
MIXED-INTEGER OPTIMIZATION APPROACH FOR
FRAGMENT-BASED DESIGN OF DRUG CANDIDATES

by

Zeynep Göksel Özsarp

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

July, 2013

ABSTRACT

Developing computational methods for drug design has become one of the most widely studied areas of life sciences in the past two decades. The main objective is to discover new drugs using the information regarding topology and interaction energies of known drugs, that minimizes binding energy and docking energy. Designing new drugs from the fragments of known drugs rather than libraries of larger molecules provides more combinations in products. From these many possible combinations it is possible to find new drugs that have lower energy values than existing molecular structures.

Experimental methods were applied to screen thousands of low-molecular-weight compounds for testing their binding to the target protein in using fragment based drug design effectively. Successful results in fragment based drug design were only obtained with the strong integration of computational techniques, innovative NMR experiments and X-Ray crystallography. Many existing computational algorithms cannot accurately predict the affinity with which fragments might or might not bind to the protein surface alone. A computational approach comprised of learning and design phases is proposed in this thesis. First, we apply a learning algorithm on known results of docking and binding energies to determine weights for the each fragment reliably. Then, in the design part, the multi-objective optimization model is formed based on the results obtained in the first step, to predict the contribution of each group and individual fragment in docking energy and binding energy and finally designs novel molecules from available fragments.

Özet

İlaç tasarımı için bilgisayar metotlarının geliştirilmesi, son yirmi yıldır en çok çalışılan alanlardan biri olmuştur. Bu çalışmalardaki asıl amaç, bilinen ilaçlardaki topoloji ve etkileşim enerji değerleri göz önünde bulundurularak bağlanma ve kenetlenme enerjileri en düşük olacak şekilde yeni ilaçlar keşfetmektir. Büyük moleküllerce oluşturulan kütüphaneleri kullanmak yerine bilinen ilaçların parçalarını kullanarak yeni ilaçlar dizayn etmek, daha çok kombinasyon oluşumu sağlamaktadır. Bilinen ilaçların parçaları kullanılarak oluşturulan bu kombinasyonlar ile, var olan moleküllerden daha düşük enerjili ilaçların tasarlanması mümkündür.

DeneySEL metotlar, parça bazlı ilaç tasarımı etkili bir şekilde uygulamak adına düşük moleküler ağırlıklı moleküllerin hedef proteine bağlanmalarını test etmek için binlerce ürün taramaktadırlar. Parça bazlı ilaç tasarımı başarılı sonuçlar ise sadece bilgisayar metotlarının yenilikçi NMR deneyleri ve X-Ray kristalografisi ile etkileşimi durumunda elde edilmektedir. Varolan birçok bilgisayar tekniği, hangi parçanın protein yüzeyine etkili bir şekilde bağlanacağını düzgün olarak tahmin edememektedir. Bu çalışmada, ilaç tasarımı öğrenme ve dizayn basamaklarından oluşan bilgisayarlı bir yöntem geliştirilmiştir. İlk olarak, yapısı bilinen moleküllerin kenetlenme ve bağlanma enerjileri kullanılarak, yapılarıdaki parçalara enerji değerlerine katkılarına göre ayrı ayrı ağırlık katsayıları belirleyen bir öğrenme algoritması uygulanır. Daha sonra dizayn kısmına geçilir ve ilk adımda elde edilen ağırlık katsayıları kullanılarak en düşük kenetlenme ve bağlanma enerjilerini veren parçalar kullanılarak tamamen yeni ilaç tasarımları yapılır.

ACKNOWLEDGMENTS

I thank Assoc. Prof. Metin Türkay for believing in me and for his patience, guidance and support during my graduate study and the thesis process. Also, I thank to Dr. Halil Kavaklı and Dr. Onur Kaya for participating in my thesis committee and for their insight.

I would like to thank my dear Buse for always being with me, and Yiğit for his support at any time. I will forever be grateful for having them together as my friends and office mates. It is also impossible to ignore the excellent friendship of my sweet Ayşe Nur and Zehra. Also, working on my thesis would be more difficult without Can. Thank you for your support and your great friendship. In addition to the many skills I gained during my research, I also had excellent friends like Doğuş, Yasin, Yahya, Emin, Gökhan, Müge, Uğur. In addition, I would like to thank my past and present house mates; Müge, İdil, İlknur, Pınar, Dilek and Bilge for their continued friendship, supports, hugs, laughs and cups of tea along the way. Thank you all.

Also, I have special friends that their support will never be forgotten. I would like to thank Mustafa for pointing me in the right direction and continuing to encourage me throughout, Talha for his moral and scientific support at any time and Abdullah for helping me in this thesis even late nights and early mornings.

I would like to thank Nafize Kocabıyık and Ayşe Morgül. Your support and presence is invaluable for me. Thank you for all the support and encouragement you give me, and the patience and unwavering faith you have in me.

I am very fortunate to have such a family, thank you for supporting me throughout the years, financially, practically and with moral support and I would like to dedicate this thesis to the memory of my dear grandfather, Bekir Yetiş who is a constant source of my inspiration. I hope he would have been proud.

TABLE OF CONTENTS

List of Tables	7
List of Figures	8
Chapter 1: Introduction	1
Chapter 2: Literature Review	4
2.1 The Drug Design Process	4
2.1.1 Principles	4
2.1.2 Molecular Interactions	6
2.2 Fragment Based Drug Design	7
2.2.1 Strategic Advances	9
2.2.2 Fragment Based Design Algorithms	11
2.3 WEKA: Data Mining Software	16
2.3.1 Prostate and Prostate Cancer	18
2.3.2 Drug Design Studies for the Treatment of Prostate Cancer . .	19
2.4 Nonlinear Programming Problems	20
2.4.1 Nonlinear Programming Solution Techniques	21
2.5 Least Square Regression Method	22
Chapter 3: Materials and Methods	24
3.1 Learning Phase	28
3.1.1 AutoDock	28
3.1.2 Decision of Scaffold for Fragment Based Drug Design	32
3.1.3 The Proposed Design Methodology	36

Chapter 4: Results and Discussion	40
4.1 First Dataset	40
4.2 Second Dataset	45
Chapter 5: Conclusion and Future Work	50
Bibliography	53

LIST OF TABLES

3.1	Possible fragments for the each position on the scaffold for the first dataset	34
3.2	Possible fragments for the each position on the scaffold for the second dataset	35
4.1	AutoDock and estimated energies of optimal combinations	43
4.2	AutoDock and estimated energies of first ten combinations	47

LIST OF FIGURES

3.1	Androgen synthesis pathway	25
3.2	Target CYP17 structure	26
3.3	Flowchart of the model	27
3.4	The main features of the grid map [1]	30
3.5	The lead compound for prostate cancer treatment	33
3.6	First scaffold	34
3.7	Second scaffold	34
4.1	. The R^2 value of the estimation error r_1 for the docking energy of the first dataset	41
4.2	The R^2 value of the estimation error r_2 for the binding energy of the first dataset	42
4.3	Optimal solution for the first dataset	42
4.4	Optimal solutions for the first dataset	44
4.5	The optimal solution of the training set	45
4.8	Optimal solution for the second dataset	45
4.6	The R^2 value of the estimation error r_1 for the docking energy of the second dataset	46
4.7	The R^2 value of the estimation error r_2 for the binding energy of the second dataset	46
4.9	AutoDock and estimated energies of optimal combinations	48
4.10	The trend between calculated and estimated docking and binding energies for the second dataset	48

Chapter 1

INTRODUCTION

Drugs are substances used for the inhibition of the protein (most cases enzymes) activation that causes diseases. The drug discovery process is aimed at bringing to market new therapeutic agents with desirable pharmacodynamic profile, favorable ADME (Absorption, Distribution, Metabolism, Elimination) and toxicological properties [2] considering the cost and the time of production. According to the annual report of PhRMA the amount of time that was needed to commercialize a single drug was in the order of 5 years and the amount of money invested was in the order of \$2,000,000,000 per drug approved before the investigation of computational techniques [2]. The huge cost and time consumption on drug design process has yielded the drug industry to investigate some computational methodologies to reduce cost and time.

In the rational drug design methods, drug discovery was a trial and error process of substances on animals or cultured cells and analyzing the effects, which is costly and time consuming. However, there is a growing realization that given the enormous size of organic chemical space (possibly $>10^{18}$ compounds), the drug discovery cannot be reduced to a simple empirical method. To this end, a battery of in vitro ADME screens has been implemented in most pharmaceutical companies with the aim of discarding compounds in the discovery phase that are likely to fail further down the line [3]. In these experimental filters, there are also some limitations such as requiring physical samples for testing, being resource-intensive and remaining time consuming as well.

For this reason, researchers choose to give the compounds making the best possible

chance of becoming drugs computationally and create a subset of 'drug-like' molecules from the vast expanse of what could possibly be synthesized [4]. So, creating such a subset by computational drug design can significantly simplify the job with decreasing the cost and time. Although, the perfect computational drug design process has not been developed yet, the structure and fragment based drug design are widely used by researchers in academia and industry.

In this thesis, computational fragment based drug design has been used for the treatment of prostate cancer. In order to achieve this, a systematic and detailed two-stage methodology is followed. The first stage is the learning phase that begins with the selection of the main scaffold, fragments and the positions of the fragments for the given scaffold for the deactivation of the target protein (CYP17 in our case). Then, all possible combinations were determined by trying each fragments for all given positions and binding and docking energies of all possible combinations were calculated by a computational tool AutoDock. The weights for each fragment were determined by linear regression in the final step of the first stage.

The second stage is the design phase that determines the best fragments for each position with minimum docking and binding energies for the given scaffold using weight coefficients determined in the learning phase. Then, a priority order is provided in the given dataset by a multi objective optimization algorithm. Finally, a subset including compounds that have the smallest docking and binding energies from the dataset with possible drug formations have been formed.

