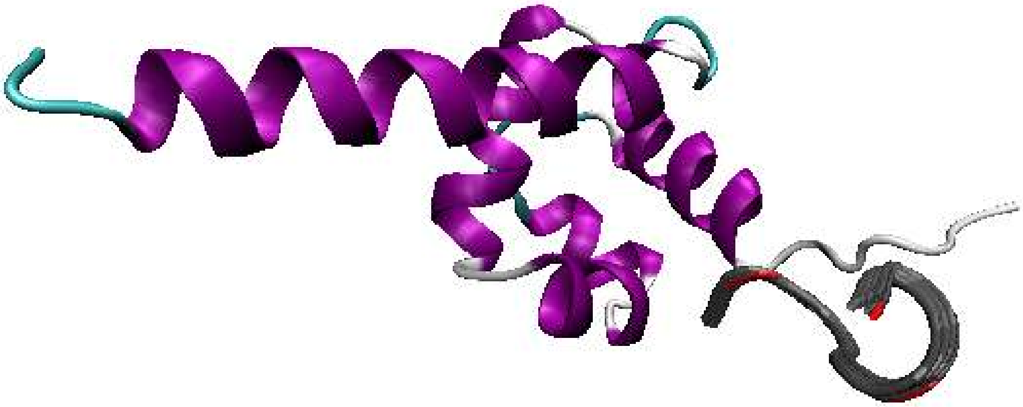


Figure 3.19: Positions of the Top 10 peptides with the reference position

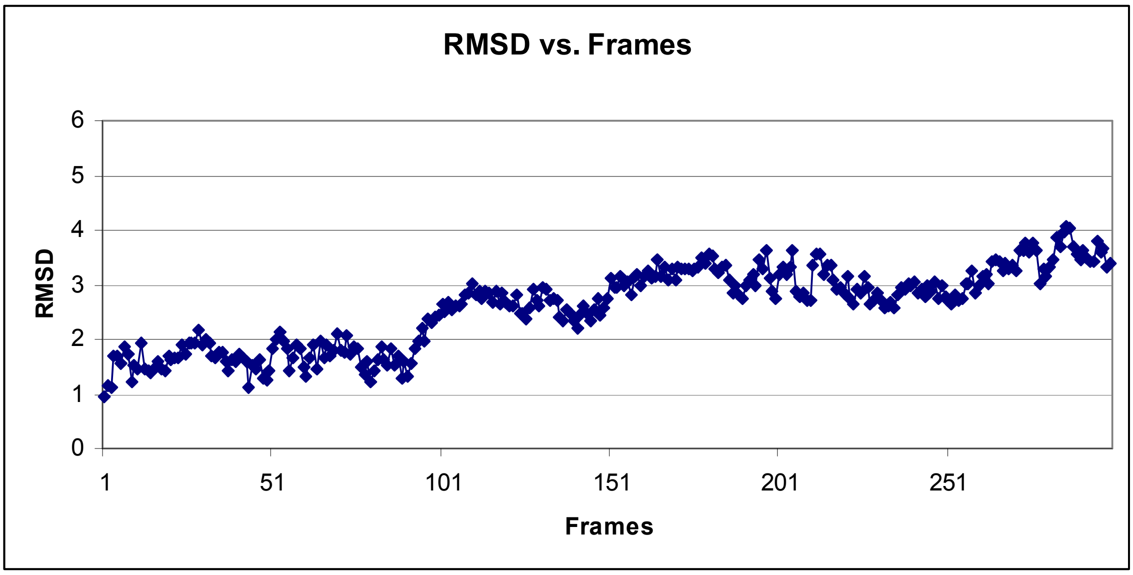


Figure 3.20: RMSD vs Frames graph for Peptide-7

RESIDUE	Average RMSD
855	4.173
856	4.557
857	3.348
858	5.501
859	2.477
860	4.055
861	2.489
862	3.755
863	3.673

Table 3.10: Average RMSD of each residue through the simulation

The RMSD graph of Peptide-9 versus simulation time is shown in Figure 3.20. The total RMSD of the peptide is below 4 Å through the simulation and, this is the indicator of stability of the peptide in cellular environment. Table 3.10 shows the average RMSD values of each residue in the peptide and this allows us to study the stability of the amino acids individually. For a 9-residues-long peptide the reported average RMSD values for the tail residues are significantly low suggesting a very stable tail structure but, residue 858 in the center of the peptide show small deviations on with average RMSD above 5 Å.

3.6.2.4 Peptide 11(854-865)

Figure 3.21 shows the docking scores versus RMSD graph of the top 1000 scoring conformations. The data points on the graph indicate that the positions of the top 10 scoring decoys are not necessarily the minimum RMSD positions. Thus, the designed Peptide-11 is predicted perform poor selectivity and binding. Figure 3.22 shows the visualization of the top 10 scoring conformations and only 3 of the 10 top scoring conformations (marked with gray color in Figure 3.22) are actually predicted to bind to the desired position (marked with red color in Figure 3.22).

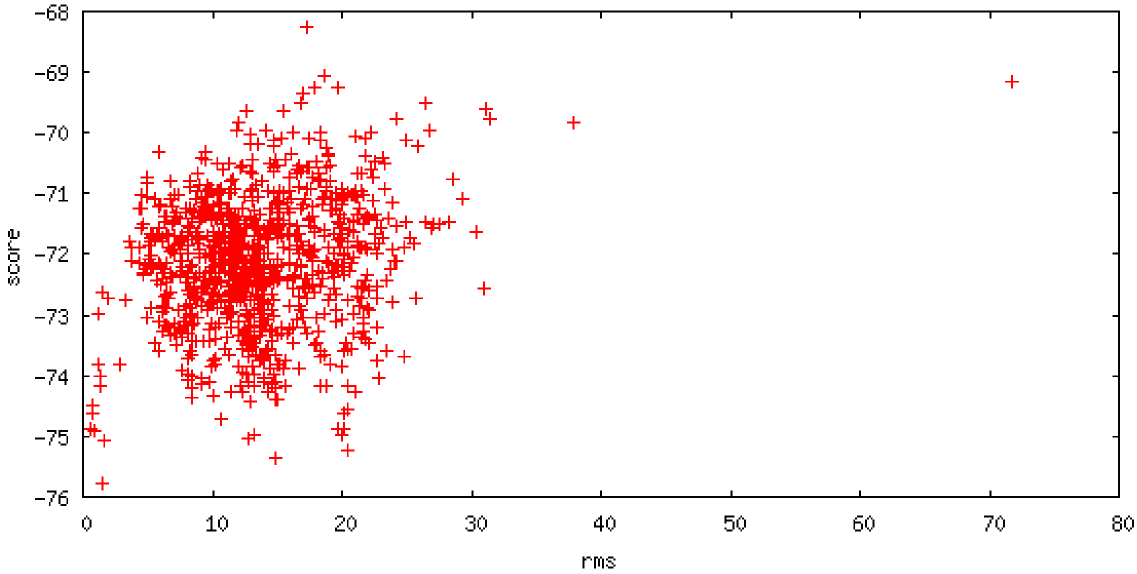


Figure 3.21: Docking score vs. RMSD graph of Peptide-11

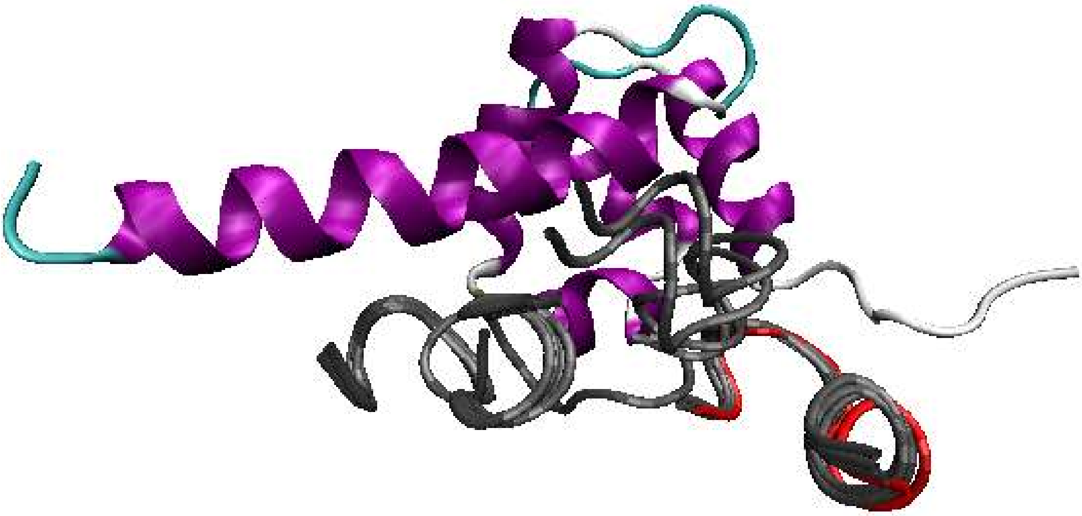


Figure 3.22: Positions of the Top 10 peptides with the reference position

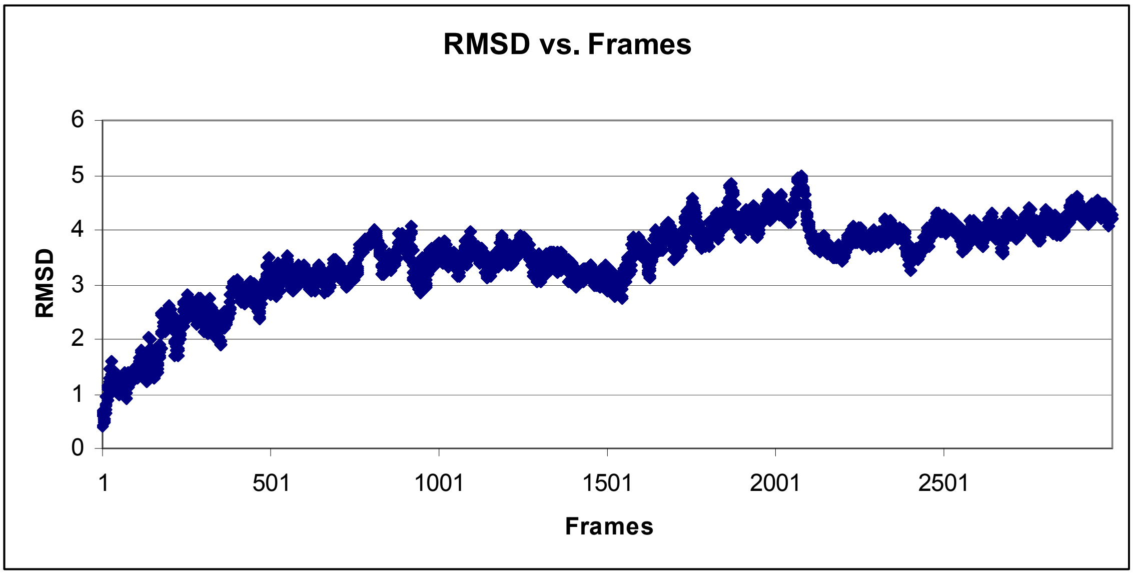


Figure 3.23: RMSD vs Frames graph for Peptide-11

RESIDUE	Average RMSD
854	6.919
855	7.034
856	5.719
857	5.796
858	5.911
859	5.139
860	5.252
861	5.230
862	5.517
863	5.243
864	6.718
865	7.014

Table 3.11: Average RMSD of each residue through the simulation

The RMSD graph of Peptide-11 versus simulation time is shown in Figure 3.23. The total RMSD of the peptide is below 5 Å through the simulation and, this is the indicator of moderate total stability of the peptide in cellular environment. Table 3.11 shows the average RMSD values of each residue in the peptide and this allows us to study the stability of the amino acids individually. For the 11-residues-long small peptide the reported average RMSD values are significantly high suggesting serious deformations of the structure.

3.6.2.5 Peptide8modified(Ala-863 to Glu-863)

Figure 3.24 shows the docking scores versus RMSD graph of the top 1000 scoring conformations. The data points on the graph indicate that the positions of the top 10 scoring decoys are not necessarily the minimum RMSD positions. Thus, the designed Peptide-8modified may perform poor selectivity and binding. Figure 3.25 also shows the visualization of the top 10 scoring conformations and 6 of the 10 top scoring conformations are actually predicted to bind to the desired position (marked with red color in Figure 3.25).