Chapter 2 provides necessary background and literature review on the methods in computational drug design. The importance of the prostate cancer and the treatment process, the principles of the drug design and computational fragment based drug design algorithms in the literature are given in this chapter.

Chapter 3 is dedicated to the methodology and algorithm designed for the computational fragment based drug design in this study. Following the aim and importance of the selection of the CYP17 protein as a target, the chapter continues with the learning phase section. Autodock tool in detail, the selection of the scaffolds and the fragments and the algorithm are also given in this section. Finally, the design phase

methodology and the application of the algorithm that gives a priority order in the given dataset is provided.

Results and the discussion of the proposed algorithm is given in the Chapter 4. First, the optimal solutions for the scaffolds are provided and the training set formations are explained. Then, training set results and providing the priority order in the datasets are given. Finally, the docking and the binding energy values of the calculated by AutoDock and estimated by the algorithm are compared and discussed. Finally, this thesis is concluded with a summary of the performed study and future work in Chapter 5.

Chapter 2

LITERATURE REVIEW

2.1 The Drug Design Process

2.1.1 Principles

ADME requirements are the milestones of the drug design process which are the adsorption, distribution, metabolism and excretion. The drug candidate must be membrane permeable, as small as possible and soluble. In addition, the drug candidate should only bind to the target, not to other molecules that have important metabolic duties. Also, the toxicity levels of the drug candidate should be minimized [5].

Since 1997, the 'rule of five' by Lipinski et al. [6] is used to characterize the drug candidates as first filtering step. According to Lipinski et al. [6], a drug-like molecule should satisfy the given five parameters below.

- i Not more than 5 hydrogen bond donors
- ii Not more than 10 hydrogen bond acceptors
- iii A molecular weight under 500g/mol
- iv A partition coefficient $\log P$ less than 5

These are the threshold values for parameters of adsorption, distribution, metabolism and excretion. Lipinski's 'rule of five' provides finding drug candidates but is not enough to discover a successful novel drug completely. After the pioneering work by

Lipinski, many other properties that affect oral bioavailability are discussed such as molecular flexibility, or polarity of the surface area [7].

Binding affinity is another important characteristic for computational drug design. This feature is related with the tendency to the binding process energetically. In binding process, the temperature that the process takes place in (T), the enthalpy change (ΔH) and the entropy change (ΔS) plays important role. Therefore, binding affinity is directly related to the Gibbs free energy of binding (ΔG) where;

$$\Delta G = \Delta H - T \quad (2.1)$$

To maximize the binding affinity that means a successful binding process, ΔG must be as small as possible. Minimizing ΔG supplies stability to the molecule. Therefore, in many cases, instead of maximizing the entropy, it is more preferable to minimize the enthalpy. Maximizing entropy provides the molecule flexibility and spontaneity, however, if this value is too high than the stability of the drug would be disturbed.

The attention is generally focused on specific activity of the receptor-like interactions or small organic ligand molecules with a well defined biological target during the search for effective drugs and other bioactive compounds [8]. Atoms and bonds properties are the local physicochemical features to control the nature and the strength of such interactions. Proper comparison of activity and local molecular properties within a single structure and between various congeneric structures are the key elements to analyze the structure-activity relationships. These properties can also be used to provide predictive statistical modeling with bioactivity parameters. In addition, such statistical modeling lead to further optimization of activity for drug mechanisms. It is easy to relate the mentality of organic and medicinal chemist with topological approaches including synthesis planning and the design of novel promising structures.

One of the most commonly used analysis method for local molecular properties is 3D-based approach. This approach starts with 3D model of atom positions. To compare different molecules, some uniform representations for structural features are proposed. Instead of individual atoms, this representation is linked to the molecular

axes of inertia or to an abstract spatial grid, thus avoiding the problem of matching the atoms of different structures [8]. 3D-based approaches reflect the actual 3D nature of the biological target and ligand better. However, generally the relation between the uniform representation and the molecular structure is not well defined and the construction of the model and the application to the virtual screening become complicated. In addition, 3D approaches are kind of info-noise and includes a lot of data on the particular details of conformational behavior, molecular structures and physicochemical parameters of the compounds.

The classical method of 3D analysis is the Comparative Molecular Field Analysis (CoMFA) technique introduced by R.Cramer et al. in 1988 (top book ref 2). This is the most commonly used method for almost 20 years for the creation of several approaches. The main aim of the method is to identify the spatial regions around the molecule where certain local properties have a positive or negative effect on activity [8]. The non-covalent interactions like van der Waals and Coulomb forces that are controlled by the shape of molecules between the organic ligand and the biotarget is the key point on the foundation of CoMFA approach.

The key feature of this approach is the assumption of considering the biological action of compounds as the electrostatic field of their molecules as a quantitative measure. A descriptor matrix is defined by calculated energies of van der Waals and Coulombic interaction of a molecule in a rectangular 3D grid. In contrast to the multiple linear regression, Partial Least Square Regression method is used which allows the predictive statistical relationship to be detected even if the number of descriptors is much greater than the number of the experimental data points.

2.1.2 Molecular Interactions

A successful drug molecules should have a strong binding affinity. Addition of the noncovalent bonds between the drug and the target protein increases the binding affinity. This more stable condition can be satisfied with addition of the atoms to the drug molecule which causes the enlargement in the molecule structure. However, Lipinski rules should be considered as well during the drug design while increasing

the binding affinity. Therefore, the balance between the increments of binding affinity without disturbing the Lipinski's constraints is important for a successful drug design.

There are different types of interactions between the drug and the target protein. Ionic bonds occur among the oppositely charged atoms. Hydrogen bonds are formed by polarization between a hydrogen acceptor and a hydrogen donor (or a hydrogen atom bonded to a N, F or I atom). Hydrogen acceptors create an electronegative nature. The final interaction is Van der Waals interaction which occurs between every atom couple that is close enough.

If the bond is strong then the energy released is more while breaking the bond. The strongest interaction is the formation of ionic bond, second is the hydrogen bond and the weakest one is the van der Waals interaction. Hydrogen bonds bear about one third of this energy, and van der Waals interactions are quite weak, having energy approximately one tenth of ionic bonds have [9].

In this thesis, considering these different types of interactions, with increasing binding affinity, minimizing the binding energy was one of the aims to create a novel drug from known molecules.

2.2 Fragment Based Drug Design

The drug discovery becomes widespread in industry and academia since these methods have been successfully generating new drug leads with high potency and improved pharmacokinetic properties. Most of the medicinal chemistry leads were developed from hits that are obtained from screening collections of compounds against functional assays with large molecules. However, it was discovered that as the molecular size decreases, the number of possible molecules decreases exponentially so it would be more efficient to screen collections of fragments and subsequently expand, merge or link them. Once fragment is identified, the most conceptually straightforward approach of advancing fragments, optimizing through chemical elaboration, generally requires highly specific or energetically favorable neighboring contacts to succeed [10]. This situation occurs if the nucleating fragment supplies much of the total necessary

binding energy. Metal coordination, mechanism based transition state analogues or very deep and well-defined pockets are the example of binding energy supplier groups. Anchoring sites are also in this class providing the necessary binding energy to enable high affinity association with a small molecule. In terms of recognizing small-molecule substrates, the enzyme molecules behave like anchoring sites. These molecules were recognized by high-throughput screen but it is suggested that they could be discovered by fragment optimization strategy as well.

Fragment optimization strategies can be facilitated by methods that guarantee the fragments bind noncompetitively with one another. Structural methods like SAR by NMR and SAR by X-ray are used to observe whether two fragments can bind simultaneously and in some cases even facilitate linking by providing the orientation of the fragments. However, structural information is not always essential because the technology like SAR by MS are used to determine two molecules that could bind simultaneously. To date, most successful applications of fragment based methods, particularly those involving fragment linking, have taken advantage of existing knowledge of the system, such as known cofactors, ligands and mechanistic consideration [10]. Structural information is used to guide the process and design the fragments.

The fragment-based drug design methods are still new and their development and validation have generally relied on using targets that often have known structures and that have already been subjected to other methods of lead discovery [10]. The identification of the hits that are worth pursuing is another hurdle for the fragment-based methods because not all inhibitors can become leads. It has been observed that some chemotypes could yield false positive results by reacting irreversibly with the target protein or interfering with colorimetric or fluorescent assays for highly colored compounds. In addition, many small molecules could form aggregates that non-specifically inhibits the protein function. Even the determination of such defects are possible, it is unclear that which molecular properties lead to this properties. Therefore, it is crucial to preserve the integrity of the fragments during new fragment based discovery process. The fragment based drug design method is open to evolve as the complement of other discovery approaches. Fragment-based methods will be

absorbed into a holistic approach to drug discovery, where fragments will be expanded and combined into libraries for functional screening, HTS hits will be dissected into component fragments for individual optimization, and the modelling techniques underpinning structure-based drug design will be called upon routinely [10].

2.2.1 Strategic Advances

Since the early 1990, there have been many scientific and technological developments in sequencing of the human genome and combinatorial chemistry fields. Fragment based drug design is a tool in combinatorial chemistry field that is developed recently.

In drug discovery field, there are two main issues for the target-based approach. First is to determine the activity of the biological target that causes the human disease must be identified. Second is to develop a therapeutic agent that will block the activity of this target without causing any toxic or hazardous effect. Competing with these problems depend on the quality and the efficiency of the drug design and discovery process. To improve the chances of finding agents that are active against these targets, technologies such as combinatorial chemistry and ultra-high-throughput screening (HTS) approaches have considerably expanded the numbers of compounds that can be evaluated for their biological activity [2]. To decrease the number of combinations, virtual ligand screening and structure-based drug design can be used. Novel drugs have been obtained by these methods but the investment to the new medicines are still the same.