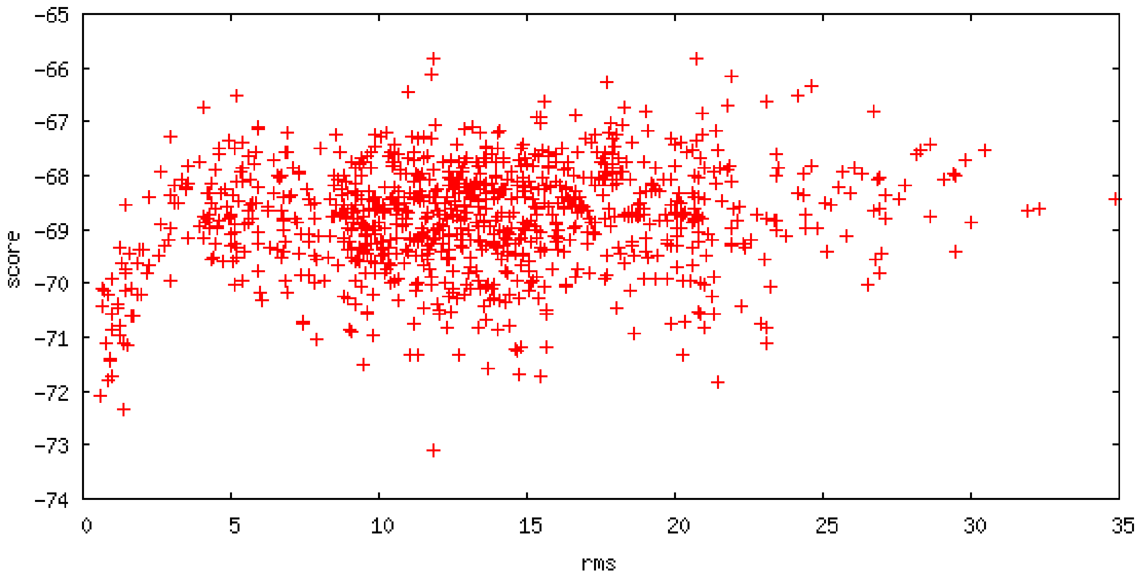


Figure 3.24: Docking score vs. RMSD graph of Peptide-8 modified

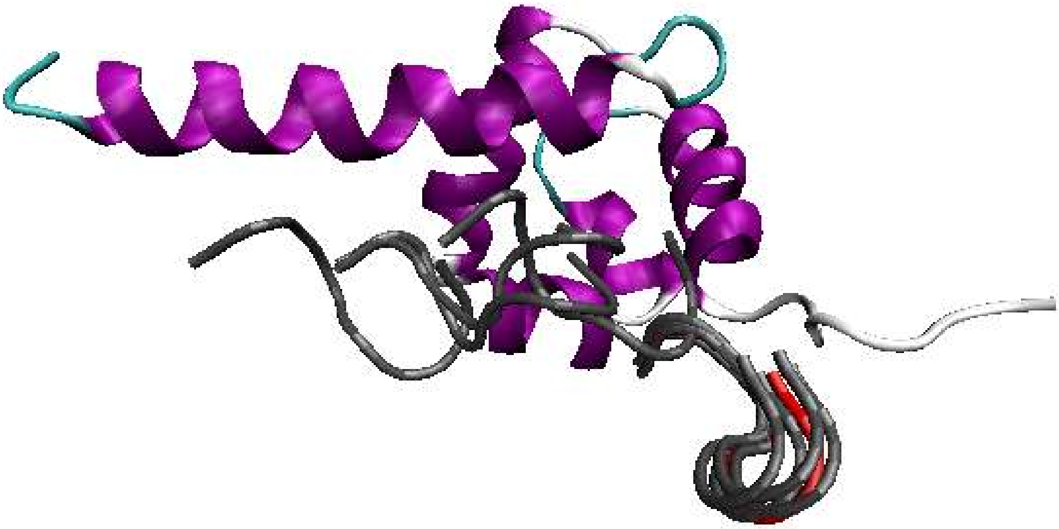


Figure 3.25: Positions of the Top 10 peptides with the reference position

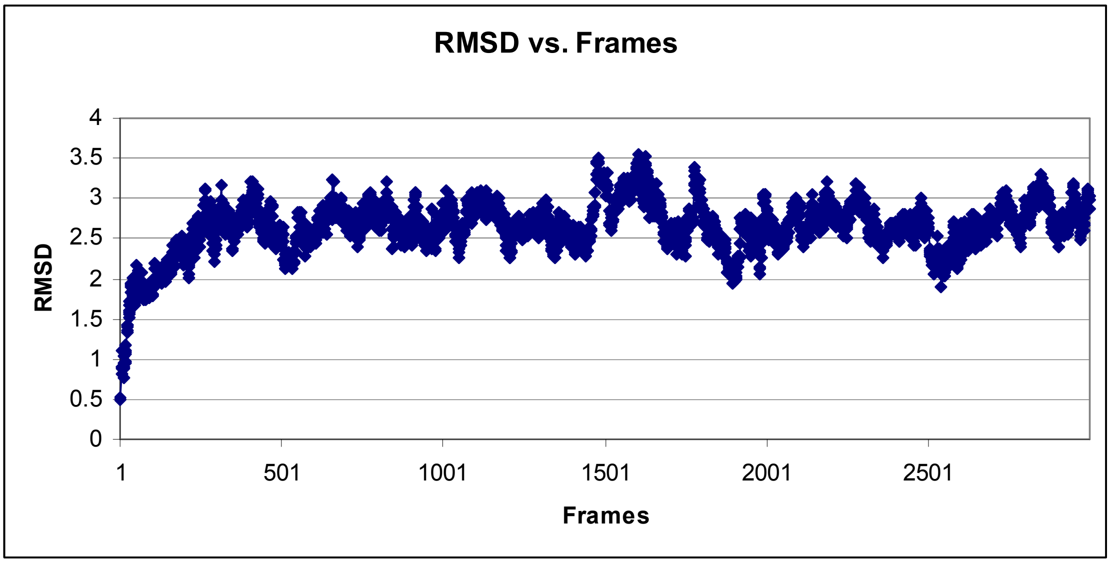


Figure 3.26: RMSD vs Frames graph for Peptide-8modified

RESIDUE	Average RMSD
856	6.387
857	3.173
858	4.660
859	2.971
860	4.051
861	3.440
862	4.010
863	5.859

Table 3.12: Average RMSD of each residue through the simulation

The RMSD graph of Peptide-8modified versus simulation time is shown in Figure 3.26. The total RMSD of the peptide is below 4.5 Å through the simulation suggesting moderate stability of the peptide in cellular environment. Table 3.12 shows the average RMSD values of each residue in the peptide and this allows us to study the stability of the amino acids individually. For an 8-residues-long small peptide the reported average RMSD values are low in the core section suggesting a stable core structure with deviations on the tails.

The designed peptides were studied in detail considering two main characteristic, the first was the selectivity, which was analyzed by peptide docking studies and, the second was stability, which was analyzed by MD simulation studies. In this study, only 5 of the several designed peptides are mentioned. Peptide-7 showed significant stability with only small deviations in the tail sites but rather moderate selectivity, Peptide-8 showed significant stability and also very selective binding, Peptide-9 also showed very good stability and very selective binding, Peptide-11 showed moderate stability with rather moderate selectivity, and the mutated peptide Peptide-8modified showed high stability with good selectivity. Thus, these peptides were chosen to be further analyzed experimentally in future studies.

Chapter 4

Cytochrome P450 C17

4.1 Introduction

Prostate cancer (PC) is the most common cancer type among men in many countries [62]. Although it is very improbable to develop PC before the age of 40, 1 out of 6 men will be diagnosed with invasive PC throughout their lifetime. 90% of the patients, who is diagnosed with PC, respond to androgen deprivation [63]. Therefore, suppressing androgen biosynthesis is an important alternative strategy for the treatment of PC. If androgen synthesis can be inhibited using CYP 17 inhibitors or combining the usage of these inhibitors with other treatments, it is possible to reduce the side effect of the other treatments (chemotherapy, surgical removal of testicles or prostate and hormonal therapy) [64].

Testosterone (TESTO) and dihydrotestosterone (DHT) are the most important androgens that are related to PC [65]. Androgen biosynthesis requires the participation of three tissues; adrenal glands, testis and prostate. Pathway starts with cholesterol, and several enzymes catalyzing the reactions leading to TESTO and DHT. CYP17 catalyzes conversion of progesterone to androstenedione (4-DIONE); 17-β HSD-3 is responsible for TESTO formation; and, finally, 5-α reductase converts TESTO to more potent androgen DHT. Ample evidence indicated that CYP17 catalyze the rate limiting step in androgen biosynthesis. Precursors of androgenic hormones, 4-DIONE and dehydro-epi-androsterone (DHEA), can be formed only by CYP17. Use of CYP17 as a drug target can improve blockage of androgen biosynthesis and inhibitors can be used as effective PC treatments [64].

The crystal structure of CYP17 has not been solved yet. Since CYP17 is a membrane bound protein, it has been difficult to crystallize. On the other hand, there are studies on computer-generated models for CYP 17. Laughton *et al.* [66] build a model for CYP17, and Lin *et al.* [67] modeled the active site of the protein. Both of these models were based on the crystal structure of P450cam, a class I P450. Lin *et al.* defines a bi-lobed substrate binding pocket [67]. The more recent model by Auchus *et al.* [68](PDB ID: 2C17) is based on a class II P450 crystal structure, P450BMP.

CYP17 is a well-recognized target for prostate cancer treatment, since selective inhibition of the enzyme exerts control over androgen synthesis. Therefore, many experimental and computational studies were reported on inhibitors of CYP17. First inhibitor designs used PREG and PROG as feeds since no other inhibitors were reported. Due to use of substrates, first generation of designed compounds was steroid based molecules with various side chains attached to 17th carbon. Even though some of these compounds showed inhibitory effects, they were not promising candidate drugs. The drawbacks of steroidal compounds were poor acid-stability, poor bioavailability, short half-life, first-pass effects and poor selectivity [64]. Ketoconazole is an antifungal agent known to reduce androgen levels in human, and has inhibitory effect on CYP17. However, it has been removed from use because of liver toxicity and its effects on other cytochrome enzymes [69]. The steroidal compound Abiraterone passed phase II clinical trials and reported to have no dose limiting toxicity [70]. Ideyama *et al.* worked on a non-steroidal inhibitor (YM116), which is able to inhibit both 17α- hydroxylase and 17-20 lyase reactions catalyzed by CYP17 with IC_{50} values lower than ketoconazole. This non-steroidal compound was 50 fold more effective in inhibiting 17-20 lyase reaction than 17α- hydroxylase reaction [71]. Nnane *et al.* presented novel steroid-based inhibitors of CYP17. The IC_{50} values for five steroid-based compounds were determined for CYP17 and 5α- reductase. Molecules L-6 and L-26 showed more potent inhibition than ketoconazole. Despite their problems in bioavailability, these compounds were found to be promising as potential agents for reducing levels of testosterone and dehydro- testosterone in patients with androgen dependent diseases [72]. Effects of a number of thiazolidinediones were reported by Arlt *et al.* Thiazolidinedione and biguanide drugs, which are used to increase insulin sensitivity in type 2 diabetes, lower serum androgen concentrations in women with polycystic ovary syndrome. In order to determine if this is a secondary effect or direct control over androgen biosynthesis, these compounds were subjected to activity assays. Thiazolidinediones were found to exert inhibition on both reactions catalyzed by CYP17, but not biguanides [73]. In another study, the effect of cinnamic-acid based derivatives of thiazolidinediones on CYP17 was analyzed. Although these studies do not focus on the treatment of prostate cancer, some of the reported compounds have shown inhibition on both reactions catalyzed by CYP17[74]. C-17-Heteroaryl steroidal compounds were rationally synthesized and tested for inhibitory and antitumor effects by Handratta *et al.* Some of these benzoazoles and pyrazines were found to be potent inhibitors of CYP17 as well as being antagonists of androgen receptors [75]. The work of Clement *et al.* uses steroid-based inhibitors and generates a pharmacophore model of human CYP17 inhibitors [69]. This model is used to retrieve hits from Maybridge, ACD and BioByteMasterFile chemical databases. Hartmann *et al.* also published a review on inhibition of CYP17 by steroidal and non-steroidal molecules as a method for androgen-dependent prostate cancer treatment [64]. Recently reported novel non-steroidal substrate mimetics reported to showed good inhibition values with good selectiviy against CYP3A4 but also showed moderate to high inhibition activity against other hepatic CYP enzymes [76, 77]. A very detailed review on CYP 17 inhibitors were also recently published by Moreira *et al.* [78].

Two compounds (VN/124-1 [75] and abiraterone [79]) were reported to be promising CYP 17 inhibitors that are tried in preclinical and clinical trials respectively. However; these two compounds are steroidal compounds, thus a successful design of non-steroidal, specific, non-toxic CYP 17 inhibitor is still an intriguing area of research. The available non-steroidal compounds were mainly designed based on mimicking known steroids that are interacting with the CYP 17 active site (i.e. progesterone, abiraterone), pharmacophore modeling (based on QSAR analysis of known inhibitors) and creating derivatives of known inhibitors of cytochrome enzymes such as: antifungal imidazole containing agents related to ketoconazole, pyridine derivatives, xanthone derivatives carbazole derivatives. The non-steroidal compounds reported in the literature did not yet show promising results in clinical trials.

In this thesis, structure-based drug design approach based on the model protein structure by Auchus *et al.* [68] was successfully applied to identify novel CYP 17 inhibitors in silico. Further experimental tests proved inhibitory activity of two novel lead compounds against CYP 17 in an HEK 293 T cell line. The leads were also tested on HeLa cell line for toxicity and the non-steroidal “lead compound” does not display toxic effects.

4.2 MD Simulation for CYP17

In this study, the output files of the MD simulation previously performed by Muhittin Emre Özdemir are used for further studies concerning CYP17 [78]. The equilibration system is sampled every 2 ps and frames are used for both root mean square deviation (RMSD) and deformation analysis. On Figure 4.1, RMSD is plotted through the simulation beginning from the end of the equilibration phase, where initial model structure is used as reference to compare generated structures. After all the restraints are released, RMSD initially increased for 2 ns and reached equilibrium for the rest of the 10 ns simulation (Figure 4.1). This behavior of RMSD is an indication of stabilized protein in simulation. When RMSD is analyzed on residue level, it can be observed that there are regions of high and low deviations. High RMSD spots are concentrated in regions where there is no strict structure; in other words, where loops are located. On the other hand, none of the alpha-helices or beta-sheets coincides with high RMSD spots. Most importantly, the I-helix passing through the core of the protein and involved in active site is in a low RMSD region. Since this helix bears the catalytically active amino acid, it is crucial to have it in a stable form. MD simulation performed on CYP17 has proved the stability of active site.

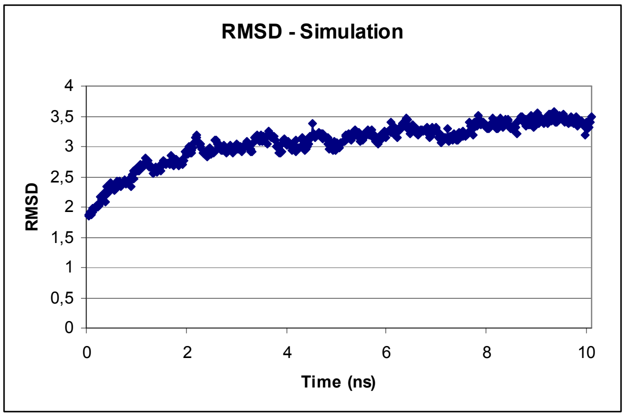


Figure 4.1: RMSD results of the MD simulation of CYP17 (pdb id: 2c17).

4.3 P450 C17 inhibitors

18 molecules were previously designed and analyzed in the MS Thesis of Muhittin Emre Özdemir [78]. N15 (non-steroidal) and S3 (steroid-based), are further characterized with inhibition test at various concentrations (Figure 4..2, 4.3). IC_{50} values for these two different compounds were determined using transformed HEK 293T cell line by AARA. For both of these compounds, concentrations covering inhibition pattern are found and shown in Table 3. IC_{50} value for N15 is 35.65 µM and for S3, IC_{50} is 46.30 µM.

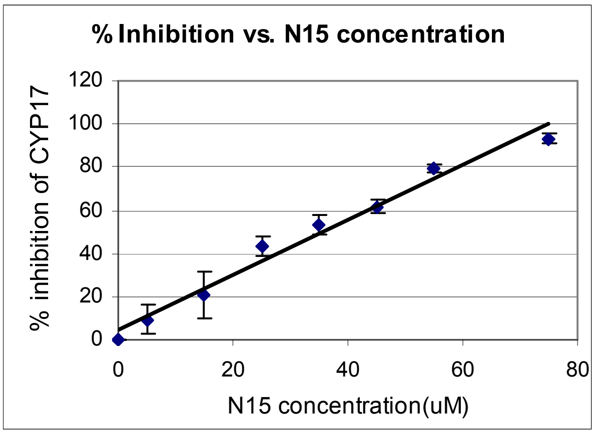


Figure 4.2: % Inhibition VS Inhibitor concentration graph for N15. Trend line is fitted and the corresponding equation is utilized to calculate the IC_{50} of N15.

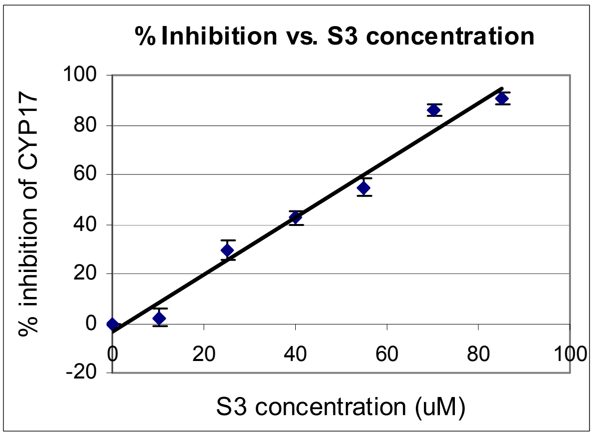


Figure 4.3: % Inhibition VS Inhibitor Concentration graph for S3. Trend line is fitted and the corresponding equation is utilized to calculate the IC_{50} of S3.

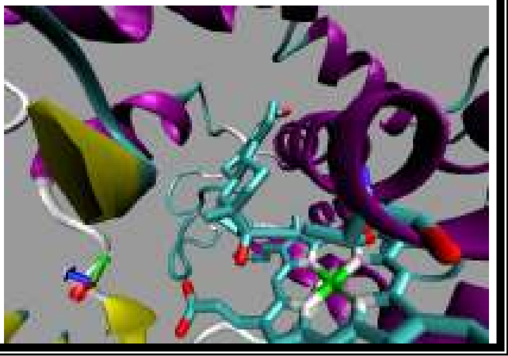
Docking Configurations for Candidate Molecules	
N15 	S3 

Table 4.1: Docking Configurations for Candidate Molecules

To calculate IC_{50} values, first by nonlinear regression a trend line is fitted, then the corresponding equation of the trend line is calculated, and the IC_{50} value is obtained. R^2 represents the goodness of the fit measure. Since the R^2 values for our calculations are above 0.97, the trend line represents our experimental data significantly well. Docking conformations of N15 and S3 are presented on Table 4.1. Then, cell viability assays were performed for N15 and S3 to study toxicity. Toxicity of a drug candidate is as important as its inhibition potential; therefore, cell viability assays are performed for N15 and S3. Results are shown in Figure 4.4 and Figure 4.5.

According to results of viability assay, TC50 values are calculated. TC50 represents the concentration of a test compound at which 50% of the cells are killed. For lead compounds, toxicity values lower than 25 µM are not favored [80]. The steroid based compound has a calculated TC50 value about 38 µM for a 24 hours assay and 68 µM for the 12 hours assay. Therefore, S3 is still a promising candidate for further optimization studies. On the other hand, N15 has fairly good viability scores above its IC_{50} value. The TC50 for N15 is calculated to be 271 µM for the 12 hours assay and 397 µM for the 24 hours assay, which is very promising. Therefore, it may used for further optimizations to improve inhibition efficiency and reduce toxicity. Note that all calculations are based on the fitted trade lines of the survival curves.

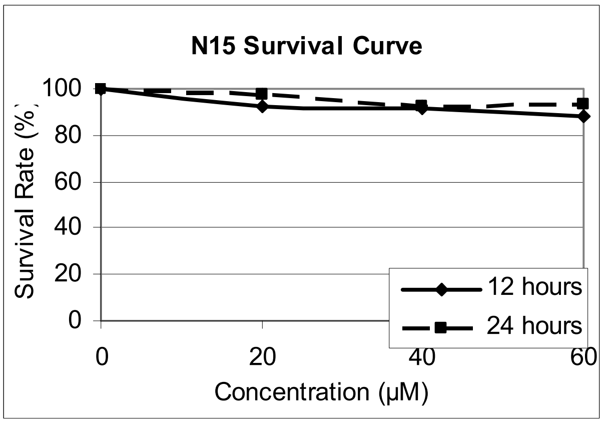


Figure 4.4: Cell Viability Assay Results for N15.

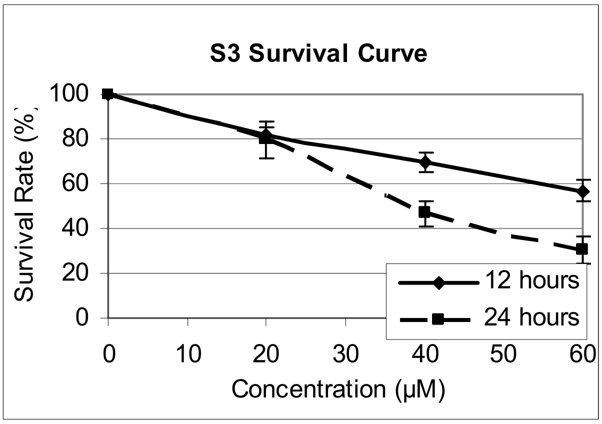


Figure 4.5: Cell Viability Assay Results for S3.

4.4 N15 derivatives

In designing the derivatives we aim to improve the binding strength of the inhibitor, while keeping the toxicity measure at minimum. Therefore; the docking position of N15 and the probable atomic interactions between N15 and the side chains of the active site residues are carefully studied. Three main interaction sites are determined in the binding pocket, which are illustrated in Table 4.2 with a representative N15 inhibitor derivative. First, as shown in Table 4.2.a the interacting side chains are neutral (Alanine and Valine), thus it should be beneficial to replace the oxygen atoms located on the double-ring with carbon atoms. Second, as shown in Table 4.2.b the positively charged iron atom at the center of the Heme may form strong non-bonded interactions or covalent bonds with a well positioned oxygen or nitrogen of the inhibitor, therefore; several derivatives are designed with accessible oxygen and nitrogen atoms in their central region. Third, the long hydrocarbon tail of the inhibitor is predicted to interact with the I-helix in the active site (Table 4.2.c) and surrounded by neutral side chains (Alanine), thus there should not be any replacements on the tail site of the inhibitor.

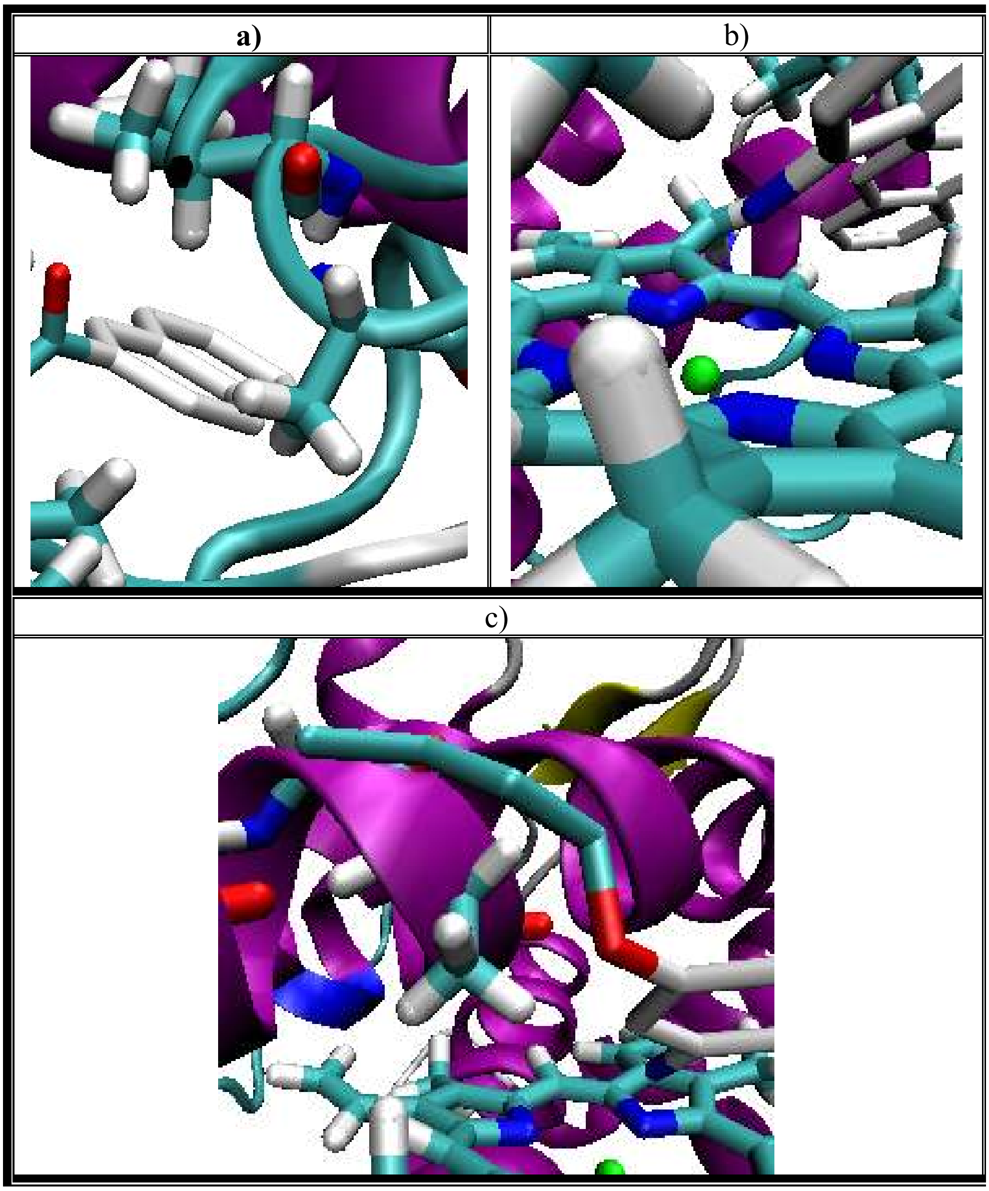
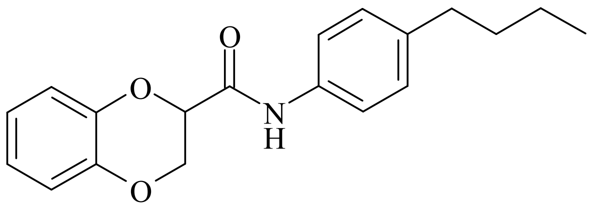
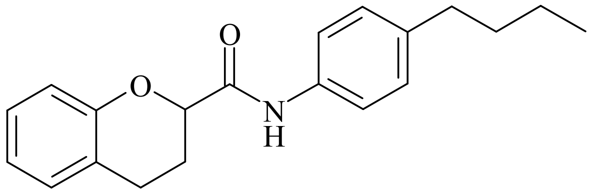
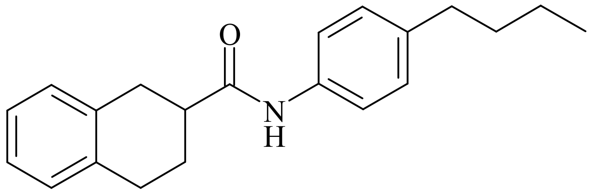
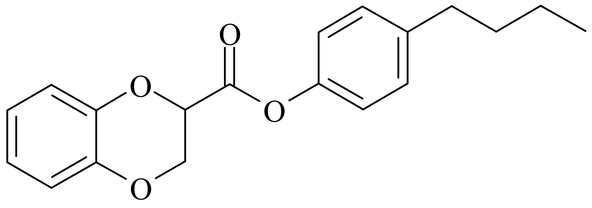
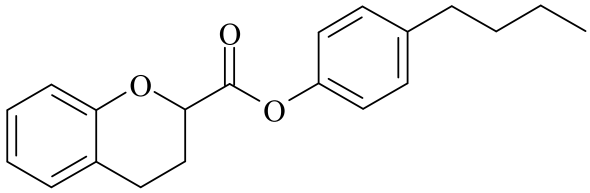
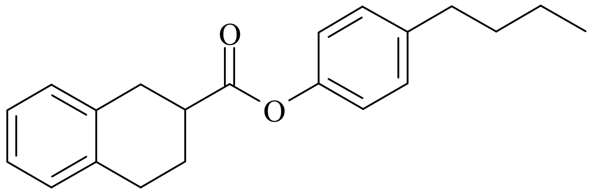
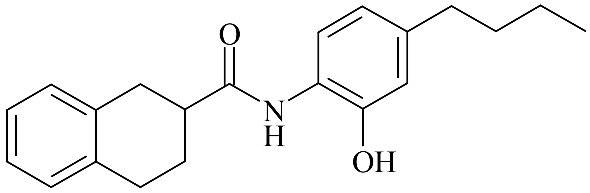
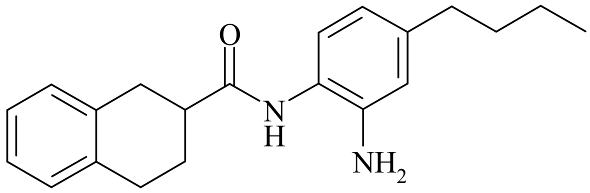
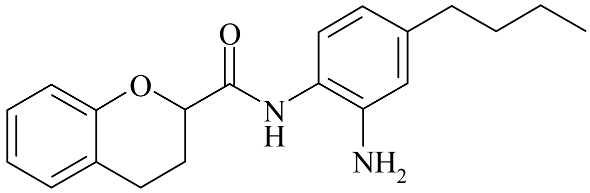
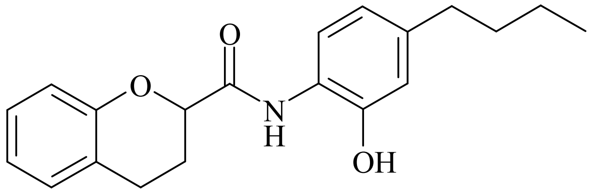


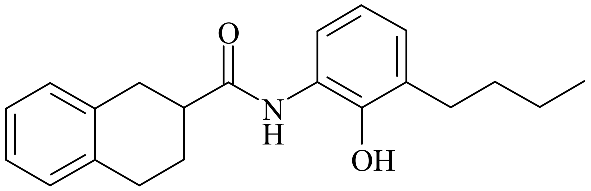
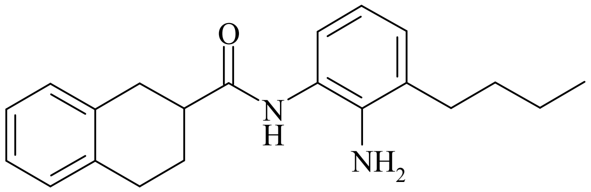
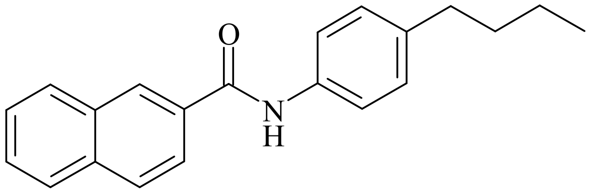
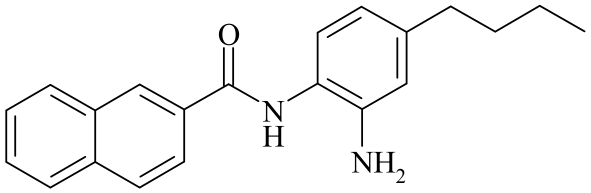
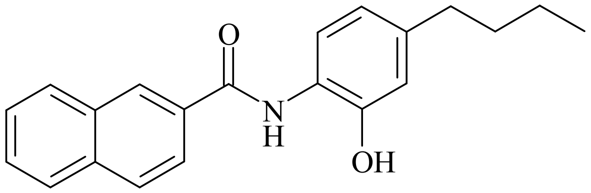
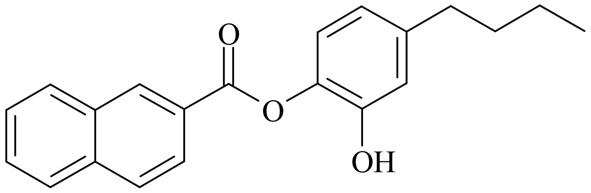
Table 4.2: Virtual representation of interaction sites between side chain residues and the inhibitor N15 derivative.