Fragment based drug design is a new approach that has the advantage of increasing productivity in drug design. It is first demonstrated a decade ago [2]. The fragmentation of drug leads into smaller pieces, or even into discrete functional groups (for example, carboxylate, amine, aryl group and so on), has been used for some time to simplify the computational analysis of ligand binding and to map out different pharmacophoric elements required for high-affinity binding [2, 11]. In this approach, instead of considering whole molecule, the optimization of binding affinity is calculated as the sum of the individual fragment interactions. However, there is no such

a computational technique that exactly predicts the binding affinity of a fragment to protein surface. Therefore, experimental methods and spectral editing are used to determine the interaction of fragments with recent technology.

Nevertheless, there was a significant amount of internal resistance to resourcing the experimental pursuit of fragment leads at Abbott, as it was commonly believed that such low-molecular-mass, low-affinity ligands, even if they could be detected, would not form a unique and stable complex with the protein that could be productively used in drug design [2]. However, in P. Hajduk and J. Greer's work, if the fragments were soluble at the test concentrations, it has been proved that stable binding modes could be observed even with millimolar ligands for the protein.

P. Hajduk and J. Greer used NMR technique to determine this structure-activity relationships (SAR). They first designed a high-affinity inhibitors of the MMPs, which is a family of zinc-dependend endopeptidases that are implicates diseases like arthritis and tumor metastasis. After fragment screening performances, they discovered that acetohydroxamate (a zinc-chelating moiety with a K_d value of 11 μM for the protein) could bind to the protein simultaneously with a number of biaryl compounds (with K_d values in the 20100 μM range) [2]. Then they started lead optimization to increase the oral bioavailability of the series and to redirect potency against MMP2 and MMP9. In ABT-518, perfect oral antitumor efficacy in animal trials was obtained. They conclude that a single molecule could be designed and medicinal chemistry optimization is ultimately yielded.

For 10 years, there has been 49 highly potent ($IC_{50} < 100$ nM) drug that are developed by fragment-based drug design method for various protein targets. With the increasing popularity of fragment-based drug design, drug companies such as AstraZeneca, ScheringPlough, and Aventis have started to use this method. However, many companies quickly developed alternative NMR-based approaches that obviated the need for isotope labeling and facilitated screening on larger numbers of targets [12, 13]. Leaders in these approaches have been Novartis²⁷, Vertex²⁸ and Pharmacia (now Pfizer) [14, 15].

Fragment-based drug design appeared as a necessity for the small-molecule drug

discovery by the medicinal chemists whose job is to produce high affinity drugs. Leaders in the pharmaceutical industry have independently realized that chasing after potency at the expense of other physicochemical properties (such as lipophilicity, polarity, charge, stability, and so on) carries serious risks of failure owing to inadequate pharmacokinetic properties of the resulting compound [16]. This has spawned a whole new movement (the 'lead-like' movement) away from the use of large lipophilic compounds as leads towards smaller compounds that will have reasonable chances of possessing good pharmacokinetic properties after the optimization process is complete [17, 18]. To achieve the best balanced combinations in terms of potency and pharmacokinetic properties, fragment-based design provides the biggest support.

The concept of the fragment-based drug design seems simple and elegant when dealing with single ligand. However, simply identifying multiple ligands will not guarantee success [19]. In general, it is not likely to find the binding pocket for multiple fragments. Therefore, in fragment-based design the designer must allow some unexpected results that can provide new opportunities for further research.

2.2.2 Fragment Based Design Algorithms

Main Characteristics of Fragment Descriptors

A tremendous number of various fragments are used in structure-property studies: atoms, bonds, topological torsions, chains, cycles, atom- and bond-centered fragments, maximum common substructures, line notation fragments, atom pairs and topological multiplets, substituent and molecular frameworks, basic sub graphs, etc. [8]. Fragment descriptors are two types; binary and integer, depending on the application area. The presence and absence of the given fragment is indicated by the binary values and they are generally used as element fingerprints for chemical database and virtual screening. However, for QSAR studies, integer values for the occurrences of fragments in structures are used.

The simplest types of fragments are the disconnected atoms. Atomic contributions assess chemical or biological properties as an additive based approach with a chemical

or biological property P:

$$P = \sum_{i=1}^N n_i \cdot A_i \quad (2.2)$$

where n_i is the number of atoms of type i , A_i is the corresponding atomic contributions. The atom types generally accounts for the hybridization, attached hydrogen atom numbers for heavy elements and occurrence in aromatic systems. Another type of simple fragment is the chemical bonds. Considering the thermodynamic properties is the only difference between the atom-based works and chemical bonds type work.

Topological torsions are defined as a linear sequence of four consecutively bonded non-hydrogen atoms [8]. The number of attached non-hydrogen atoms, pi-electron pairs and type of corresponding chemical element are the parameters to describe the atoms. Qualitative predictions for the presence or absence of topological torsions are made by structure-activity studies.

The main aim is to obtain the fast assessment of usefulness of molecules according to the given rules and eliminate compounds with unfavorable pharmacodynamic, pharmacokinetics and toxic properties in terms of drug design. Binding of the drug-like organic fragments to chosen biological target is considered by pharmacodynamics. In addition, pharmacokinetics is mostly related to absorption, distribution, metabolism and excretion related properties: octanol-water partition coefficients ($\log P$), solubility in water ($\log S$), blood-brain coefficient ($\log BB$), partition coefficient between different tissues, skin penetration coefficient etc. [8]. To filter the large database, the easiest way is to detect the undesirable molecular fragments by structural alerts like toxicity, mutagenicity and carcinogenicity. However, Lipinski rule of five is the most popular filter used in the drug design area.

The use of information technology and management has become a critical part of the drug discovery process. Mixing of those information resources to transform data into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization is the key point for the computational drug design methods. In this thesis, the main aim is to find out the best fragment

combination to create a novel drug with fragment based drug design method from the fragments of known drugs by minimizing the binding and docking energies. To minimize these energies, an optimization model is created and the new structure is found out computationally. It is crucial to summarize the computational works done up to now to show that this thesis includes completely a new idea on computational fragment based drug design.

Computational approaches on ligand design appeared first in the late 80s, as an alternative to the high throughput screening methods. The main usage was to find the docking energy of the drug to the target protein. Recently, there are many de novo design algorithms that use various types of building blocks, search methods and scoring functions [20]. However, in these approaches, it is difficult to filtrate the successful products among all. The filtration process again needs to evaluate all the drugs in terms of combinatorial difficulties and pharmacokinetics which is time consuming and costly.

Techniques for Finding Fragments

Developing libraries of fragments is the first step of the fragment-based drug discovery. There are some unique considerations for fragment libraries. For example, fragment should be smaller than typical HTS compounds because the fragments will ultimately be elaborated. Scientists at Vertex Pharmaceuticals computationally dissected known drugs into fragments corresponding to molecular frameworks and side chains, these analysis demonstrated that most drugs can be represented by a relatively small set of molecular architectures [10]. After this investigation, a small library including fewer than 200 fragments, especially designed for NMR screening, called SHAPES library [21]. The compounds were chosen from known drugs as be soluble at high concentrations, nonreactive and commercially available. Then, Lewell and colleagues described the use of a retrosynthetic combinatorial analysis procedure *RECAP* to identify recurring fragments from known drugs [22]. Fesik and colleagues have proposed enriching fragment libraries with privileged molecules, such as biphenyls, that have been experimentally shown to bind to proteins frequently [23]. Fragment library

design has been reviewed more recently [24] with the computational deconstruction of drugs into fragments remaining an active research focus [25]. In terms of the targets with fairly rigid sites, the virtual screening methods can be used to augment default libraries with fragments selected on the basis of their structural complementarity to the protein.

Here are some methods used for the fragment-based drug design algorithms.

LUDI [26] is a fragment-based design algorithm developed in 1991. As input, PDB structure, position of active site and possible fragments to bond the active site are necessary for the program. The working principle is to create hydrogen bonds between the fragments that are chosen from a fragment library as initial step. This library includes energetically optimized conformations calculated by available energy field CVFF [27]. According to bond distance and angle preferences, algorithm creates all of the conformations of aliphatic, aromatic, hydrogen donors and acceptors in 'influence region'. Then, all of the conformations are tried on active site to find the best fit. Finally, the successfully fitted fragments are linked with 'bridges' from another library. These bridges are used to connect the closest fragments of a closest hydrogen atoms and the heaviest atom. The most suitable bridge for the situation is determined by the program.

HOOK [28] is another program that requires the coordinates of the active sites, 'skeleton' database and functional regions to achieve the fragment-based drug design. The fragments are called as skeletons in this program and the main aim is to hook the skeletons to the active site by non-covalent interactions. If there is no attraction above a certain level on a grid point, that point is called 'vacant' and if the certain attraction level is satisfied, then that grid point is called 'donor', 'acceptor' or 'hydrophobic'. Then, the program calculates an overlap score by comparing the hooks on skeletons and the geometry of the functional sites using Lennard-Jones potential which gives a certain interaction threshold to be satisfied.

LigBuilder [29] is another fragment-based design algorithm. In the first step, POCKET [30] is used to determine the active binding sites as grids. The algorithm automatically determines a structure from fragments initially. Hydrogen atoms of

both active site and the fragment are determined and covalent bonds are formed between the heavy atoms and hydrogen atoms. Then, energetically the most favorable structure which has the lowest ΔG value is searched by turning the bond by 15 increment. However, because of every favorable conformation is kept in memory as a different molecule, the number of possible structures is too much. Therefore, genetic algorithm is used which keeps 10% of the best members from old generation in new generation. The rest in the new generation is selected randomly from the fragment library. The final generation is the best drug candidate.

As a different methodology than the others, CONCERTS [31] by Pearlman et al. can be shown. This method uses molecular dynamics on database of fragments together within the active site. In that algorithm, the first step is to provide interactions between the active site and the fragments preventing the interactions between the fragments. Then, in each iteration, bonds between fragments are formed. Finally, the algorithm stops when the energetically favorable conformation is found.