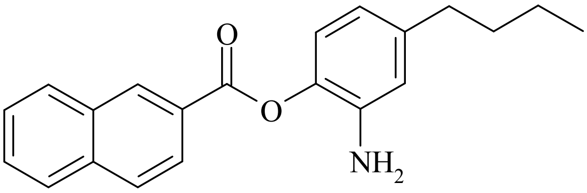
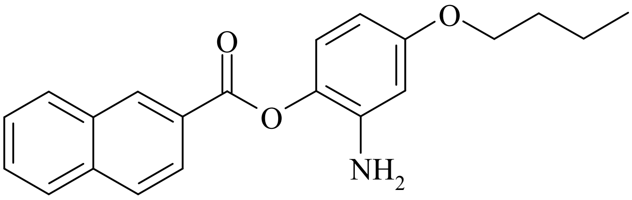
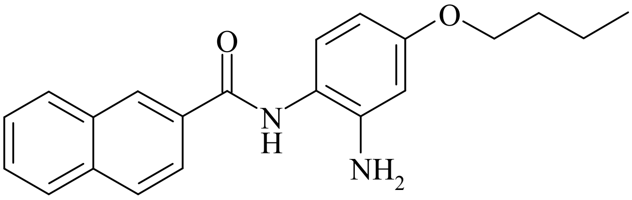
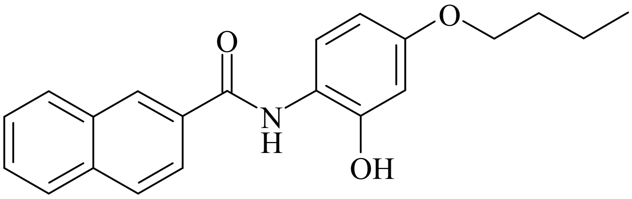
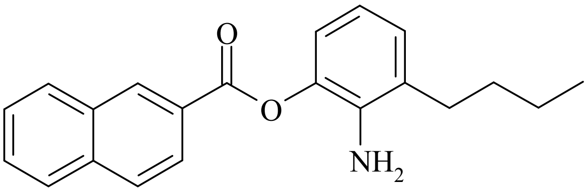
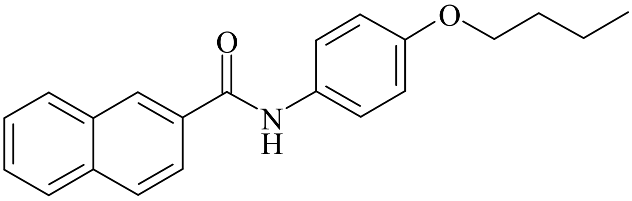
4.5 Analysis of N15 inhibitor derivatives

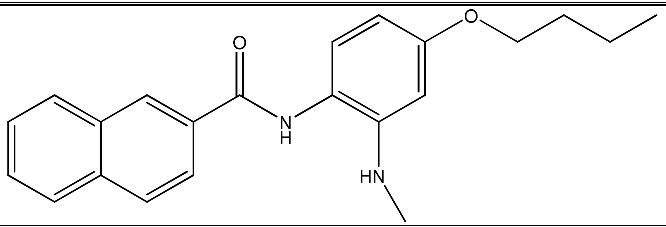
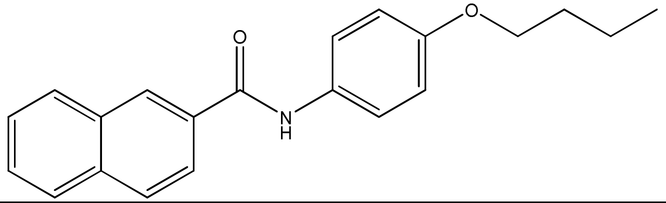
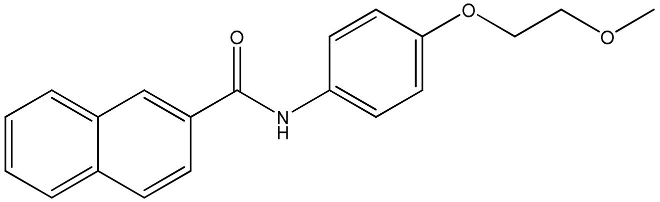
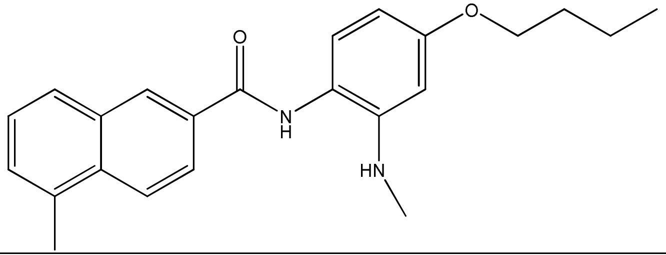
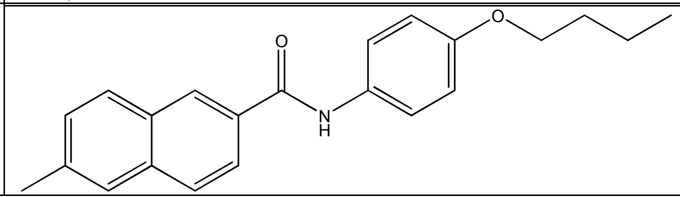
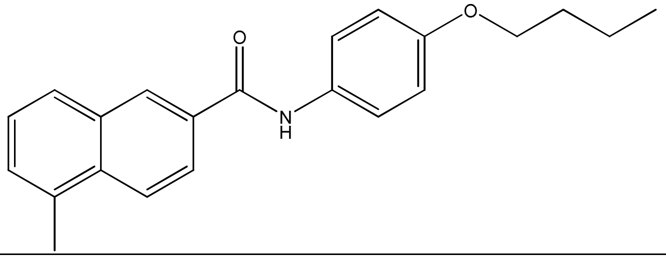
The derivatives that are designed by the strategies discussed in section 4.4 are listed on Table 4.3 with calculated docking and biding scores. 13 of the given derivatives are calculated to improve the docking score and 12 of them are calculated to improve the binding scores.

	LEAD COMPOUND	BINDING	DOCKING
N15		-7.70	-9.60
	LEAD COMPOUND DERIVATIVES	BINDING	DOCKING
D1		-7.61	-9.43
D2		-7.85	-9.5
D3		-6.71	-8.54

D4		-7.39	-8.98
D5		-7.17	-8.75
D6		-7.6	-9.29
D7		-7.46	-9.28
D8		-7.33	-9.31
D9		-7.3	-9.23

D10		-6.48	-8.61
D11		-6.39	-8.5
D12		-7.77	-9.89
D13		-8.47	-10.56
D14		-8.05	-10.13
D15		-7.91	-9.81

D16		-8.26	-10.14
D17		-8.15	-10.1
D18		-8.39	-10.66
D19		-8.1	-10.38
D20		-7.45	-9.4
D21		-7.63	-9.85

D22		-8.11	-10.15
D23		-7.63	-9.85
D24		-7.34	-9.62
D25		-7.47	-9.61
D26		-7.8	-9.48
D27		-7.68	-9.29

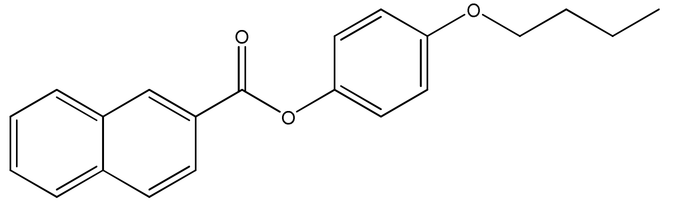
D28		-7.87	-9.05

Table 4.3: N15 derivatives with docking and binding scores

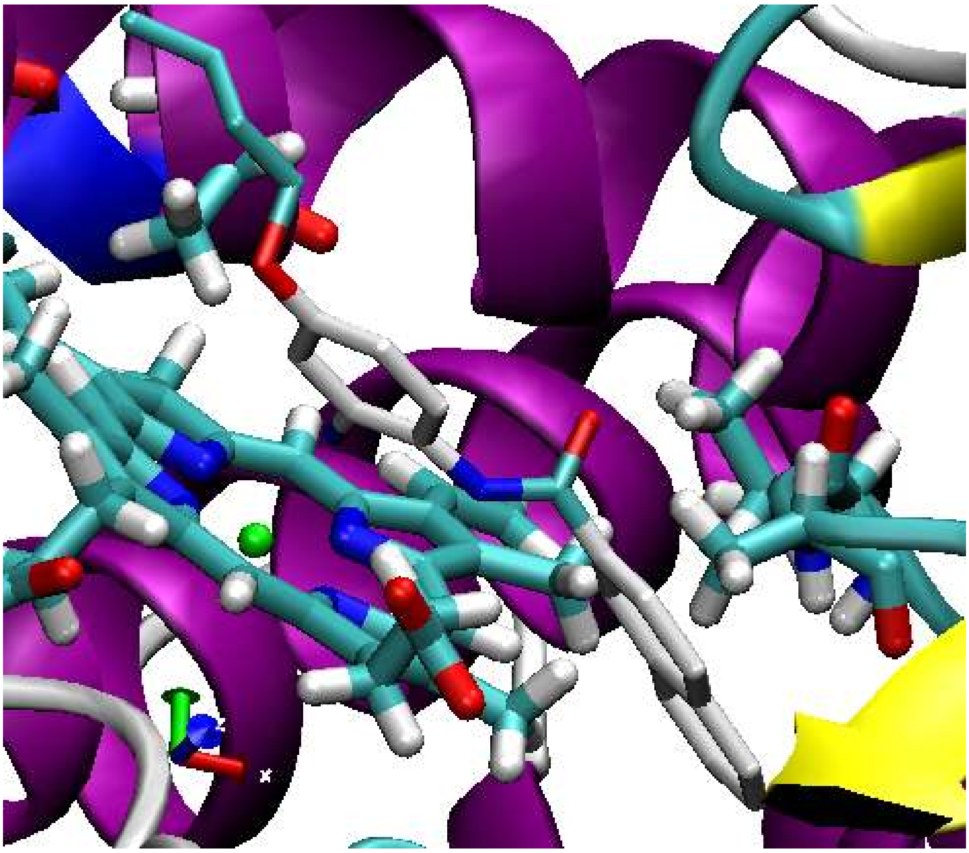


Figure 4.6: Docking position of best scoring derivative D18 in the binding pocket

The hydrophobicity measures (average logP) were also calculated to analyze and predict toxicity potentials of the inhibitors. Table 4.4 shows the predicted logP values for the selected 28 molecules with five different prediction algorithms and average logP values. The molecules within the drug-likeness range suggested by Ghose *et. al.* [57] are highlighted.

	logP	**logP-err**
Lead	**3.99**	0.67
D1	**4.78**	0.65
D2	**5.23**	0.65
D3	**4.62**	0.66
D4	**5.56**	0.73
D5	5.94	0.7
D6	**5.02**	0.66
D7	**4.33**	0.58
D8	**3.82**	0.53
D9	**4.73**	0.61
D10	**4.9**	0.64
D11	**4.25**	0.58
D12	**5.5**	0.63
D13	**4.57**	0.48
D14	5.86	0.65
D15	5.99	0.68
D16	**5.44**	0.76
D17	**5.15**	0.73
D18	**4.27**	0.58
D19	**5.32**	0.64
D20	**5.35**	0.9
D21	**5.12**	0.53
D22	**4.61**	0.51
D23	**5.12**	0.53
D24	**3.83**	0.57
D25	**4.9**	0.51
D26	**5.39**	0.5
D27	**5.39**	0.5
D28	5.79	0.47

Table 4.4: Hydrophobicity measure (logP) of N15 derivatives

Chapter 5

Classification of Drug Molecules

5.1 Overview of QSAR

At the initial stages of drug discovery and design, there are often millions of candidate drug molecules under consideration. Therefore, the early prediction of activity for drug candidates using computational methods is very important to save time and resources. Due to importance of early prediction of activity of drug candidates on the target protein, a large number of computational methods were developed. QSAR (Quantitative Structure-Activity Relationship) analysis is one of the most widely used methods to relate structure to function. QSAR analysis can be described as the quantitative effort of understanding the correlation between the chemical structure of a molecule and its biological and chemical activities such as biotransformation ability, reaction ability, solubility or target activity[81]. QSAR assumes that structurally similar molecules should have similar activities, which draws attention to the importance of detecting the most significant chemical and structural descriptors of the drug candidates. The drug activity behavior can be predicted using a wide range of descriptors.

Some of the most widely used 3D QSAR methods can be listed as follows: comparative molecular field analysis (CoMFA), comparative molecular similarity indices analysis (CoMSIA), the eigenvalue analysis (EVA). In CoMFA, molecular descriptors are calculated and selected by calculating the electrostatic and steric potential energies between a positively charged carbon atom located at each vertex of a rectangular grid and a series of molecules embedded within the grid [82]. The sensitivity to small changes in the alignment of compounds is reduced and hydrogen-bonding and hydrophobic fields are introduced to in CoMSIA[83]. In these methods the aligning of the structures is essential, therefore EVA was used due to the fact that methods that are sensitive to 3D structure but do not require superposition were introduced[84]. The generation of descriptors in EVA is based on molecular vibrations, where a normal mode calculation is required to simulate the IR spectrum of a molecule [85].

In this study, E-Dragon [86-88], which is a remote version of the DRAGON descriptor calculation program, was used to calculate the molecular descriptors for drugs. It applies the calculation of molecular descriptors developed by Todeschini *et.al.* [89] and provides more than 1,600 molecular descriptors, which are divided into 20 blocks, including atom types, functional group and fragment counts, topological and geometrical descriptors, autocorrelation and information indices, 3D molecular descriptors, molecular properties [86-88]. DRAGON incorporates two steps; the first step eliminates low-variable descriptors, the second step optimizes the descriptor subset using a Q^2- guided descriptor selection by means of a genetic algorithm using several data analysis methods: Unsupervised Forward Selection (UFS) [90], Associative Neural Network (ASNN) [91, 92], Polynomial Neural Network (PNN) [93, 94] and Partial Least Squares (PLS) [86-88].

In most studies, Partial least squares (PLS)[95] is used to develop QSAR models by reducing the number of attributes in the descriptor set to a small number of attributes correlated with the defined property being modeled.

In our approach, to classify activities of drug compounds we used the mixed-integer programming (MILP) based hyper-boxes method that take the molecular descriptors in QSAR models as input. The comparison of our classification accuracies with the classification methods available in the WEKA data mining package [16] is also made. WEKA contains 63 different classification methods, but here only 16 of those, which had the best classification accuracies for the data sets considered in this thesis are discussed. Brief overview of these classifiers is further presented in Methods section.

In this study, the problem of QSAR analysis and the classification of drug candidates are addressed based on their published IC_{50} values by introducing an algorithm that combines PLS regression with mixed integer linear programming based hyper-boxes classification method. The strenght of the algorithm not only comes from combining regression with classification but also the ability to improve the classification accuracies by its iterative approach.The algorithm that links QSAR descriptor model generation with inhibitory activity classification was applied to inhibitors of Acetylcholinesterase (ACHE), Benzodiazepine Receptor (BZR), Dihydrofolate Reductase (DHFR) and Cyclooxygenase-2 (COX-2) and finally for Cytochrome P450 C17 (CYP17).

5.2 Classification Methodology

In this thesis, we present an integrated approach combining statistical analysis and MILP based hyper-boxes classification method for early prediction of drug behavior targeting Ache, BZR, Cox-2, DHFR_TG, DHFR_RL, DHFR_PC, and finally Cytochrome P450 C17.

The approach used in this thesis is composed of five main steps. In the first step, molecular structures of the drug candidates is built and optimized the by Marvin Sketch[96]. Then, the molecular descriptors of these drug candidates are obtained using the web server E-Dragon [86-88]. The second step consists of building the regression model using PLS, which will result in selecting the most significant descriptors. Then drug candidates are classified based on the most significant descriptors that are obtained by the previous step, using MILP based hyper-boxes method. This primary classification may result in relatively lower classification accuracy due to the existence of a few insignificant descriptors in the model; therefore, a significance testing analysis is conducted in order to determine the insignificant descriptors that might interfere with our classification accuracy in fourth step. If there are insignificant descriptors in the model we replace the insignificant descriptors with more significant ones; then return to the third step where we classify the drug activities again with the new model that is obtained in step five. After the significance tests if all of the descriptors are significant we build our model with the most significant ones, and report the classification results.

We use an iterative algorithm such that, some of the steps can be repeated when the significance tests give unsatisfactory results for the selected descriptors of a particular model. Less significant descriptors are replaced with a more significant ones affecting the final classification of the drugs at each iteration, thus improves the success of the study. The outline of our method is given in Figure 5.1.

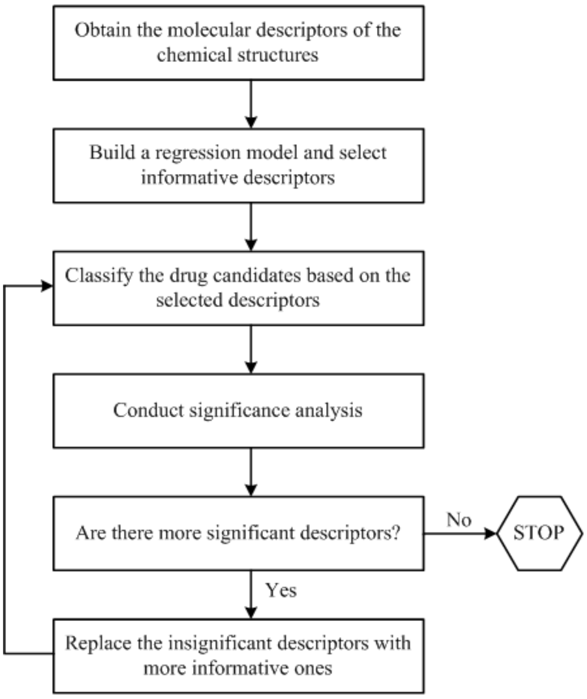
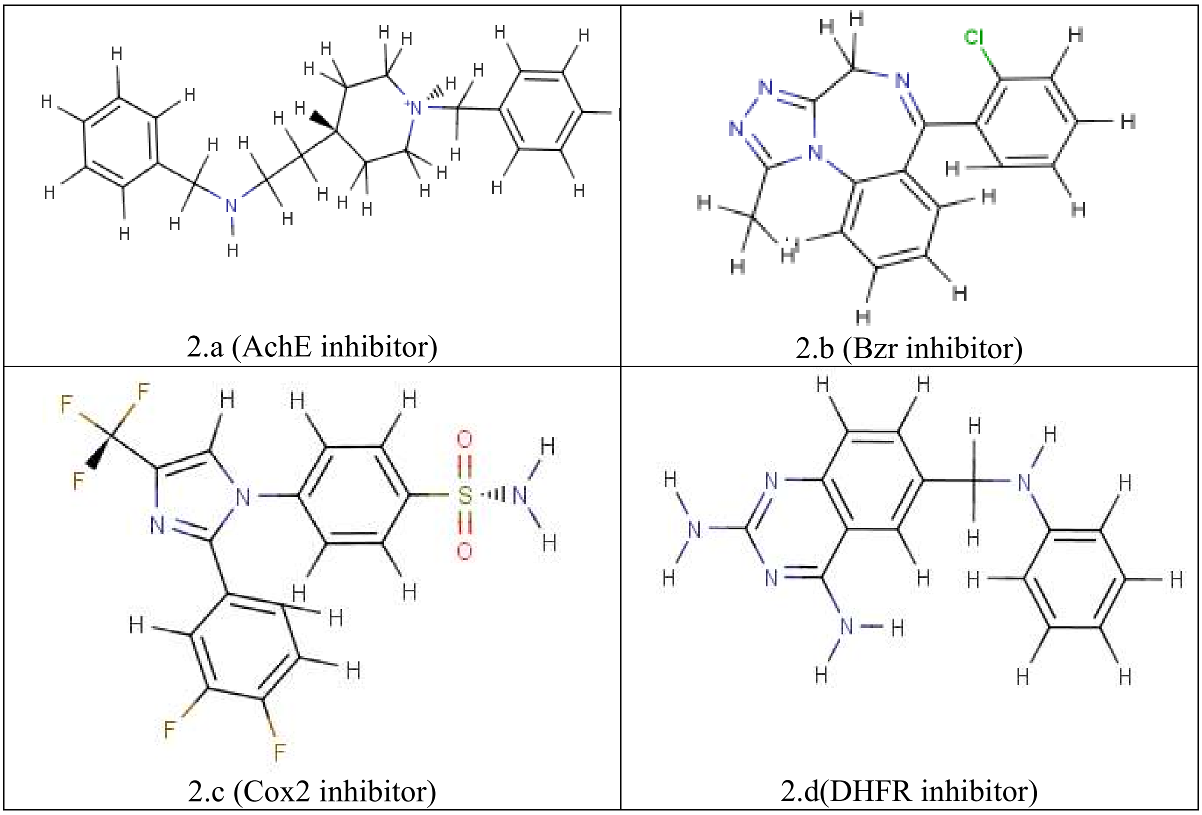

Obtain the molecular descriptors of the chemical structures
Build a regression model and select informative descriptors
Classify the drug candidates based on the selected descriptors
Conduct significance analysis
Are there more significant descriptors?
No
STOP
Yes
Replace the insignificant descriptors with more informative ones

Figure 5.1: Outline of classification approach

5.2.1 Data sets

We applied our algorithm to widely known QSAR data sets available in literature. Dihydrofolate Reductase (DHFR), Acetylcholinesterase (AchE), Benzodiazepine Receptor (BZR) and Cyclooxygenase-2 (COX2) inhibitor sets are used for classification. We also introduce a new dataset of Cytochrome P450 C17 inhibitors, which we have derived from the literature and calculated their 3D structures.



2.a (AchE inhibitor)

2.b (Bzr inhibitor)

2.c (Cox2 inhibitor)

2.d(DHFR inhibitor)

Figure 5.2: Representative compounds from each QSAR data

Seven data sets were used for the validation of our methodology by applying the algorithm on these large and known data sets and comparing our classification accuracy on these data sets with the other widely used classifiers available in the WEKA data mining package. Representative compounds from each data set are shown in Figure 5.2. The experimental IC_{50} values for the dihydrofolate reductase (DHFR) inhibitor set were calculated and reported [97-100] for the DHFR enzyme from three different species: *P. carinii* (PC), *T. gondii* (TG) and rat liver (RL), where the activity of the DHFR inhibitors to the enzymes from different species differ. Therefore, activities of the inhibitors towards the enzymes from these three species for DHFR inhibitors are studied separately in this study. A set of 397 dihydrofolate reductase inhibitors (DHFR) were used for *P. carinii* DHFR with IC_{50} values from 0.31 nM to 3700 µM, a set of 378 inhibitors were used for *T. gondii* DHFR with values from 0.88 nM to 392 µM and 397 inhibitors were used for rat liver DHFR with values from 0.156 nM to 7470 µM. A set of 111 acetylcholinesterase (AchE) inhibitors were used with experimentally calculated IC_{50} values, reported by within the range of 0.3 nM to 100 µM[100-103]. The data set of the benzodiazepine receptor (BZR) inhibitors consisted of 163 inhibitors, whose IC_{50} values were calculated experimentally from 1.2 nM to 5 µM[100, 104]. The 322 molecules of cyclooxygenase-2 (COX2) inhibitor set were derived such that IC_{50} values from 1 nM to 100 µM [100, 105, 106]. The QSAR sets used in this study were also used in a comparison study of QSAR methods by Sutherland et al[100]. We also compared the R^2 values of our 3D descriptor models, which were calculated by the Minitab PLS runs in the first phase of our algorithm, with the reported R^2 values by Sutherland et al [27] for several PLS models on the same data sets.

5.2.2 Structure building and obtaining the descriptor model

As outlined above, in our study the first step is finding molecular descriptors for the drug candidates. Therefore, Marvin Sketch [96] was used to calculate the molecular structures of each drug candidate should be constructed by building their structure and optimize their energy by minimization to determine their confirmation in 3-D space. Next, the optimized 3-D structures are loaded to E- Dragon and molecular descriptors are calculated by using the web server.

E-Dragon suggests many descriptor blocks, each of which contains parameters that describe the characterization of molecules, and the ones that are used in this study can be listed as follows: constitutional descriptors (48), topological descriptors (119), connectivity indices (33), information indices (47), edge adjacency indices (107), topological charge indices (21), geometrical descriptors (74), 3D-MoRSE descriptors (160), functional group counts (154), atom-centered fragments (120), molecular properties (29)[89]. Therefore, the total number of descriptors considered is 912 while building our QSAR descriptor model. PLS[95] is selected for regression analysis because the number of instances is much smaller than the number of attributes (descriptors) by using MINITAB[107]. As we mentioned before, PLS is widely used to develop QSAR models by reducing the number of attributes in the descriptor set to a small number of attributes correlated with the defined property being modeled, which is experimental IC_{50} values in our study.

5.2.3 Model building with PLS for the selection of the most informative descriptors

The main purpose of the regression analysis is to determine the model that predicts the activity (IC_{50}) of the drug candidates in terms of the descriptors. PLS can be referred as an MLR method closely related to principal component regression. Basically, by conducting a PLS study we can predict a set of dependent variables Y based on a set of independent variables X by MINITAB [107],which gave us the PLS runs automatically based on the upper limit we determined on the number of most significant descriptors. Each PLS run provides a linear model of the dependent variable (IC_{50} values) with respect to the independent variables (most significant descriptors). At this point, the relevant model is built and the most significant descriptors are determined. Next step would be the initial classification of the drugs based on the descriptors. The choice of the significant descriptors by the first PLS runs may not be the most effective ones in classification. Therefore, we perform significance tests on the selected descriptors by the regression analysis to increase the classification accuracies.

5.2.4 Classification of drug candidates with MILP based hyper-boxes method

The third step is devoted to the classification of drugs; we apply the MILP based hyper-boxes method [108, 109] by using the selected descriptors from the previous step.

In data classification problem, the objective is assigning the data points, which are described with certain number of attributes, into predefined classes. The strength of hyper-boxes classification method is from its ability to use more than one hyper-box when defining a class as shown in Figure 5.3, and this ability prevents overlapping in the classes, which would not be prevented if the classes were defined with a single hyper-box only[108].

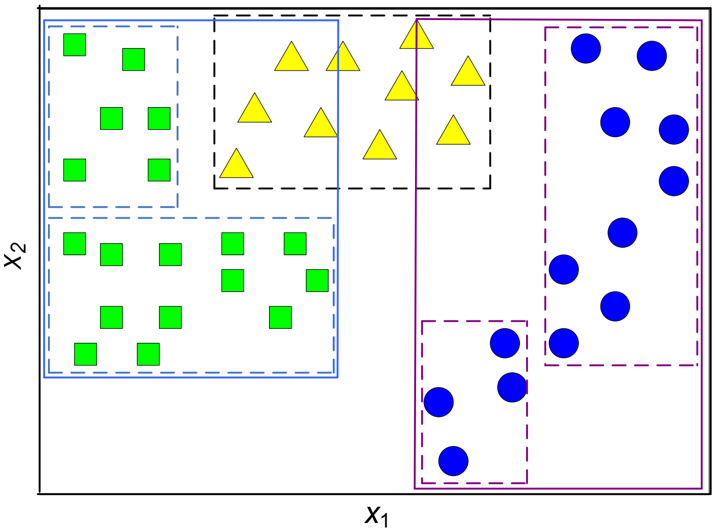


Figure 5.3: Schematic representation of multi-class data classification using hyper-boxes

The data classification problem is solved in two steps: training step and testing step. In the training step, the boundaries of the classes are formed by the construction of hyper-boxes, where as the effectiveness of the constructed classes are tested in the testing step[108].

The MILP problem for the classification is constructed such that the objective function is the minimization of the misclassifications in the data set with the minimum number of hyper-boxes in the training step. The minimization of the number of hyper-boxes, i.e. the elimination of unnecessary use of hyper-boxes, is enforced by penalizing the existence of a box with a small scalar in the objective function. In the training part the upper and lower bound of each hyper-box also calculated by the data points enclosed in that hyper-box[108].

In the testing step, the data points are assigned to classes by calculating the distance between the data point to the each box, and determining the box that is closest to the data point. Finally, the original and assigned classes of test data points are compared and the effectiveness of the classification is obtained by means of correctly classified instances[108].

Solving the proposed MILP problem to optimality is computationally challenging for large datasets due to the large number of binary variables. Hence, a three-stage decomposition method for obtaining optimal solutions of large data classification problems is developed [109]. Instances that are difficult to classify are identified in the first stage that we refer to as preprocessing. Moreover, seeds are determined for each class to improve the computational efficiency. With greater emphasis given to these observations, a solution to the problem is obtained in the second stage with the modified model. Last, final assignments and intersection eliminations are carried out in the third step[109].

In this thesis, we apply this method described above in classifying the activities of drug molecules for the data sets considered. We perform 10-fold cross validation while choosing the training and test sets, where we partition the datasets randomly into 10 subsamples with equal number of members. From these 10 subsamples 9 of them are combined and used as the training set, and the remaining 1 subsample is used as the test set. Then the classification is performed 10 times with each of the 10 subsamples used exactly once as the test set. Finally, the accuracy of the classification is reported as the average of these 10 classifications.

We classify each of the drug candidates in the test set as having a low or high IC_{50} value. In this iterative study, this classification step is performed several times: first with the initial set of descriptors then using the enhanced set of descriptors derived from significance analysis.

5.2.5 Significance analysis

In the fourth step, significance tests are performed. After the PLS runs it is possible to conclude a descriptor as significant while it is not in reality and this problem is resolved by conducting significance tests after primary classification. The main idea behind the significance test is as follows: If Z is the whole set of drug candidates, assume after the classification it is divided into two classes, A and B. For a successful classification, the variances of descriptor values should be smaller within classes *A* and *B* than it is for the whole population, *Z*.

The equation given below in Eq. 2.1 exhibits the *F* distribution.

$$\frac{S_{ij}^2 / \sigma_i^2}{S_k^2 / \sigma_i^2} = S_{ij}^2 / S_{ik}^2 = f_{\nu\eta} \tag{5.1}$$

where, S_{ij}^2 is the sample variance of values for descriptor *i* for drug set *j,* ν=*n*-1 and η=*m*-1 are degrees of freedom, and *n* is the number of values of descriptor *i* for the drug set *j*, and *m* is the number of values of descriptor *i* for the drug set *k*.

Then the hypothesis testing is performed by the null hypothesis $S_{ij}^2 = S_{ik}^2$, which suggests that the variance of the whole set of drug candidates is equal to the variance of the drugs within the same class. Since the variance of the whole set of drugs should be larger than the variance within the class, we define our alternative hypothesis as: $H_a = S_{ij}^2 \text{ f } S_{ik}^2$, where j is a member of a whole data set and k is a member of the class. Note that the *p*-value of $f_{\nu\eta}$ in the current should be smaller than the *p*-value of $f_{\nu\eta}$ in the previous model to accept the alternative hypothesis.

5.2.6 Building the new classification model

This last step is performed when we conclude that there are overestimated descriptors in the model during step four. Therefore, a total number of 3 models are constructed through regression analysis by selecting 7,10 and 15 descriptors respectively as representative variables of each model, and the significance analysis is applied to all of the descriptors in these 3 models. If we conclude existence of an insignificant variable in one of these models, we replace them with the ones that are significant in the other models. This adjustment is proved to improve our classification accuracy. When we are replacing the less significant ones, the remaining 880 descriptors that are eliminated during the PLS analysis are ignored, since these 7, 10, and 15 attributes were chosen by the PLS regression analysis and have a proven strength in describing the IC_{50} values. The main purpose of the PLS regression study in fact is eliminating the statistically meaningless features, and provide us with the most meaningful sample space to further work with.

The results obtained by our method are compared with all of the 63 classification methods available in WEKA, and 16 best WEKA classifiers reported with the results obtained by our algorithm in Table 5.3, with the corresponding classification accuracy. The attributes used in WEKA classifiers are the same descriptors that are found after the significance tests, and 10-fold cross validation was applied to each classifier including our classification method.

WEKA is a powerful data mining tool to use for comparison purposes, since it includes all widely known machine-learning algorithms among its 63 classifiers. The success of these existing machine learning algorithms in binary classification of active and inactive compounds based on their descriptor values were also previously reported [110]. Following is a brief overview of the best performing data classification methods available in WEKA. A Bayesian network[111] $B = < N, A, \Phi >$ is a directed acyclic graph <N,A> with a conditional probability distribution attached to each node, collectively represented by Φ.

Each node $n \in N$ represents a dataset attribute, and each arc $a \in A$ between nodes represents a probabilistic dependency. The Naive Bayes classifier assumes that all of the variables are independent of each other, where the classification node is represented as the parent node of all other nodes[112]. Naive Bayes Simple uses the normal distribution for the modelling of the attributes and handle numeric attributes using supervise discretization, where as Naive Bayes Updateable is an incremental version, which processes one instance at a time, and uses a kernel estimator instead of discretization.

The Logistic classifier[112] builds a two-class logistic regression model. It is a statistical regression model, where logistic regression assumes that the log likelihood ratio of class distributions is linear in the observations. The Simple Logistic classifier builds linear logistic regression models based on a single attribute[112]. The model is a generalized model of the ordinary least squares regression model. Multilayer perceptron[112] is a neural network that uses back propagation. The perceptron, which is a processing element, computes a single output, a nonlinear activation function of linear combination of multiple inputs, whose parameters are learned through the training phase. SMO (sequential minimal optimization)[113], also called the WEKA SVM(support vector machine), is a method to train a support vector classifier using polynomial kernels by breaking a large quadratic programming optimization problem into smaller QP optimization problems.

IB1[112] is listed as a lazy classifier, in a sense that it stores the training instances and it does not really do any work until the classification time. IB1 is an instance based learner. It finds the training instance closest in Euclidian distance to the given test instance. IBk is a k-nearest-neighbor classifier that uses the same idea.