Converting Fragments into Hits and Leads

After identifying the fragments, converting fragments into hits and leads become the final stage. There are three broad converting strategies for converting fragments into hits and leads: fragment optimization, fragment merging or linking and in situ fragment assembly. Fragment optimization is like traditional medicinal chemistry where different substitutions or expansions are made to the fragment to develop the affinity and other properties and it is easy to optimize the smaller fragments. Although these fragments may have low intrinsic affinities, they generally possess binding specificity sufficient to serve as viable anchors for subsequent derivatization [10]. When found, the fragment's potency can be optimized.

Fragment merging or linking includes combining elements from fragments with elements from a known substrate, inhibitor or another fragment by a hybrid molecule. The physicochemical and ADME *absorption, distribution, metabolism and excretion* properties can be improved by this approach. Finally, in situ fragment assembly uses the target as a template for the synthesis of inhibitors from fragments. In practice,

all of these methods have considerable overlap. For instance, *in situ* method could include fragment linking and fragment optimization where fragment linking method may involve the fragment optimization method as well.

In most successful examples of the fragment optimization, the anchoring fragments bound using discrete, specific contacts and their binding modes were preserved throughout the optimization process. [10]. Helping to circumvent undesired qualities is provided by the known inhibitors or substrates by initial fragment for screening and guided optimization. Deconstructing known leads and reassembling are the most successful applications of the fragment based design methods to generate a new chemical series with improved properties.

Using *in situ* fragment assembly is an efficient method for techniques such as NMR or crystallography to identify fragments that can bind to nearby sites in a mutually compatible manner. Combinations of fragments for binding concurrently as potential candidates for linking are provided by the competition studies. However, the linking and merging process of fragments is a technical challenge which is open to development when the target protein has a flexible binding surface. Selection and combination of the fragments *in situ* have now been explored by considering the cross-linking of the proteins. In addition, there are certain criteria to new lead discovery in this method. The molecules must be compatible with the chemical functionality in amino acid side chains. The reactions should be fast enough to enable a reasonable number and the coupling reaction would ideally be under thermodynamic control to prevent highly reactive intermediates from dominating the product distribution [10]. After these criterias are satisfied, then the product would be reasonably compact, soluble and stable.

2.3 WEKA: Data Mining Software

WEKA is a comprehensive toolbench for machine learning and data mining that its main strengths lie in the classification area with a clean, object-oriented Java class hierarchy [32]. Regression, association rules and clustering algorithms are also

included in WEKA software. To start using the software, a set of data items, the dataset is enough. The given dataset is equivalent to a database table or a two dimensional spreadsheet. Instances class in Java is used in WEKA and each instance is equivalent to collection of examples. Each instance consists of a number of attributes, any of which can be nominal (one of a predefined list of values), numeric (a real or integer number) or a string (an arbitrary long list of characters, enclosed in double quotes) [32].

For a basic classifier, a routine that evaluates the generated model on an unseen test dataset or generates a probability distribution for all classes and a routine generates a classifier model from training dataset. The usage of this classifier in terms of mapping or model differs from classifier to classifier. The performance of the classifiers can be determined by various approaches like counting the proportion of correctly predicted examples in an unseen test dataset. The easiest way is to create a training set and a test set by randomly reordering and then splitting into these two sets with collecting all estimates on the test data and calculating average and standard deviation of accuracy.

Cross validation is more detailed method. In cross validation, n fold is specified and dataset is split n folds of equal size after reordering randomly. In each iteration, one fold is used for testing and the other $n-1$ folds are used for training the classifier and the test results are collected and averaged over all folds [32]. Same class distributions in each fold should be obtained in the complete dataset. The cross-validation estimate of accuracy is calculated by this method. On the other hand, it is more useful to deal with small datasets because the small datasets uses the greatest amount of training data from the dataset.

WEKA uses a filter package that filters the classes that transforms datasets and provides useful support for data preprocessing by removing examples, removing or adding attributes and resampling dataset. This package includes both supervised and unsupervised filtering by subdividing into instance and attribute filtering. Supervised filtering takes advantage of the class information. However, unsupervised filtering creates non-stratified subsample of the given dataset and classes should not be assigned

here. After filtering, WEKA contains clusters for finding groups of similar instances in a dataset. Clusters can be visualized and compared to true clusters. In addition, Apriori algorithm for learning association rules and works only discrete data. This algorithm can identify statistical dependencies between groups of attributes. And compute all rules that have a given minimum support and exceed a given confidence.

Attribute selection method is used to investigate the attributes as the most predictive ones. The method includes two parts. The first part is search method that includes the best-first, forward-selection, random, exhaustive, genetic algorithm and ranking attributes. The second part is evaluation method including correlation-based, wrapper, information gain attributes. WEKA is very flexible allowing almost arbitrary combinations of these two methods. In addition, visualization is another useful tool that helps to determine difficulty of the learning problem. Experimenter makes it easy to compare the performance of different learning schemes for classification and regression problems. Finally, the results can be written into file or database. The evaluation options are the cross-validation and drawing learning curve and iterations can be made over different parameter settings.

2.3.1 Prostate and Prostate Cancer

As a glandular tissue, the prostate is a male productive organ which stores and helps to produce seminal fluid. In primal phases, prostate cancer spreads to local structures and nodes in prostate. In the late phases, metastasis mostly occurs to local nodes, bone, supradiaphragmatic lymph nodes and lung [19].

Recently, prostate cancer has been ranked as the 2nd leading cause of cancer deaths among males [33]. 1 out of 6 men will be diagnosed with invasive prostate cancer during their lifetime but it is not likely possible to have before the age 40. Dietary, cultural mediated and genetic differences cause the variations in incidence rate of prostate cancer [34]. African- American population has the highest rate and the lowest rate belongs to Japanese people.

Early diagnosis is very important for prostate cancer treatments [19]. However, most of the patients do not suffer symptoms in the early stages, and if diagnosis is

done after symptoms are seen, metastasis is common [35]. PSA test and DRE are two methods to detect the prostate cancer. PSA means prostate specific antigen which is a glycoprotein produced by prostate. This method is not completely efficient because 20% of the prostate cancer patients have normal prostate specific antigen levels. To increase the efficiency, digital rectal examination (DRE) has done after PSA. DRE shows the irregularities in prostate gland. Although 50% of irregularities are due to cancer, this combined method is the most effective screening tool up-to-date [19].

In treatment process, chemotherapy, radiation therapy, surgical procedures and hormonal therapy are the options. Chemotherapy and radiation therapy use radiation or cytotoxic drugs to kill cancer cells. Therefore, these methods have lots of side effects. Orchiectomy and prostatectomy are the surgical methods for prostate cancer treatment. Orchiectomy is a method that removes the testicles. In addition, prostatectomy is the removal of malignant tissue and all or part of prostate gland. As another method, hormone therapy is the replacement of androgens with anti androgens or antagonists of androgens. Also, controlling androgens by a regulator of androgen synthesis called gonadorellin is another hormonal therapy method.

2.3.2 Drug Design Studies for the Treatment of Prostate Cancer

CYP17 is a well known target for the treatment of prostate cancer since the inhibition of this enzyme exerts control over the androgen synthesis. There are many studies on the computer-generated models on CYP17. Laughton *et al.* [36] built a model for CYP17 and Lin *et al.* had modeled the active site of the protein and defined a bi-lobed substrate binding pocket based on the crystal structure of a class I P450 [37]. The more recent model by Auchus *et al.* is based on a class II P450 [38].

First generation of designed compounds was steroid based molecules like PREG and PROG with various side chains attached to the 17th carbon. These steroidal compounds have some disadvantages like poor acid stability, poor bio-availability, short half life, first pass effects and poor selectivity [39]. Ketoconazole is an anti-fungal 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 ef-

ffects on other cytochrome enzymes [40]. The steroidal compound Abiraterone passed phase II clinical trials and reported to have no dose limiting toxicity [41]. Nnane et al. presented novel steroid-based inhibitors of CYP17. The IC₅₀ values for five steroid-based compounds were determined for CYP17 and 5- 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 [42]. The effect of cinnamic-acid based derivatives of thiazolidinediones on CYP17 was analyzed and have shown inhibition on both reactions catalyzed by CYP17 [43] as a different study. 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 [44]. A novel non-steroidal substrate mimetics reported to showed good inhibition values with good selectivity against CYP3A4 but also showed moderate to high inhibition activity against other hepatic CYP enzymes [45]. The available non-steroidal compounds like progesterone and abiraterone were mainly designed based on mimicking known steroids that are interacting with the CYP 17 active site and creating derivatives of known inhibitors of cytochrome enzymes like pyridine derivatives, xanthone derivatives and carbazole derivatives. The non-steroidal compounds reported in the literature did not yet show promising results in clinical trials. In the work done by Pelin Armutlu, structure-based drug design approach based on the model protein structure by Auchus *et al* [38]. 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 [46].

2.4 Nonlinear Programming Problems

In many cases, especially real world problems, the objective function may not be a linear function, or some of the constraints may not be linear constraints. These kinds of optimization problems involving nonlinearities are called nonlinear programming (NLP) problems. Nonlinear programming models have the same general form as the linear programming models except the objective function with the nonlinear constraints. Generally, the solution procedures are much more complex and no guaranteed procedure exists for all nonlinear programming models.

Solution is often not on the boundary of the feasible solution space unlike linear programming. The solution is not simply on the solution space boundary but other points on the surface of the objective function should be considered as well. These differences greatly complicate solution approaches and make the solution techniques very complex.

2.4.1 Nonlinear Programming Solution Techniques

The Substitution Method

The substitution method is the least complex method for solving nonlinear programming problems. The method includes the constraint equation which is solved for one variable in terms of another and then substituted into the objective function. The constraints are eliminated by substituting the new expression in the objective function and an unconstrained model is formed. This method becomes difficult if the constraint is complex. Therefore, the problems without complex objective functions and with fewer constraints are preferred for this method. Otherwise, Lagrange multipliers method can be used which is not as restricted as the substitution method.