Logit Boost[114] uses additive logistic regression. The algorithm can be accelerated by assigning a specific threshold for weights. Multi Class Classifier[115] uses four distinct two-class classification methods for multiclass problems. The Threshold Selector[112], which is a meta learner optimizes the F-measure by selecting a probability threshold on the classifiers output.

Random forest and LMT are decision tree methods. Random Forest generates random trees by collecting ensembles of random trees, where as LMT builds logistic model trees, and uses cross validation to determine the number of iterations while fitting the logistic regression functions at each node. OneR (one rule)[112] builds a one-level decision tree and learns a rule from each attribute and selects the rule having the smallest error rate as the one rule.

5.3 Classification Results

To determine the threshold values, which divide the low and high classes, for all datasets the IC_{50} values were statistically analyzed. In this study, we consider 6 datasets, of which IC_{50} values and structures were reported [97-106, 116]. In addition to these datasets we introduced a new dataset for P450 C17 inhibitors that we collected from the literature. P450 C17 is a well-recognized target for prostate cancer treatment, since selective inhibition of the enzyme exerts control over androgen synthesis [117].

5.3.1 PLS Results

After building the descriptor models by e-Dragon [88], three models were constructed during the PLS analysis as: 7, 10 and 15 descriptor models. The reason that we build 3 models with different number of variables is due to the fact that we might come up with insignificant descriptors within one of these models, so that we can replace them by a more significant one from the other models.

	CoMFA*	CoMSIA basic*	CoMSIA extra*	EVA*	HQSAR*	2D*	2.5D*	eDragon-7	eDragon-10	eDragon-15
AchE	0.88	0.86	0.86	0.96	0.72	0.40	0.38	0.84	0.90	0.95
BZR	0.61	0.62	0.62	0.51	0.64	0.51	0.52	0.51	0.67	0.79
COX-2	0.70	0.69	0.69	0.68	0.70	0.62	0.68	0.53	0.61	0.73
DHFR_RL	0.79	0.76	0.75	0.81	0.81	0.61	0.65	0.42	0.53	0.64
DHFR_PC	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.44	0.54	0.65
DHFR_TG	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.40	0.51	0.66
Cytochrome P450 C17	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.84	0.91	0.95

* PLS results reported by Sutherland et al.

Table 5.1: Comparison of R^2 values for PLS models.

In Table 5.1, the QSAR models with the most significant descriptors, as they were concluded as a result of the initial PLS study for the 7, 10 and 15 attribute models are listed above with their R^2 values. Table 5.1 shows the optimal R^2 values of our PLS models given by Minitab[107] with predefined number of descriptors from the descriptors calculated by e-Dragon software, and the R^2 values of the PLS models calculated by Sutherland *et al.*[100] with the same data sets but different methods and models.

The R^2 values shows that, the models we developed with 10 and15 descriptors for Ache BZR and COX-2 are stronger than or at least as strong as the other models reported by Sutherland et al [100] in representing the IC_{50} values in terms of selected descriptors, but our model for DHFR_RL is not as good as the other reported models. The high R^2 values of Cytochrome P450 C17 models also suggest good prediction of the IC_{50} values and a promising initial model for classification.

It is worth to note that, our study is not simply a regression study, but we develop these regression models in order to use the selected descriptors from this step as attributes for accurate classification.

5.3.2 Iterations

At the end of the initial runs of the hyper-boxes classification method, classification results are obtained. The next step is the significance analysis and the improvement of the classification accuracies by iterations. By the significance analysis, the weakest and the strongest descriptors were calculated and, the weakest descriptor in the current model was replaced by the most significant one from other models at each iteration. The classification runs are repeated after each replacement, by MILP based hyper-boxes method. When the classification accuracy is not improved at the end of an iteration, the algorithm stops and our final results are reported (Table 5.2).

Classification Accuracies					
		Iter #0	**Iter #1**	**Iter #2**	**Iter #3**
ACHE	**7 Attributes**	91.89	100.00		
	10 Attributes	86.48	89.19	91.89	
	15 Attributes	86.05	89.18		
BZR	**7 Attributes**	90.90	96.36		
	10 Attributes	92.73	94.55		
	15 Attributes	90.09	92.73		
COX2	**7 Attributes**	94.39	95.33	97.20	98.13
	10 Attributes	91.58	97.20		
	15 Attributes	88.78	89.72	90.65	
DHFR-1	**7 Attributes**	94.73	96.99		
	10 Attributes	93.98	97.74		
	15 Attributes	94.73			
DHFR-2	**7 Attributes**	95.23	96.83	97.62	
	10 Attributes	94.44	95.24	98.41	
	15 Attributes	92.06	93.65		
DHFR-3	**7 Attributes**	96.24	97.74		
	10 Attributes	93.23	93.98	96.24	
	15 Attributes	96.24	97.74		
P450 C17	**7 Attributes**	86.36	90.00	97.20	100.00
	10 Attributes	100.00			
	15 Attributes	100.00			

Table 5.2: Classification Accuracies of each iteration

While choosing the weakest descriptor to leave the model, the descriptor with the maximum p-value (failed to reject H_0 with the greatest error, see methods section for our hypothesis) for one of the high or low classes was selected. The weakest descriptor was replaced by the strongest one. The strongest descriptor defined as the attribute whose maximum p-value for high and low classes is the minimum among the other descriptors. An instance of the significance analysis is presented in Table 5.4.

5.3.3 Final Classification

As shown in Table 5.3, we compared the classification accuracies of our model with the results that calculated using all of the classification methods in WEKA. We report only the results of the16 best performing WEKA classifiers. Our method performed better than all of the other classifiers for every model of each dataset. Our integrated approach of regression and classification for Ache and P450 C17 inhibitors datasets displayed an activity prediction accuracy of 100%. The activity of BZR inhibitors was calculated with the accuracy of 96.36%. We were able to predict the activities of COX-2 inhibitors with 98.13% in a 7-descriptor model. In addition, the prediction accuracy of activity of DHFR_RL, DHFR_PC, and DHFR_TG inhibitors were 97.74%, 98.41% and 97.74% respectively. The best performing WEKA classifiers are also highlighted in Table 5.3.

ACHE	**% accuracy**			**BZR**	**% accuracy**		
	7-att	**10-att**	**15-att**		**7-att**	**10-att**	**15-att**
MILP based hyper-boxes method	**100**	**91.89**	**89.19**	**MILP based hyper-boxes method**	**96.36**	**94.55**	**92.73**
Bayes Network	79.28	77.48	78.38	Bayes Network	77.91	77.3	73.62
Naive Bayes	80.18	80.18	81.08	Naive Bayes	80.37	77.91	66.26
Naive Bayes Simple	81.08	80.18	81.98	Naive Bayes Simple	79.14	77.3	68.71
NaiveBayesUpdatable	80.18	80.18	81.08	NaiveBayesUpdatable	80.37	77.91	66.26
Lojistic	79.28	***84.68***	80.18	Lojistic	***83.44***	80.98	80.98
Multilayer Perceptron	82.88	81.08	81.08	Multilayer Perceptron	79.75	80.98	79.14
SimpleLogistic	***83.78***	82.88	79.28	SimpleLogistic	80.98	***82.82***	79.14
SMO (WEKA SVM)	79.28	80.18	80.18	SMO (WEKA SVM)	79.14	77.91	77.91
IB1	70.27	80.18	77.48	IB1	72.39	74.85	75.46
IBk	70.27	80.18	77.48	IBk	72.39	74.85	75.46
Logit Boost	82.88	81.08	***82.88***	Logit Boost	78.53	77.3	77.91
Multi Class Classifier	79.28	***84.68***	80.18	Multi Class Classifier	***83.44***	80.98	***80.98***
Threshold Selector	47.75	68.47	60.36	Threshold Selector	78.53	76.69	75.46
LMT	***83.78***	82.88	79.28	LMT	80.98	***82.82***	79.14
RandomForest	80.18	80.18	81.98	RandomForest	77.3	79.75	***80.98***
OneR	81.08	72.97	72.97	OneR	74.85	74.23	79.14

DHFR_TG	**% accuracy**			**COX2**	**% accuracy**		
	7-att	**10-att**	**15-att**		**7-att**	**10-att**	**15-att**
MILP based hyper-boxes method	**97.74**	**96.24**	**97.74**	**MILP based hyper-boxes method**	**98.13**	**97.2**	**90.65**
Bayes Network	77.33	78.09	73.05	Bayes Network	67.2	67.2	66.88
Naive Bayes	76.57	***79.35***	72.54	Naive Bayes	71.66	70.06	64.65
Naive Bayes Simple	75.57	78.84	67	Naive Bayes Simple	72.29	70.06	64.65
NaiveBayesUpdatable	76.57	***79.35***	72.54	NaiveBayesUpdatable	71.66	70.06	64.65
Lojistic	75.82	78.84	75.57	Lojistic	72.29	70.38	70.06
Multilayer Perceptron	76.32	77.08	75.06	Multilayer Perceptron	***72.61***	72.29	***75.16***
SimpleLogistic	74.56	77.83	75.31	SimpleLogistic	72.29	71.97	68.47
SMO (WEKA SVM)	72.54	79.09	72.54	SMO (WEKA SVM)	71.02	69.43	69.43
IB1	75.31	79.09	75.82	IB1	69.11	71.02	70.06
IBk	75.31	79.09	75.82	IBk	69.11	71.02	70.06
Logit Boost	77.33	78.34	78.34	Logit Boost	71.66	70.06	70.7
Multi Class Classifier	75.82	78.84	75.57	Multi Class Classifier	72.29	70.38	70.06
Threshold Selector	69.77	74.81	73.55	Threshold Selector	68.47	65.29	64.65
LMT	76.07	76.57	77.83	LMT	71.34	71.02	68.15
RandomForest	***77.58***	79.09	***80.35***	RandomForest	71.97	***74.2***	70.06
OneR	69.77	69.77	70.53	OneR	70.7	70.38	70.06

DHFR_RL	% accuracy			DHFR_PC	% accuracy		
	7-att	10-att	15-att		7-att	10-att	15-att
MILP based hyper-boxes method	**96.99**	**97.74**	**94.73**	**MILP based hyper-boxes method**	**97.62**	**98.41**	**93.65**
Bayes Network	63.72	71.78	70.5	Bayes Network	80.42	80.42	78.04
Naive Bayes	63.97	68.76	71.7	Naive Bayes	82.54	81.48	80.95
Naive Bayes Simple	63.97	67.75	71	Naive Bayes Simple	82.8	79.89	81.22
NaiveBayesUpdatable	63.98	68.77	71.78	NaiveBayesUpdatable	82.54	81.48	80.95
Lojistic	***69.52***	73.8	78.58	Lojistic	81.75	83.33	81.75
Multilayer Perceptron	62.72	76.57	77.58	Multilayer Perceptron	82.8	82.8	84.13
SimpleLogistic	66.75	73.55	78.33	SimpleLogistic	80.42	***84.13***	81.22
SMO (WEKA SVM)	64.99	73.05	79.59	SMO (WEKA SVM)	82.28	83.33	79.1
IB1	62.97	75.06	81.11	IB1	82.28	80.16	81.75
IBk	62.97	75.06	***81.11***	IBk	82.28	80.16	81.75
Logit Boost	64.99	75.06	77.33	Logit Boost	83.33	81.48	81.48
Multi Class Classifier	69.52	73.8	78.59	Multi Class Classifier	81.75	83.33	81.75
Threshold Selector	64.99	69.52	78.59	Threshold Selector	83.33	79.1	81.22
LMT	65.24	***77.33***	77.83	LMT	***83.6***	83.07	***85.19***
RandomForest	68.51	77.08	77.83	RandomForest	82.8	80.95	83.07
OneR	61.46	66	62.72	OneR	79.89	79.89	80.16

Table 5.3: Comparison of classification accuracies of best WEKA classifiers with the MILP based hyper-boxes classification

5.4 Detailed analysis: Cytochrome P450 C17 inhibitors

We applied our approach to classify activities of drug molecules in a new data set (P450 C17) that is constructed from data in literature [117, 118]. This approach may be utilized for the new molecules that inhibit activity of P450 C 17 before channeling them into experiment.

For the 7, 10 and 15 attribute models the selected most significant descriptors as a result of the initial PLS study are listed with R^2 values (Table 5.1) of 0.946, 0.907 and 0.838 respectively.

When the hyper-boxes model was implemented, 10 and the 15-attribute models reached 100% accuracy, by 10-fold cross validation. The 7-attribute model, however, still needed to be improved since the classification results reached an average accuracy of 96.35%. This led us to conclude that there may be some overestimated descriptors actually having low significance in terms of classifying the drug activity. Therefore, significance tests were performed after the preliminary classification runs.

	Iteration 1	**Iteration 2**	**Iteration 3**
Descriptor Leaving the 7 desc. Model	maxmax1	maxmax2	maxmax3
	C-027	*EEig01d*	*Mor22m*
	0.96416	0.9491	0.67855
Descriptor Entering the 7 desc. Model	minmax1	minmax2	minmax3
	EEig04x	*nHAcc*	*Mor14e*
	0.5455	0.5783	0.5946

Table 5.4: The descriptors leave the 7 descriptor model and the descriptors replacing them

Table 5.4 shows the *p*-value calculation results for the descriptors for each iteration. At iteration 1, C-027 was detected since it had the largest *p*-value among the other descriptors. Then from the significance analysis of 10 descriptor model, EEig04x was chosen to replace it, since its maximum *p*-value is the minimum among the other descriptors. After each replacement, the hyper-boxes classification model was built and performed with the new attributes and, average classification accuracy was determined. The runs were stopped after iteration 3 since we reached 100% accuracy. The classification results are reported in Table 5.5.

P450 C17 Classification Method	**% accuracy**		
	7-attribute	**10-attribute**	**15-attribute**
MILP based hyper-boxes method	**100.00**	**100.00**	**100.00**
Bayes Network	***81.25***	***81.25***	***81.25***
Naive Bayes	62.50	71.88	53.13
Naive Bayes Simple	62.50	68.75	50.00
Naive Bayes Updatable	62.50	71.88	53.13
Lojistic	71.88	56.25	62.50
Multilayer Perceptron	62.50	71.88	59.38
SimpleLogistic	75.00	75.00	***81.25***
SMO	***81.25***	***81.25***	***81.25***
IB1	59.38	59.38	***81.25***
IBk	59.38	59.38	62.50
Logit Boost	71.88	62.50	62.50
Multi Class Classifier	71.88	56.25	62.50
Threshold Selector	43.75	40.63	62.50
LMT	75.00	75.00	***81.25***
RandomForest	75.00	68.75	65.63
OneR	75.00	71.88	75.00

Table 5.5: Comparison of classification accuracies of best WEKA classifiers with MILP based hyper-boxes classification on P450 C17 inhibitors.

The results of the final run of hyper-boxes classification for the 7-descriptor model showed that the effect of changing the less significant descriptors with the more significant ones improved the accuracy of the classification from 96.36% to 100%. Since we have reached 100 % accuracy in 7–descriptors models, the significant ones may be included in this model among 912-descriptors that are initially calculated by e-DRAGON. Brief explanations of the descriptors can be found in Table 5.6 [89].

Mor10m	3D-MoRSE - signal 10 / weighted by atomic masses
DISPp	d COMMA2 value / weighted by atomic polarizabilities
Mor14e	3D-MoRSE - signal 14 / weighted by atomic Sanderson electronegativities
Mor08m	3D-MoRSE - signal 08 / weighted by atomic masses
nHAcc	number of acceptor atoms for H-bonds (N. O. F)
EEig04x	Eigenvalue 04from edge adj. matrix weighted by edge degrees
DISPv	d COMMA2 value / weighted by atomic van der Waals volumes

Table 5.6: The brief explanations of the most significant descriptors

5.5 Discussion

Early analysis and estimation of the drug activities by computational methods are widely studied in order to narrow down drug candidates for further experimental tests. The accuracy comparison of our algorithm with other QSAR algorithms suggests that drug activities can be classified with a significantly higher accuracy with the method introduced in this study.

After model building by E-dragon QSAR software, the PLS runs were performed to determine the best model in representing the depended variables (IC_{50} values) in terms of the independent variables (the attributes). The corresponding R^2 values were calculated to determine the reliability of the PLS models, where a model with a higher R^2 value can be regarded as a more reliable model to proceed to the classification step. The R^2 values for the 15, 10 and 7 descriptor models of P450 C17 were obtained by PLS runs and, with a considerable strength in representing the IC_{50} values we accepted the initial models. While the high R^2 values of the Ache inhibitor models also were promising on its own to build the classification model, the initial models of BZR and COX2 inhibitor sets were chosen after the comparison of PLS results with the results reported in the literature as presented in Table 5.3. For DHFR inhibitors data sets such comparison was not also possible, therefore the models with the best R^2 values in PLS studies were chosen among all other possible models calculated. R^2 value directly depends on the values of attributes (the descriptors) and the number of attributes in the corresponding model.

We first applied our iterative algorithm to the large and widely used QSAR data sets in order to validate our methodology. The strength of our algorithm was presented by comparing our classification accuracies with the classification accuracies of 63 WEKA classifiers, on 7 inhibitor sets. The attribute sets prepared as the input for WEKA classifiers were the same as the ones, by which we built the last iteration of our MILP based hyper-boxes classification model. In other words, those were the most significant attributes that we used to develop the final classification models and reached our best accuracies. Our approach outperformed all of the classifiers available in WEKA for each model of the all of the 7 data sets, even reaching 100% accuracy in predicting the activity classification of the inhibitor sets, Ache inhibitors and Cytochrome P450 C17. A total number of 21 QSAR models were built in this study for 7 inhibitor sets, and in 18 of them the accuracy of our methodology exceeded the accuracy of the second best classifier with more than 10%. Through all of the 21 models, the smallest difference in the accuracies is 6.31% and the largest difference is 27.47%.

The higher prediction accuracy of the model not only comes from the choice of initial models by PLS analysis but also the characteristics of MILP based hyper-boxes method. The MILP based hyper-boxes approach allows using more than one hyper-box in order to define a single class [108]. Moreover, this approach deals with problematic and non-problematic instances separately and prevents overlapping of final hyper-boxes [109]. Therefore, these strengths significantly improve the accuracy and efficiency of the MILP based hyper-boxes approach compared to other data classification methods.

Chapter 6

CONCLUSIONS

In this study, structure based rational drug design approach was employed to discover novel inhibitors for two different target proteins for cancer research. The first target protein was Cytochrome P450 C17, the key enzyme in androgen synthesis and the second target protein was XPF, a key member of the DNA excision repair system. In designing an inhibitor targeting Cytochrome P450 C17, we aim to decrease androgen levels in the cells and therefore prevent the progression of the prostate cancer. However, targeting the XPF-ERCC1 pair was not a direct chemotherapeutic agent design approach, but the aim was to design XPF-ERCC1 inhibitors as chemotherapeutic agent "aids" that prevent DNA repair and resistance development in cancerous cells against the chemotherapeutic agents. Finally a novel QSAR approach to predict the activity level of the inhibitors was presented, since a priori analysis of the activity of inhibitors on the target protein by computational approaches can be useful in narrowing down drug candidates for further experimental tests.

As a method of treatment for prostate cancer, reducing levels of androgens was adopted. In order to achieve this, CYP17 is selected as the target enzyme due to its key role in androgen biosynthesis. First inhibitor designs were based on PREG and PROG, since no other inhibitors were reported. Due to use of substrates, first generation of designed compounds was steroid based molecules with various side chains attached to 17th carbon. Even though some of these compounds showed inhibitory effects, they were not promising candidate drugs. The drawbacks of steroidal compounds were poor acid-stability, poor bioavailability, short half-life, first-pass effects and poor selectivity[64]. Next, ketoconazole was shown to reduce androgen levels in human, and has inhibitory effect on CYP17. However, it has been removed from use because of liver toxicity and its effects on other cytochrome enzymes[69]. Recently, two steroidal compounds (VN/124-1 and abiraterone) were reported to be promising CYP 17 inhibitors and showing success in preclinical and clinical trials respectively[78]. However; these two compounds are steroidal compounds, thus a successful design of non-steroidal, specific, non-toxic CYP 17 inhibitor is still an intriguing area of research.

The non-steroidal compounds that are reported in the literature are mainly based on mimicking known steroids that are interacting with the CYP 17 active site (i.e. progesterone), pharmacophore modeling (based on QSAR analysis of known inhibitors) and creating derivatives of known inhibitors of cytochrome enzymes such as: antifungal imidazole containing agents related to ketoconazole, pyridine derivatives, xanthone derivatives carbazole derivatives. The non-steroidal compounds reported in the literature did not yet show promising results in clinical trials.

In this study, we show that structure-based drug design approach based on the model protein structure by Auchus *et al.*[68] (PDB id: 2C17) can be successfully applied to identify novel CYP 17 inhibitors. Molecular dynamics simulation was used to improve the protein structure by Auchus *et al.* to further use in docking studies. The identified steroidal and non-steroidal compounds that inhibit the biological activity of CYP 17 in silico, further tested for toxicity, and the assay indicated that the non-steroidal "lead compound" does not display toxic effects on HeLa Cell line in 24 hours at a concentration equals to its IC_{50} value. Therefore; the identified lead molecule (N15) may be used as a promising lead for further optimizations to improve inhibition efficiency and discover a novel non-steroidal CYP17 inhibitor.

In designing the derivatives we aim to improve the binding strength of the inhibitor, while keeping the toxicity measure at minimum. The docking position of N15 and the probable atomic interactions between N15 and the side chains of the active site residues were carefully studied and 28 derivatives were designed. The derivatives were further analyzed by docking simulations to study inhibition strength and logP calculations to study bioavailability.

Next, structure based rational drug design approach was applied to design inhibitors to prevent XPF-ERCC1 pair formation. Since XPF-ERCC1 has a key role in nucleotide excision repair, XPF-ERCC1 inhibitors were designed as chemotherapeutic agent aids that prevent DNA repair and resistance development in cancerous cells.

The target binding pocket on XPF was determined by several experimental and computational studies. Experimental data suggests that the deletion of Phe-293 eliminates binding to XPF, where Phe-293 is naturally located in a pocket surrounded by XPF residues 837-905. HotSprint reported Pro-837 Phe-840 Leu-841 Met-844 Met-856 Ile-862 Leu-865 and Phe-889 as hot spot residues on XPF, and Arg-234 Leu-239 Ser-259 and Phe-293 on ERCC1, which agrees with the experimental data previously reported. Computational alanine scanning was also applied to the XPF-ERCC1 pair to determine important interactions in the protein-protein interface. The inhibitor target site is determined on XPF to be the pocket surrounded by residues Leu-841, Met-856, Val-859, Lys-860, interacting the key interface residue Phe-293.

Virtual screening with MD simulations were applied to determine drug candidates from a 6 million compound library and the best 28 molecules were determined based on docking scores (i.e. binding affinity to the active site) and logP scores (i.e. bioavailability score) to further test experimentally.

Finally, a computational classification approach was created to enhance the drug design process by eliminating the number of drug candidates to be tried experimentally. The accuracy comparison of our algorithm with other QSAR algorithms suggests that drug activities can be classified with a significantly higher accuracy with the method introduced in this study.

After model building by E-dragon QSAR software, the PLS runs were performed to determine the best model in representing the depended variables (IC_{50} values) in terms of the independent variables (the attributes). The corresponding R^2 values were calculated to determine the reliability of the PLS models, where a model with a higher R^2 value can be regarded as a more reliable model to proceed to the classification step. The R^2 values for the 15, 10 and 7 descriptor models of P450 C17 were obtained by PLS runs and, with a considerable strength in representing the IC_{50} values we accepted the initial models. While the high R^2 values of the Ache inhibitor models also were promising on its own to build the classification model, the initial models of BZR and COX2 inhibitor sets were chosen after the comparison of PLS results with the results reported in the literature as presented in Table 5.3. For DHFR inhibitors data sets such comparison was not also possible, therefore the models with the best R^2 values in PLS studies were chosen among all other possible models calculated. R^2 value directly depends on the values of attributes (the descriptors) and the number of attributes in the corresponding model.

We first applied our iterative algorithm to the large and widely used QSAR data sets in order to validate our methodology. The strength of our algorithm was presented by comparing our classification accuracies with the classification accuracies of 63 WEKA classifiers, on 7 inhibitor sets. The attribute sets prepared as the input for WEKA classifiers were the same as the ones, by which we built the last iteration of our MILP based hyper-boxes classification model. In other words, those were the most significant attributes that we used to develop the final classification models and reached our best accuracies. Our approach outperformed all of the classifiers available in WEKA for each model of the all of the 7 data sets, even reaching 100% accuracy in predicting the activity classification of the inhibitor sets, Ache inhibitors and Cytochrome P450 C17. A total number of 21 QSAR models were built in this study for 7 inhibitor sets, and in 18 of them the accuracy of our methodology exceeded the accuracy of the second best classifier with more than 10%. Through all of the 21 models, the smallest difference in the accuracies is 6.31% and the largest difference is 27.47%.

The higher prediction accuracy of the model not only comes from the choice of initial models by PLS analysis but also the characteristics of MILP based hyper-boxes method. The MILP based hyper-boxes approach allows using more than one hyper-box in order to define a single class. Moreover, this approach deals with problematic and non-problematic instances separately and prevents overlapping of final hyper-boxes. Therefore, these strengths significantly improve the accuracy and efficiency of the MILP based hyper-boxes approach compared to other data classification methods.

Bibliography

1. Bohacek, R.S., C. McMartin, and W.C. Guida, *The art and practice of structure-based drug design: A molecular modeling perspective.* Medicinal Research Reviews, 1996. **16**(1): p. 3-50.
2. Kuntz, I.D., *Structure-Based Strategies for Drug Design and Discovery.* Science, 1992. **257**(5073): p. 1078-1082.
3. Klebe, G., *Recent developments in structure-based drug design.* Journal of Molecular Medicine-Jmm, 2000. **78**(5): p. 269-281.
4. Turkay, M., *Structural Synthesis of Small Molecule Drug Candidates*, in *FOCAPD 2004*. 2004. p. 395-398.
5. Alder, B.J. and T.E. Wainwright, *Studies in Molecular Dynamics. I. General Method.* J. Chem. Phys. , 1959. **31**(2): p. 459.
6. Streett, W.B., D.J. Tildesley, and G. Saville, *Multiple time-step methods in molecular dynamics* Molecular Physics, 1978. **35**(3): p. 639 - 648
7. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules.* Nature Structural Biology, 2002. **9**(9): p. 646-652.
8. Vitkup, D., et al., *Solvent mobility and the protein 'glass' transition.* Nature Structural Biology, 2000. **7**(1): p. 34-38.
9. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.
10. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
11. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-&.
12. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking.* Current Opinion in Structural Biology, 1996. **6**(3): p. 402-406.
13. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: Methods and applications.* Nature Reviews Drug Discovery, 2004. **3**(11): p. 935-949.
14. Jorgensen, W.L., *Rusting of the Lock and Key Model for Protein-Ligand Binding.* Science, 1991. **254**(5034): p. 954-955.
15. Wei, B.Q., et al., *Testing a flexible-receptor docking algorithm in a model binding site.* Journal of Molecular Biology, 2004. **337**(5): p. 1161-1182.
16. Jain, A.N., *Scoring functions for protein-ligand docking.* Current Protein & Peptide Science, 2006. **7**(5): p. 407-420.

17. Bohm, H.J., *Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs.* Journal of Computer-Aided Molecular Design, 1998. **12**(4): p. 309-323.
18. Muegge, I., *PMF scoring revisited.* Journal of Medicinal Chemistry, 2006. **49**(20): p. 5895-5902.
19. Mintseris, J. and M.B. Eisen, *Design of a combinatorial DNA microarray for protein-DNA interaction studies.* Bmc Bioinformatics, 2006. **7**: p. -.
20. Janin, J., et al., *CAPRI: A Critical Assessment of PRedicted Interactions.* Proteins-Structure Function and Genetics, 2003. **52**(1): p. 2-9.
21. Katchalskikatzir, E., et al., *Molecular-Surface Recognition - Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(6): p. 2195-2199.
22. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* Journal of Molecular Biology, 2003. **331**(1): p. 281-299.
23. Ritchie, D.W., *Macromolecular Docking Using Spherical Polar Fourier Correlations* Hex 4.2 User Manual 2003.
24. Eisen, J.A. and P.C. Hanawalt, *A phylogenomic study of DNA repair genes, proteins, and processes.* Mutation Research-DNA Repair, 1999. **435**(3): p. 171-213.
25. Ogrunc, M., et al., *Nucleotide excision repair in the third kingdom.* Journal of Bacteriology, 1998. **180**(21): p. 5796-5798.
26. Walker, G.C., *Understanding the complexity of an organism's responses to DNA damage.* Cold Spring Harbor Symposia on Quantitative Biology, 2000. **65**: p. 1-10.
27. Sancar, A., et al., *Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints.* Annual Review of Biochemistry, 2004. **73**: p. 39-85.
28. Sancar, A., *DNA excision repair.* Annual Review of Biochemistry, 1996. **65**: p. 43-81.
29. Wood, R.D., *Nucleotide excision repair in mammalian cells.* Journal of Biological Chemistry, 1997. **272**(38): p. 23465-23468.
30. Reardon, J.T. and A. Sancar, *Nucleotide excision repair.* Progress in Nucleic Acid Research and Molecular Biology, Vol 79, 2005. **79**: p. 183-235.
31. Sancar, A. and W.D. Rupp, *A Novel Repair Enzyme - Uvrabc Excision Nuclease of Escherichia-Coli Cuts a DNA Strand on Both Sides of the Damaged Region.* Cell, 1983. **33**(1): p. 249-260.
32. Huang, J.C., et al., *Human Nucleotide Excision Nuclease Removes Thymine Dimers from DNA by Incising the 22nd Phosphodiester Bond 5' and the 6th Phosphodiester Bond 3' to the Photodimer.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(8): p. 3664-3668.

33. Mu, D., D.S. Hsu, and A. Sancar, *Reaction mechanism of human DNA repair excision nuclease.* Journal of Biological Chemistry, 1996. **271**(14): p. 8285-8294.
34. Wakasugi, M. and A. Sancar, *Order of assembly of human DNA repair excision nuclease.* Journal of Biological Chemistry, 1999. **274**(26): p. 18759-18768.
35. Clugston, C.K., et al., *Binding of Human Single-Stranded-DNA Binding-Protein to DNA Damaged by the Anticancer Drug Cis-Diamminedichloroplatinum(Ii).* Cancer Research, 1992. **52**(22): p. 6375-6379.
36. Reardon, J.T., D. Mu, and A. Sancar, *Overproduction, purification, and characterization of the XPC subunit of the human DNA repair excision nuclease.* Journal of Biological Chemistry, 1996. **271**(32): p. 19451-19456.
37. Evans, E., et al., *Mechanism of open complex and dual incision formation by human nucleotide excision repair factors.* Embo Journal, 1997. **16**(21): p. 6559-6573.
38. Wakasugi, M. and A. Sancar, *Assembly, subunit composition, and footprint of human DNA repair excision nuclease.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(12): p. 6669-6674.
39. Gaillard, P.H.L. and R.D. Wood, *Activity of individual ERCC1 and XPF subunits in DNA nucleotide excision repair.* Nucleic Acids Research, 2001. **29**(4): p. 872-879.
40. Reardon, J.T., L.H. Thompson, and A. Sancar, *Rodent UV-sensitive mutant cell lines in complementation groups 6-10 have normal general excision repair activity.* Nucleic Acids Research, 1997. **25**(5): p. 1015-1021.
41. Al-Minawi, A.Z., N. Saleh-Gohari, and T. Helleday, *The ERCC1/XPF endonuclease is required for efficient single-strand annealing and gene conversion in mammalian cells.* Nucleic Acids Research, 2008. **36**(1): p. 1-9.
42. McHugh, P.J., *Repair of DNA interstrand crosslinks: molecular mechanisms and clinical relevance. .* Lancet Oncol. , 2001. **2**: p. 483-490.
43. Torresgarcia, S.J., et al., *Correlation of Resistance to Nitrogen Mustards in Chronic Lymphocytic-Leukemia with Enhanced Removal of Melphalan-Induced DNA Cross-Links.* Biochemical Pharmacology, 1989. **38**(18): p. 3122-3123.
44. Thompson, L.H., et al., *A Screening Method for Isolating DNA Repair-Deficient Mutants of Cho Cells.* Somatic Cell Genetics, 1980. **6**(3): p. 391-405.
45. Hoy, C.A., et al., *Defective-DNA Cross-Link Removal in Chinese-Hamster Cell Mutants Hypersensitive to Bifunctional Alkylating-Agents.* Cancer Research, 1985. **45**(4): p. 1737-1743.
46. de Laat, W.L., et al., *Mapping of interaction domains between human repair proteins ERCC1 and XPF.* Nucleic Acids Research, 1998. **26**(18): p. 4146-4152.
47. Tsodikov, O.V., et al., *Crystal structure and DNA binding functions of ERCC1, a subunit of the DNA structure-specific endonuclease XPF-ERCC1.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(32): p. 11236-11241.