The Method of Lagrange Multipliers

The Lagrange multipliers method is a mathematical technique for solving a problem with nonlinear objective function and one or more linear or nonlinear constraint equations. In this method, constraints as multiples of a multiplier λ are subtracted from

the objective function, which is then differentiated with respect to each variable and solved [47]. This method can encompass problems with more than two variables and is more flexible than the substitution method. However, the mathematics becomes very difficult as the size of the problem increases. The Lagrange multiplier λ in nonlinear programming problems is analogous to the dual variables in a linear programming problems and it reflects the approximate change in the objective function resulting from a unit change in the quantity (right-hand side) value of the constraint equation. A one-unit increase in the right-hand side of the constraint equation results in a λ increase in the objective function which is the same interpretation with the dual variable in linear programming. When λ is positive, the optimal objective function value will increase if the quantity (absolute) value in the constraint equation is increased and will decrease if the quantity (absolute) value is decreased [47]. Otherwise, when λ is negative, the optimal objective function will decrease if the quantity (absolute) value is increased.

2.5 Least Square Regression Method

The least square regression method was independently developed in the late 1700s and the early 1800s by the mathematicians Karl Friedrich Gauss, Adrien Marie Legendre and Robert Adrain working in Germany, France and America respectively. It is the most widely used modeling method to fit a model to a dataset where each explanatory variable is multiplied by an unknown parameter; there is at most one unknown parameter with no corresponding explanatory variable and all of the individual terms are summed to produce the final function value. In statistics, the function that meets these criteria is called linear function. The term linear is used, even though the function may not be a straight line because if the unknown parameters are considered to be variables and the explanatory variables are considered to be known coefficients corresponding to those variables, then the problem becomes a system (usually overdetermined) of linear equations that can be solved for the values of the unknown parameters.

Regression methods are for the estimation of the values for the given parameters. A regression model is required to show how a variable varies in a systematic manner with another variable(s). There are a number of advancements for the linear regression model. The regression model of straight line relationship is

$$y_i = \beta_0 + \beta_1 \cdot x_i + e_i \quad (2.3)$$

where x_i is the independent variable, y_i is the dependent variable values in the i^{th} trial, β_0 and β_1 are the parameters (i.e. intercept and slope) and e_i is the random error term. To find the good estimates for the parameters β_0 , β_1 , the least square regression method is used. In another words, least square method is able to minimize quantity with minimum error rates.

The least square regression method is a primary tool because of its effectiveness and completeness. Many processes in science and engineering are well-described by linear models. The estimates of the unknown parameters obtained from linear least square regression are the optimal estimates from a broad class of possible parameter estimates under the usual assumptions used for process modeling [?]. Therefore, successful results can be obtained by small data and using least square regression is very efficient. In addition, the theory behind the method is very well-understood and it is easy to construct different types of statistical intervals for predictions with clear answers.

The main disadvantages of the least square regression method are the poor extrapolating properties and limitations in the shapes that linear models. Therefore, the linear models may not be effective for extrapolating the results of a process for which data cannot be collected in the region of interest, so extrapolation is potentially dangerous regardless of the model type and it is very sensitive to the presence of unusual data points in the data used to fit the model [?].

Chapter 3

MATERIALS AND METHODS

Computational methods have increasingly been used for drug design in the past two decades due to advances in understanding the molecular behavior of proteins. The main objective of these computational methods is to discover new drugs using the information on the interaction energies of known drugs that minimizes docking and binding energy. Designing new drugs from fragments reduces the combinatorial search for effective drug candidates. Successful results in fragment based drug design have only been obtained with the employment of computational techniques.

In this thesis, this novel method is applied to design drug candidates for the treatment of prostate cancer. Prostate cancer is the most common cancer type among men in the world [48]. One out of six men will be diagnosed with invasive prostate cancer in their lifetime. Suppressing androgen biosynthesis is an important strategy for prostate cancer treatment because 90% of the prostate cancer patients respond to androgen deprivation. Androgens are steroid based molecules and they are responsible for sex-specific organ development and control. They are the major growth factor prostate cancer cells; especially testosterone and dihydrotestosterone. Also, androgen levels are directly proportional to the prostate cancer risk. CYP17 is the enzyme that plays a key role on testosterone and dihydrotestosterone synthesis. Figure 3.1. shows the role of CYP17 enzyme in the androgen synthesis.

If androgen synthesis can be inhibited using Cytochrome P450 (CYP17, PDB access code 2C17) inhibitors or combining the use of these inhibitors with other treatments, it is possible to reduce the side effects of the other treatments (chemotherapy, surgical removal of testicles or prostate and hormonal therapy) [49]. The molecular structure of the CYP17 is given in Figure 3.2.

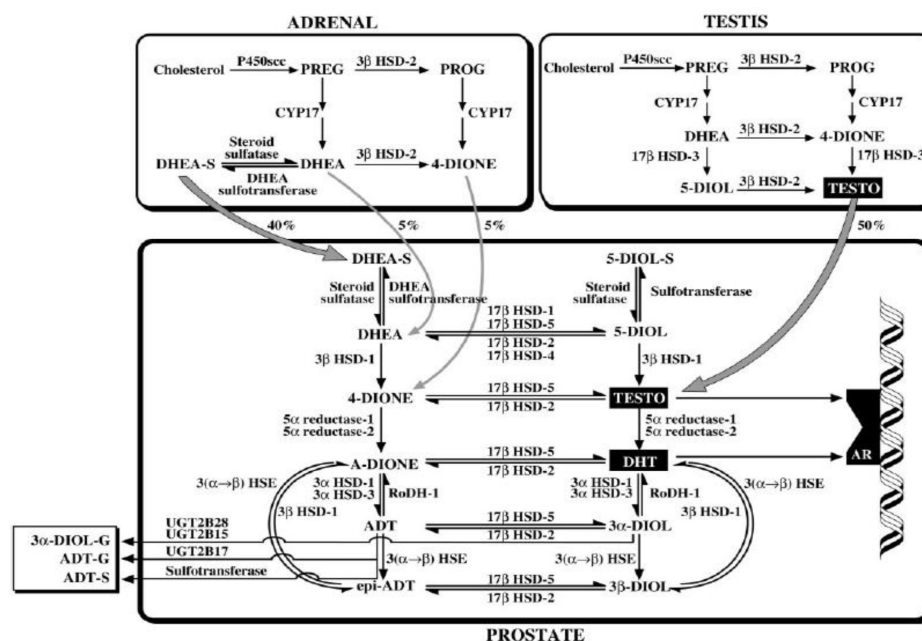


Figure 3.1: Androgen synthesis pathway

Cytochrome P450 is a very large and diverse superfamily of hemoproteins (heme containing), and a member of an even wider superfamily of heme-containing monooxygenases, besides secondary amino monooxygenase SAMO, and heme chloroperoxidase CPO [50]. There are three important characteristics of CYP17 protein binding sites. The first one is the heme region; the catalytic reactions are performed in this region, orient substrates towards the heme. Therefore, by fixing the inhibitor in the catalytic region of the enzyme, the heme iron can be used as an anchor point for blocking the binding site. The second characteristic is the lipophilicity, planarity and rigidity of the steroidal backbone which is successfully maintained in the developed aromatic systems, leading to good inhibitory results [50]. Finally, the third characteristic is the oxygen functionality at the carbon 3 position that provides stronger binding of electron donor and acceptor groups.

The most important androgens related to prostate cancer are testosterone and dihydrotestosterone. The biosynthesis of these androgens takes place in the prostate, testis and adrenal glands. Several enzymes catalyze the reactions starting with chole-

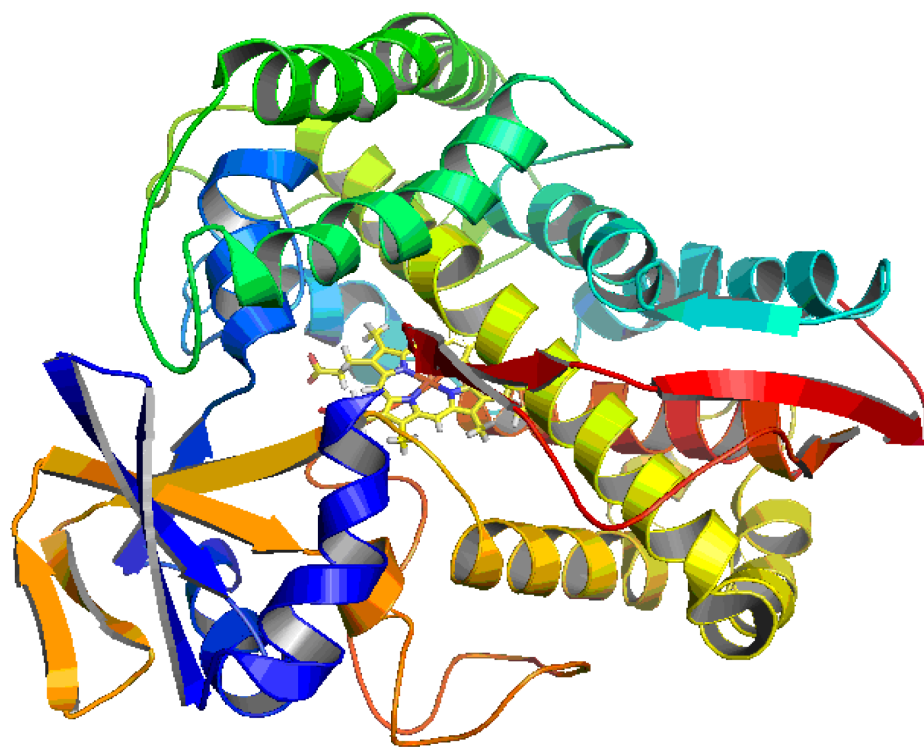


Figure 3.2: Target CYP17 structure

terol 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 indicates 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 [49].