48. Kortemme, T., D.E. Kim, and D. Baker, *Computational Alanine Scanning of Protein-Protein Interfaces.* 2004.

49. Kortemme, T. and D. Baker, *A simple physical model for binding energy hot spots in protein-protein complexes.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(22): p. 14116-14121.

50. Lazaridis, T. and M. Karplus, *Effective energy function for proteins in solution.* Proteins-Structure Function and Genetics, 1999. **35**(2): p. 133-152.

51. Kortemme, T., A.V. Morozov, and D. Baker, *An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes.* Journal of Molecular Biology, 2003. **326**(4): p. 1239-1259.

52. Kuhlman, B. and D. Baker, *Native protein sequences are close to optimal for their structures.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(19): p. 10383-10388.

53. Kubinvi, H., *Nonlinear dependence of biological activity on hydrophobic character: the bilinear model.* Farmaco [Sci], 1979. **34**(3): p. 248-76.

54. Eisenberg, D. and A.D. Mclachlan, *Solvation Energy in Protein Folding and Binding.* Nature, 1986. **319**(6050): p. 199-203.

55. Miyamoto, S. and P.A. Kollman, *What Determines the Strength of Noncovalent Association of Ligands to Proteins in Aqueous-Solution.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(18): p. 8402-8406.

56. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Advanced Drug Delivery Reviews, 2001. **46**(1-3): p. 3-26.

57. Ghose, A.K., V.N. Viswanadhan, and J.J. Wendoloski, *A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases.* Journal of Combinatorial Chemistry, 1999. **1**(1): p. 55-68.

58. Janin, J. and C. Chothia, *The Structure of Protein-Protein Recognition Sites.* Journal of Biological Chemistry, 1990. **265**(27): p. 16027-16030.

59. Clackson, T. and J.A. Wells, *A Hot-Spot of Binding-Energy in a Hormone-Receptor Interface.* Science, 1995. **267**(5196): p. 383-386.

60. Livnah, O., et al., *Functional mimicry of a protein hormone by a peptide agonist: The EPO receptor complex at 2.8 angstrom.* Science, 1996. **273**(5274): p. 464-471.

61. Satoh, T., et al., *Bioactive peptide design based on protein surface epitopes - A cyclic heptapeptide mimics CD4 domain 1 CC' loop and inhibits CD4 biological function.* Journal of Biological Chemistry, 1997. **272**(18): p. 12175-12180.

62. Mettlin, C., *Recent developments in the epidemiology of prostate cancer.* European Journal of Cancer, 1997. **33**(3): p. 340-347.

63. Carter, B.S., H.B. Carter, and J.T. Isaacs, *Epidemiologic Evidence Regarding Predisposing Factors to Prostate-Cancer.* Prostate, 1990. **16**(3): p. 187-197.

64. Hartmann, R.W., et al., *Inhibition of CYP 17, a new strategy for the treatment of prostate cancer.* Archiv Der Pharmazie, 2002. **335**(4): p. 119-128.

65. Mcconnell, J.D., *Physiological-Basis of Endocrine Therapy for Prostatic-Cancer.* Urologic Clinics of North America, 1991. **18**(1): p. 1-13.

66. Laughton, C.A., et al., *A Molecular-Model for the Enzyme Cytochrome-P45017-Alpha, a Major Target for the Chemotherapy of Prostatic-Cancer.* Biochemical and Biophysical Research Communications, 1990. **171**(3): p. 1160-1167.

67. Lin, D., et al., *Modeling and Mutagenesis of the Active-Site of Human P450c17.* Molecular Endocrinology, 1994. **8**(3): p. 392-402.

68. Auchus, R.J. and W.L. Miller, *Molecular modeling of human P450c17 (17 alpha-hydroxylase/17,20-lyase): Insights into reaction mechanisms and effects of mutations.* Molecular Endocrinology, 1999. **13**(7): p. 1169-1182.

69. Clement, O.O., et al., *Three dimensional pharmacophore modeling of human CYP17 inhibitors. Potential agents for prostate cancer therapy.* Journal of Medicinal Chemistry, 2003. **46**(12): p. 2345-2351.

70. Cougar Biotechnology, I. (2007) **Volume**,

71. Ideyama, Y., et al., *YM116, 2-(1H-imidazol-4-ylmethyl)-9H-carbazole, decreases adrenal androgen synthesis by inhibiting C17-20 lyase activity in NCI-H295 human adrenocortical carcinoma.* Japanese Journal of Pharmacology, 1999. **79**(2): p. 213-220.

72. Nnane, I.P., et al., *Inhibition of androgen synthesis in human testicular and prostatic microsomes and in male rats by novel steroidal compounds.* Endocrinology, 1999. **140**(6): p. 2891-2897.

73. Arlt, W., R.J. Auchus, and W.L. Miller, *Thiazolidinediones but not metformin directly inhibit the steroidogenic enzymes P450c17 and 3 beta-hydroxysteroid dehydrogenase.* Journal of Biological Chemistry, 2001. **276**(20): p. 16767-16771.

74. Arlt, W., et al., *Cinnamic acid based thiazolidinediones inhibit human P450c17 and 3 beta-hydroxysteroid dehydrogenase and improve insulin sensitivity independent of PPAR gamma agonist activity.* Journal of Molecular Endocrinology, 2004. **32**(2): p. 425-436.

75. Handratta, V.D., et al., *Novel C-17-heteroaryl steroidal CYP17 inhibitors/antiandrogens: Synthesis, in vitro biological activity, pharmacokinetics, and antitumor activity in the LAPC4 human prostate cancer xenograft model.* Journal of Medicinal Chemistry, 2005. **48**(8): p. 2972-2984.

76. Mendieta, M.A.E.P.B., et al., *Synthesis, biological evaluation and molecular modelling studies of novel ACD- and ABD-ring steroidomimetics as inhibitors of CYP17.* Bioorganic & Medicinal Chemistry Letters, 2008. **18**(1): p. 267-273.
77. Jagusch, C., et al., *Synthesis, biological evaluation and molecular modelling studies of methyleneimidazole substituted biaryls as inhibitors of human 17 alpha-hydroxylase-17,20-lyase (CYP17). Part I: Heterocyclic modifications of the core structure.* Bioorganic & Medicinal Chemistry, 2008. **16**(4): p. 1992-2010.
78. Moreira, V.M., et al., *CYP17 inhibitors for prostate cancer treatment - An update.* Current Medicinal Chemistry, 2008. **15**(9): p. 868-899.
79. Potter, G.A., et al., *Novel Steroidal Inhibitors of Human Cytochrome P450(17-Alpha) (17-Alpha-Hydroxylase-C-17,C-20-Lyase) - Potential Agents for the Treatment of Prostatic-Cancer.* Journal of Medicinal Chemistry, 1995. **38**(13): p. 2463-2471.
80. Benbow, J.W., et al., *Synthesis and evaluation of dinitroanilines for treatment of cryptosporidiosis.* Antimicrobial Agents and Chemotherapy, 1998. **42**(2): p. 339-343.
81. Golender, V.E. and E.R. Vorpagel, *Computer Assisted Pharmacophore Identification* in *3D QSAR in Drug Design: Theory, Methods and Applications*. 1993. p. 137-49.
82. Cramer, R.D., D.E. Patterson, and J.D. Bunce, *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins.* J. Am. Chem. Soc., 1988. **110**: p. 5959-5967.
83. Klebe, G. and U. Abraham, *Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries.* Journal of Computer-Aided Molecular Design, 1999. **13**(1): p. 1-10.
84. Turner, D.B. and P. Willett, *The EVA spectral descriptor.* European Journal of Medicinal Chemistry, 2000. **35**(4): p. 367-75.
85. Ferguson, A.M., et al., *EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis.* Journal of Computer-Aided Molecular Design, 1997. **11**(2): p. 143-52.
86. Tetko, I.V., *Computing chemistry on the web.* Drug Discovery Today 2005. **10**(22): p. 1497-500.
87. Tetko, I.V., et al., *Virtual Computational Chemistry Laboratory – Design and Description.* Journal of Computer-Aided Molecular Design, 2005. **19**(6): p. 453-63.
88. VCCLAB. *Virtual Computational Chemistry Laboratory*. 2005 [cited; Available from: http://www.vcclab.org.
89. Todeschini, R. and V. Consonni, *Handbook of Molecular Descriptors*. 2000: Wiley–VCH, Weinheim.

90. Whitley, D.C., M.G. Ford, and D.J. Livingstone, *Unsupervised forward selection: a method for eliminating redundant variables.* Journal of Chemical Information & Computer Sciences, 2000. **40**(5): p. 1160.

91. Tetko, I.V., *Neural Network Studies, 4. Introduction to Associative Neural Networks.* Journal of Chemical Information & Computer Sciences, 2002. **42**(3): p. 717-728.

92. Tetko, I.V., *Associative neural network.* Neural Proc. Lett, 2002. **16**(2): p. 187-99.

93. Aksyonova, T.I., V.V. Volkovich, and I.V. Tetko, *Robust Polynomial Neural Networks in Quantitative-Structure Activity Relationship Studies.* SAMS, 2003. **43**(10): p. 1331-9.

94. Tetko, I.V., et al., *Polynomial neural network for linear and non-linear model selection in quantitative-structure activity relationship studies on the internet.* SAR and QSAR in environmental research, 2000. **11**: p. 263-80.

95. Garthwaite, P.H., *An interpretation of partial least squares.* Journal of the American Statistical Association, 1994. **89**(425): p. 122-7.

96. ChemAxon, *Marvin 4.1.7.* 2005.

97. Broughton, M.C. and S.F. Queener, *Pneumocystis carinii dihydrofolate reductase used to screen potential antipneumocystis drugs.* Antimicrobial Agents and Chemotheraphy, 1991. **35**(7): p. 1348-55.

98. Gangjee, A., et al., *Synthesis and Biological Activities of Tricyclic Conformationally Restricted Tetrahydropyrido Annulated Furo[2,3-d]pyrimidines as Inhibitors of Dihydrofolate Reductases.* Journal of Medical Chemistry, 1998. **41**(9): p. 1409 -16.

99. Rosowsky, A., J.B. Hynes, and S.F. Queener, *Structure-activity and structure-selectivity studies on diaminoquinazolines and other inhibitors of Pneumocystis carinii and Toxoplasma gondii dihydrofolate reductase.* Antimicrobial Agents and Chemotheraphy, 1995. **39**(1): p. 79–86.

100. Sutherland, J.J., L.A. O'Brien, and D.F. Weaver, *A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships.* Journal of Medical Chemistry, 2004. **47** (22): p. 5541 -5554.

101. Sugimoto, H., et al., *Synthesis and structure-activity relationships of acetylcholinesterase inhibitors: 1-benzyl-4-[(5,6-dimethoxy-1-oxoindan-2-Yl)methyl]piperidine hydrochloride and related compounds.* Journal of Medical Chemistry, 1995. **38**: p. 4821-29.

102. Sugimoto, H., et al., *Novel piperidine derivatives.Synthesis and anti-acetylcholinesterase activity of 1-benzyl-4-[2-(N-benzoylamino)ethyl]piperidine derivatives.* Journal of Medical Chemistry, 1990. **33**: p. 1880-87.

103. Sugimoto, H., et al., *Synthesis and structure-activity relationships of acetylcholinesterase inhibitors: 1-benzyl-4-(2-phthalimidoethyl)piperidine and related derivatives.* Journal of Medical Chemistry, 1992. **35**: p. 4542-48.
104. Haefely, W., et al., *Recent advances in the molecular pharmacology of benzodiazepine receptors and in the structure-activity relationships of their agonists and antagonists.* Advances in Drug Research, 1985. **14**: p. 165-322.
105. Carter, J.S., et al., *Design and synthesis of sulfonylsubstituted 4,5-diarylthiazoles as selective cyclooxygenase-2 inhibitors* Bioorganic & Medicinal Chemistry letters, 1999. **9**(8): p. 1167-70.
106. Huang, H.C., et al., *Diaryl indenes and benzofurans - novel classes of potent and selective cyclooxygenase-2 inhibitors.* Bioorg. Med. Chem. Lett., 1995. **5**: p. 2377-80.
107. College, M.I.-S., *MINITAB Statistical Software, Release 14 for Windows*. 2003: Pennsylvania.
108. Üney, F. and M. Türkay, *A mixed-integer programming approach to multi-class data classification problem.* European Journal of Operational Research, 2006. **173**(3): p. 910-20.
109. Yuksektepe, F.Ü., O. Yilmaz, and M. Turkay, *Prediction of Secondary Structures of Proteins using a Two-Stage Method.* Computers & Chemical Engineering, 2008. **32**(1-2): p. 78-88.
110. Plewczynski, D., S.A.H. Spieser, and U. Koch, *Assesing Different Classification Methods for Virtual Screening.* J. Chem. Inf. Model., 2006: p. 1098-1106.
111. Cheng, J. and R. Greiner, *Comparing Bayesian Network Classifiers*. 1999, Department of Computing Science University of Alberta: Alberta.
112. Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2nd Edition ed, ed. M. Kaufmann. 2005, San Francisco.
113. Pinardan, P.J., *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Editors. 1998, MIT Press.
114. Landwehr, N., M. Hall, and E. Frank, *Logistic Model Trees*, in *14th European Conference on Machine Learning, ECML 2003*. 2003: Croatia.
115. Tax, D.M.J. and R.P.W. Duin, *Using two-class classifiers for multiclass classification.* 16th International Conference on Pattern Recognition (ICPR'02), 2002. **2**: p. 20124.
116. Gangjee, A., O. Adaira, and S.F. Queener, *Synthesis of 2,4-Diamino-6-(thioarylmethyl)pyrido[2,3-d]pyrimidines as dihydrofolate reductase inhibitors.* Bioorganic & Medicinal Chemistry 2001. **9**(11): p. 2929-35.

117. Clement, O.O., et al., *Three Dimensional Pharmacophore Modeling of Human CYP17 Inhibitors. Potential Agents for Prostate Cancer Therapy.* Journal of Medical Chemistry, 2003. **46**(12): p. 2345-51.

118. Handratta, V.D., et al., *Novel C-17-Heteroaryl Steroidal CYP17 Inhibitors/Antiandrogens: Synthesis, in Vitro Biological Activity, Pharmacokinetics, and Antitumor Activity in the LAPC4 Human Prostate Cancer Xenograft Model.* Journal of Medical Chemistry, 2005. **48**(8): p. 2972-84.

VITA

Pelin Armutlu was born in Mersin, in May 26, 1983. She is a graduate of Icel Anatolian High school and received her Bachelor of Science degree from the department of Industrial Engineering at Koc University, Istanbul, in June 2006. She was a research and teaching assistant at Koc University from 2003 to 2008 at the College of Engineering.

Her research interests include biophysics and computational biology mainly focusing on QSAR models and drug discovery.