In this work, a computational method that is composed of training and design phases is proposed. The flowchart of the proposed approach is given in Figure 3.3.

First, we use known results which include the discovery of the topological properties that predict the IC50 value, binding energy and docking energy reliably. All the energy values of the combinations are calculated with Autodock tool. From known molecule and known results, a common structure is formed and common fragment

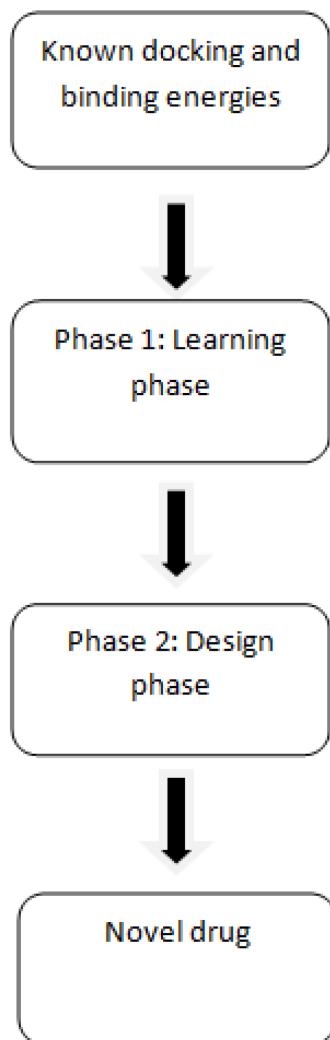


Figure 3.3: Flowchart of the model

were determined. The dataset that includes energy and IC50 values were tested on a computational program called WEKA which gives the best algorithm for data analysis. In our case, least square regression method was chosen. According to least square regression method, weight coefficients for common fragments were determined.

Then, in the design phase, the least square regression model is formed based on the weight coefficients obtained in the first step, to predict the contribution of each group and individual fragment in IC50 value, docking energy and binding energy and

finally the model designs entirely new molecules based on the fragments available by fragment-based drug design. Then a priority order among drug candidates dataset is formed by help of an algorithm with MATLAB.

3.1 Learning Phase

3.1.1 AutoDock

AutoDock is a computational program that provides a method to observe the interaction of a biomolecular target with ligands. Difficulties in the design and development of bioactive compounds for computer aided drug design motivated scientists to create computational tools to analyze the interaction of ligands with the target protein. AutoDock first analyzes the interactions of molecules, like proteins and nucleic acids. Then, this analysis may lead to identification of the promising candidates.

The need for an accurate procedure to evaluate the interaction between two molecules is important. The optimum result is to find the interaction energy between the target protein and the substrate at the minimum with all available degrees of freedom for the system(1). AutoDock uses two techniques: manually-assisted and automated docking. The most used technique in AutoDock is manually-assisted docking where the internal and orientational degrees of freedom can be controlled successfully. The protein is included in a three-dimensional grid in the program. For each type of atom in the substrate, like carbon, oxygen, nitrogen and hydrogen, the affinity grid is calculated. The energy interaction of a single probe atom and protein is assigned to each grid point. For this calculation, the Poisson-Boltzmann finite difference method or using point charge of +1 as the probe procedures are used.

Calculating the energy by a gridding approach is proportional to the number of atoms in the substrate and is independent of the number of atoms in the protein. The energy of a particular substrate is calculated by interpolating the electrostatic potential.

Using the advantage of large search space, the Monte Carlo (MC) simulated annealing technique is used with a grid-based molecular affinity potential approach. In

this simulation, the substrate molecule walks around the protein molecule randomly at each step. Changes in the translation of the substrate center of gravity, orientation and rotation around each flexible internal dihedral angles of substrate position happen for all degrees of freedom stepwise. Each step determines a new configuration with new energy. If the new energy is smaller than the energy calculated before, the new configuration is accepted [1]. If not, the new configuration is accepted or cancelled depending on the probability of acceptance given below.

$$P(\Delta E) = \exp^{-\frac{\Delta E}{k_B T}} \quad (3.1)$$

Here T is the user defined temperature, ΔE is the energy difference between the previous step and the current step and k_B is the Boltzmann constant. As temperature increases, the probability of a conformation being accepted increases.

In each cycle for all individual steps, each specified temperature is used. The accepting or rejecting decision of the conformation is based on the current temperature of that specific step. Then, the next cycle starts with a temperature lowered by a specified schedule as [1]:

$$T_i = gT_{i-1} \quad (3.2)$$

where T_i is the temperature at cycle i and g is a constant between 0 and 1.

Theory

In each binding step, there is a thermodynamic cycle for the binding of an enzyme E and inhibitor I . Taking the hydrophobic effect as a basis, there is an entropic effect for each cycle. Hydrogen bonds between E and I determines the enthalpic stabilization and estimates the energy function. According to Hess's law of heat, the free energy of binding can be calculated by the equation below;

$$\Delta G_{binding,solution} = \Delta G_{binding,vocuo} + \Delta G_{solvation}(EI) - \Delta G_{solvation}(E+I) \quad (3.3)$$

From docking simulation by AutoDock $\Delta G_{binding,vacuo}$ can be calculated. $\Delta G_{solvation(EI)}$ and $\Delta G_{solvation(E+I)}$ can be estimated for both separated and complex forms. Also, calculation of the $\Delta G_{binding,solvation}$, the binding of the inhibitor to the enzyme is possible. Thus, the inhibitor constant K_i can be estimated for the inhibitor I .

Linear regression analysis is used for inhibition constant data with this method. The best fit to the observed inhibition constant data is found and there is no need to modify these coefficients in each step. Docking can be extremely fast by using pre-calculated grid maps in AutoGrid. A grid map is formed by the three dimensional lattice of regularly spaced points. Each point stores the potential energy of a probe from all of the atoms in the macromolecule.

Docking can be extremely fast by using pre-calculated grid maps in AutoGrid. A grid map is formed by the three dimensional lattice of regularly spaced points. Each point stores the potential energy of a probe from all of the atoms in the macromolecule.

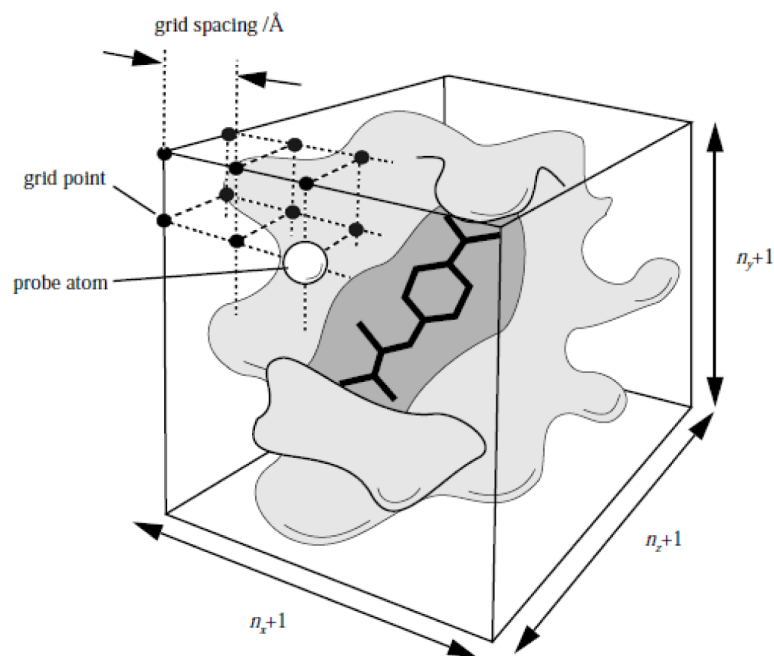


Figure 3.4: The main features of the grid map [1]

Figure 3.4. displays the main features of the grid map. The ligand is buried

inside the active site of the protein in the center of the grid map. The parameters for this grid map are stored in the grid parameter file (GPF). The Van der Waals potential energy is also calculated with the parameters. Van der Waals potential energy is calculated by the pairwise potential energy between two non-bonded atoms as a function of internuclear separation, r . The potential energy equation is

$$V(r) \approx \frac{C_n}{r^n} - \frac{C_m}{r^m} = C_n r^{-n} - C_m r^{-m} \quad (3.4)$$

where m and n are integers, and C_n and C_m are constants whose values depend on the depth of the energy well and the equilibrium separation of the two atoms' nuclei.

The hydrogen bonds are also important in ligand binding. The hydrogens that are bonded to carbon atoms are called polar hydrogens whereas the hydrogens that are bonded to heteroatoms like nitrogen and oxygen are called polar hydrogens. Generally, polar hydrogens of the ligands are used to conserve the disk space. The interactions by hydrogen bonds are calculated by AutoDock.

With these polar hydrogens, AutoDock also needs the electrostatic potential grid map. The electrostatic potential is calculated by;

$$\epsilon(r) = A + \frac{B}{1 + ke^{-\lambda Br}} \quad (3.5)$$

where: $B = \epsilon - A$; ϵ = the dielectric constant of bulk water at 25°C = 78.4; $A = -8.5525$, $\lambda = 0.003627$ and $k = 7.7839$ are parameters.

AutoDock Setup

To start running AutoDock, four input files need to be prepared: PDBQ file for ligand molecule, PDBQS file for macromolecule, GPF file for grid parameters and DPF file for docking parameters.

Docking starts with preparing the macromolecule by adding polar hydrogen atoms. Then, the Kollman charges are added to the macromolecule and the final structure is stored in the PDBQS file format.

After preparing the PDBQS file for the molecule, the second step is to prepare the file for the ligand molecule with same steps. The non polar hydrogens are merged where the polar hydrogens are added. If the ligand is a peptide, the Kollman charges are added, otherwise Gasteiger charges are added. In addition, by calculating the angle between consecutive C atoms, planar and non planar C atoms are marked as well as the torsional freedom of the bonds. The information for the ligand given above are then saved in the PDBQS format as well.

The grid maps that are created to define the active site for the docking are included in the GPF file. Active sites of the protein are defined in a box in AutoDock. In the GPF file, the grid size as number of points, the spacing between two grid points and the grid center coordinates of this box and the numbers of the grid maps are included. The number of the required grid maps depends on the type of atoms that are present in the ligand molecule. Because of that, this study involves the large scale virtual screening of a compound set including hundreds of atoms; grid maps are prepared for each type of atom that may be present in drug like molecules similarly.

The GPF file for virtual screening provides saving from the computational time but lower resolution maps compared to detailed docking. The grid box, including the entire active site, is defined with 0.375 Å spacing in these low resolution maps.

The DPF file contains the population size, number of generations, number of runs, crossover rate, mutation rate and number of evaluations which are the setup for run parameters of the Lamarckian- Genetic Algorithm. In the DPF file the run parameters for virtual screening are defined as follows: the population size is 50, the number of generations is 2.7×10^4 , the crossover rate is 0.8, the mutation rate is 0.02, the number of runs is 10, and the number of evaluations is 1×10^6 .

3.1.2 Decision of Scaffold for Fragment Based Drug Design

According to the work done for deactivating the CYP17, the lead structure is determined as given in Figure 3.5. This scaffold was designed after analyzing the docking position of the lead compound from previous work (LDD). In the work done by Pelin Armutlu, first the molecular dynamics of CYP17 protein was made and the

active sites of the protein were determined. Then, by many docking processes, many lead compound structures were determined that deactivated the active sites of the CYP17 protein. After measuring the inhibitory effect of the drug candidate and MTT toxicity assays of drugs, the numbers of possible structures were eliminated. Finally, with detailed docking processes the lead compound scaffold given in the Figure 3.5. was obtained.

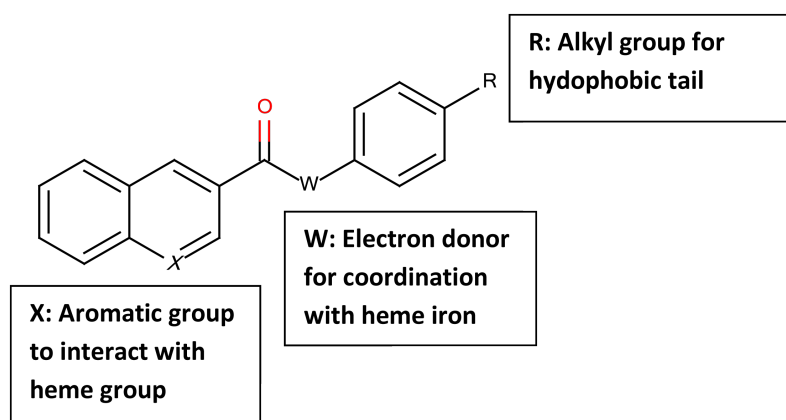


Figure 3.5: The lead compound for prostate cancer treatment

In the lead compound, three different parts that have roles to deactivate the CYP17 protein are present. The first part is the aromatic group that interacts with the heme group of the protein. The *X* position of the aromatic group is suitable for the fragment base drug design. The second part is the position of the electron donor that is shown by *W* in Figure 3.5. This group is for the coordination with heme iron in the CYP17 protein. The final part is shown with *R* in Figure 3.5. which is the alkyl group for the hydrophobic tail of the target.

According to the lead structure determined by previous work, we proposed two different scaffolds. The first scaffold is given in Figure 3.6. and the fragments for positions in the first scaffold are given in Table 3.1.

For the first scaffold, 5 different positions have been selected: R1, R2, G1, G2 and G3. For each position, a specific fragment is determined.

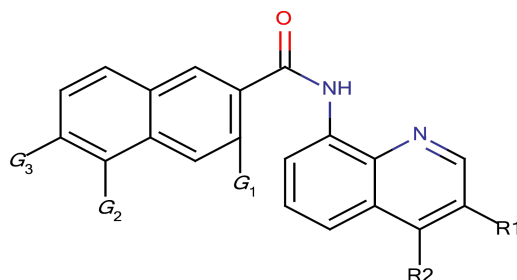


Figure 3.6: First scaffold

R1	R2	G1	G2	G3
		-CH ₃	-CH ₃	-CH ₃
		-OCH ₂	-OCH ₂	-OCH ₂
		-CH ₂ OH	-CH ₂ OH	-CH ₂ OH

Table 3.1: Possible fragments for the each position on the scaffold for the first dataset

The second scaffold is given in Figure 3.7. and the fragments for positions in the second scaffold are given in Table 3.2.

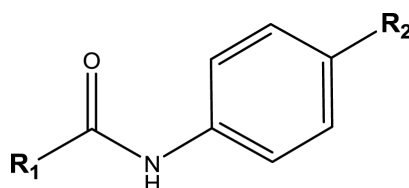


Figure 3.7: Second scaffold

For the second scaffold, 2 different positions have been selected; R1, R2. For each position, a specific fragment is determined.

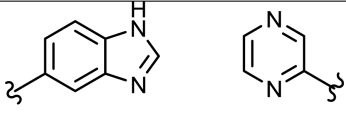
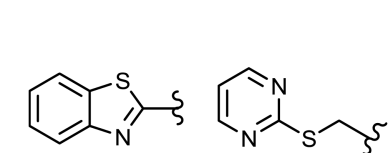
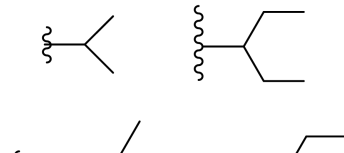
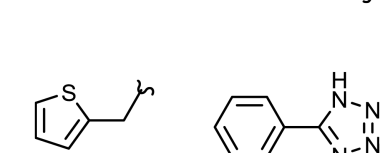
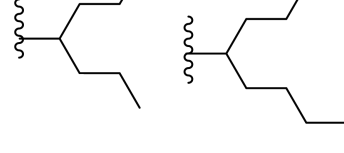
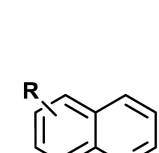
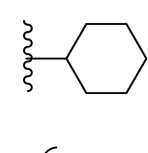
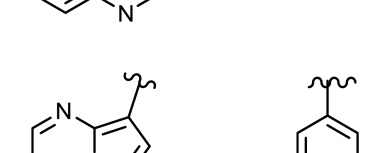
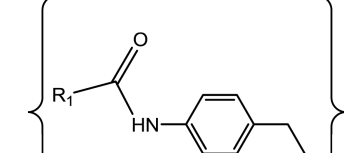
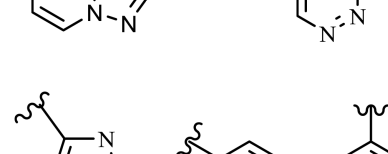
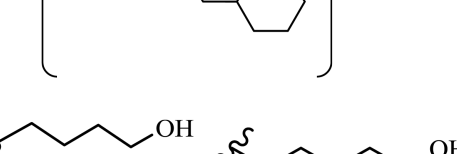
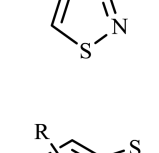
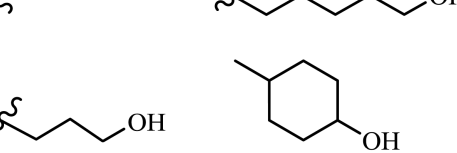
R1:Heme Binding Carboxylic Acids	R2:Alkyl Anilines
	$\zeta-(\text{CH}_2)_n$ $n=1,2,3,4,5,6$
	
	
 <p>R: positios 2,3,4,5,6,7, and 8</p>	
	
	
 <p>R: positios 2,3,4,5,6,7</p>	

Table 3.2: Possible fragments for the each position on the scaffold for the second dataset

To analyze these molecules for docking by AutoDock, all of the compounds were drawn by scientific computational software called MarvinSketch. This program allows saving the structures in PDB format which can be read by AutoDock for docking. Docking and binding energies of all 688 combinations were calculated by AutoDock for both datasets.

3.1.3 The Proposed Design Methodology

We proposed a 2-stage method to design novel chemical compounds to maximize the interaction energy (minimize the docking and binding energy) between the specific target protein CYP17 and the ligand.

Learning Algorithm: Characterization of Data

Mathematical programming is a branch of optimization theory in which a single-valued objective function of n real variable $x_1..x_n$ is minimized(or maximized),possibly subject to a finite number of constraints, which are written as inequalities or equations [51]. When one or more of the functions that appear in mathematical programming are non-linear, the programming becomes non-linear programming. In the complex situations that include decisions in broad sense like in engineering, economics, mathematics, physical sciences have nonlinear programming problems. Nonlinear curve fitting approach is used for the learning part of the model. Optimal values of the parameters in the least square sense are calculated which are the sum of squares of the experimental deviations from the theoretical curve. The least square regression method is a mathematical model that gives the best-fit curve and has the minimal sum of the least square error from a given set of data. To learn the characteristics of our data with known docking and binding energy values, least square regression method was used.

In addition,the least square regression method results the highest R^2 value among all regression methods being searched by the WEKA software. WEKA is software that gives the best characterization, data analysis and algorithm method for the present data set. The model for our data set is given below:

$$\text{Min } z = \sqrt{r_1^2 + r_2^2}$$

s. t.

$$DE_i = \sum_j x_j t(i, j) + r_{1i} \forall i \quad (3.6)$$

$$BE_i = \sum_j x_j t(i, j) + r_{2i} \forall i \quad (3.7)$$

$$x_j, y_j \in R^n \quad (3.8)$$

$$r_{1i}, r_{2i} \in R^n \quad (3.9)$$

$$t(i, j) \in \{0, 1\} \quad (3.10)$$

$$(3.11)$$

where DE is docking energy value, BE is the binding energy value, r_{1i} , r_{2i} are the estimation errors, x_j , y_j are weights and $t(i, j)$ are binary variables. The binary variable table $t(i, j)$ has been designed with the idea of whether the given fragment is bound to the given position or not. If the fragment is bound to the given state, the corresponding value is 1, if not, the corresponding value is zero. The model is created and calculated by GAMS software. According to the r_{1i} , r_{2i} values calculated, the R^2 values are determined. If the dataset has an R^2 value that closes to 1, then the algorithm continues with the design phase. The same algorithm is applied to both first and second datasets.

Design Phase

The logic of the modeling phase for both the first dataset and the second dataset are the same. With the weight coefficients x_j , y_j that are calculated in the learning phase for each dataset, the prediction of the contribution of each group and individual fragment was estimated separately. The main aim in the design phase is to find out the best combination for determined positions for the scaffold determined above with the objective of minimizing energy. Multiobjective optimization approach is used to find the best solution.

Let J be the set of fragments and there are two subsets $Region_1 \subset J$ and $Region_2 \subset J$.

$$Region_1 \cap Region_2 = \emptyset \quad (3.12)$$

$$Region_1 \cup Region_2 = J \quad (3.13)$$

Let i be a combination $i = \{(j, k) | j \in Region_1, k \in Region_2\}$ and $I = \{i | i = (j, k), i \in Region_1, j \in Region_2\}$.

DE is the set of docking energies calculated for each combination using the weight coefficients r_1 and r_2 from the learning part and similarly BE is the set of binding energies.

$$\text{Min } BE_i = mBE$$

$$\text{Min } DE_i = mDE$$

$$i^* = \text{argmin}\{DE_i | BE_i = mBE\}$$

$$m = DE_{i^*}$$

$$\text{Results} = \{i^*\}$$

while $m \geq mDE$ **do**

 Find $tempBE = \min\{BE_i | DE_i < m\}$

 and $k = \text{argmin}\{DE_i | BE_i \leq tempBE\}$

if $k \notin \text{Results}$ **then**

$\text{Results} = \text{Results} \cup K$

end if

$m = m - \varepsilon$

end while

For the first dataset, *region1* represents the region that includes G1, G2 and G3 positions and *region2* represents the region that includes R_1 and R_2 positions on

the scaffold. The algorithm was also designed to choose at most two positions for the *region1* and one position for the *region2*.

For the second dataset, *region1* represents the R_1 position and *region2* represents the R_2 position on the scaffold. The algorithm was also designed to choose at most one fragment for each R_1 and R_2 positions.

Chapter 4

RESULTS AND DISCUSSION

4.1 *First Dataset*

All positions on the common scaffold and all the fragments were determined from the work done by Pelin Armutlu et al. Aromatic group, electron donor for coordination with heme iron and alkyl group for hydrophobic tail were selected according to main structure determined by their work for learning phase. On the aromatic group, three different positions were selected: G1, G2 and G3. Three different groups were chosen to provide interactions between the aromatic group of drug candidate and the heme group of CYP17 active site as well. NH group was determined as electron donor for the coordination with heme iron. For hydrophobic tail of the active site; carbon chains with 3, 4 and 5 carbons were used on R1 and R2 positions. After determination of the fragment and positions, docking and binding energies of all possible 288 combinations were calculated by AutoDock tool.

The analysis of dataset with energy values was done by WEKA which is a collection of visualization tools and algorithms for predictive modeling. According to WEKA, the best method for learning phase was least square regression method. In our learning phase, the objective function is designed to minimize the summation of estimation errors for both docking and binding energies of all molecules. This idea worked for a better regression model with smaller estimation errors. The constraints includes weight for each fragment, their existence parameter and estimation errors. The weight coefficients were multiplied by existence parameter, which is 1 if fragment exists on the given position and 0 otherwise, added with estimation errors and equated with

both docking and binding energies in different equations.

The algorithm was non-linear programming. After calculation of weight coefficients and estimation errors, next step was to calculate the R^2 values of estimation errors. R^2 value was the key point for the accuracy of estimation error calculation. All the estimation error values of docking energies and binding energies were plotted. For docking energy constraint, the R^2 value of estimation error r_1 was 94.53%, which shows enough accuracy.

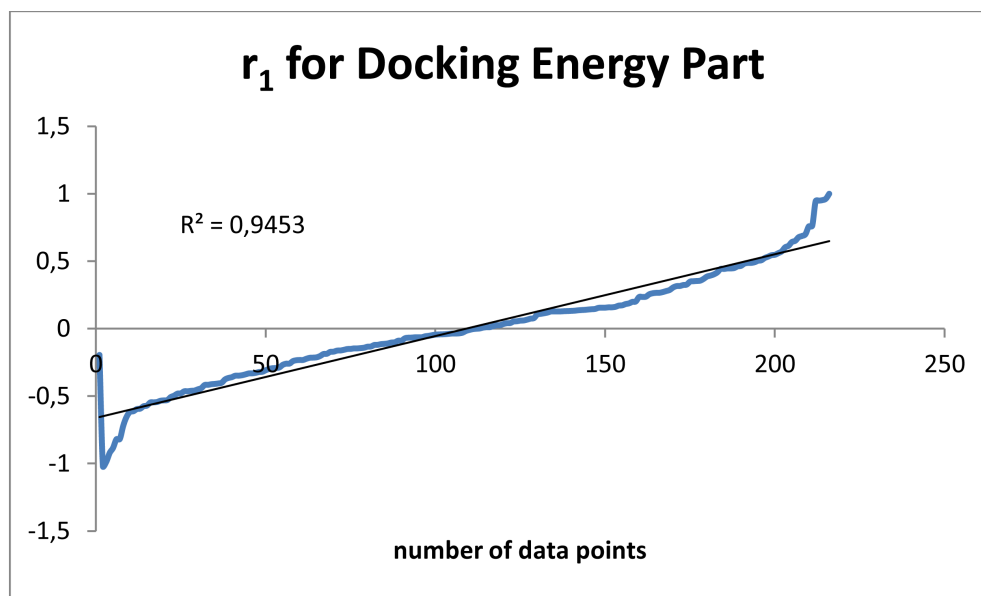


Figure 4.1: . The R^2 value of the estimation error r_1 for the docking energy of the first dataset

For the binding energy constraint of learning phase, the accuracy was 92.92%.

After R^2 value calculation, the design phase was started. In the design phase, the weight coefficients were taken from the learning phase and applied to the new algorithm. In the design phase, the objective was calculated by minimizing the weighted sums of the energy values. In the constraint part, the main idea was to force the algorithm to choose the correct fragment satisfying the objective. Algorithm automatically gives 1 to z_j if the fragment j is chosen for the given position, 0 otherwise. In this phase, the algorithm is the multi-objective integer programming problem.

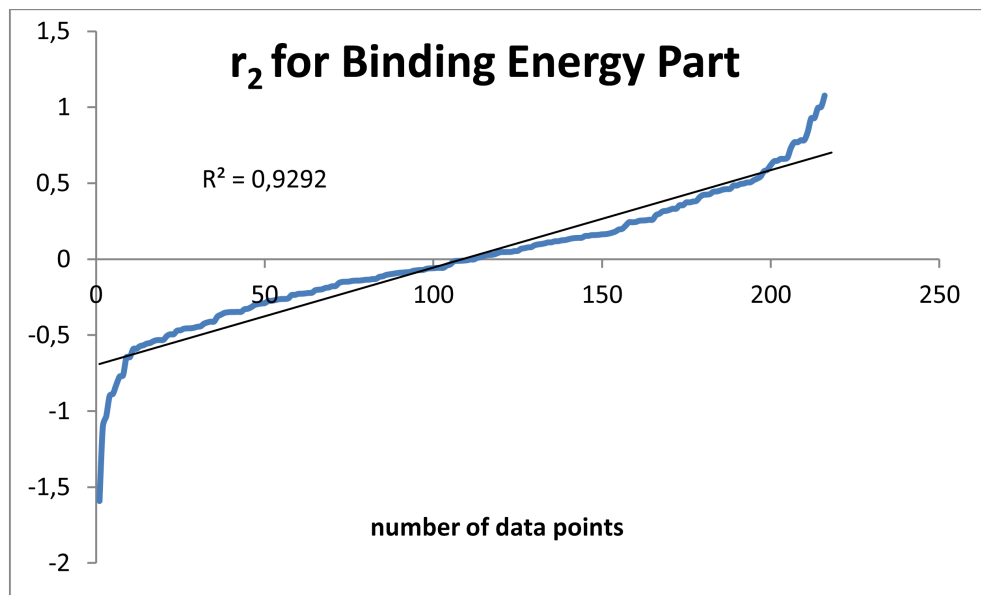


Figure 4.2: The R^2 value of the estimation error r_2 for the binding energy of the first dataset

For the first dataset, the optimal solution after the design phase is given in Figure 4.3.

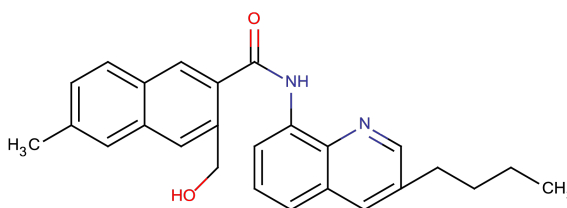


Figure 4.3: Optimal solution for the first dataset

In the algorithm, the constraints were limited by selection of at most two binding groups for G1, G2, G3 positions and only one binding group for R1 and R2 positions. The reason of this restriction is the difficulties in synthesising the molecule that has three fragments in close positions. In the first optimal solution, methyl group for G3 position, hydroxyl methyl group for G1 position and butyl group for R1 position were selected.

One of the main aims of this proposed algorithm was to provide a priority order among a drug candidate dataset. The priority order among first dataset with 288 combinations was obtained by applying an algorithm written in MATLAB. The algorithm cuts the current optimal solution from the current dataset, and then it is possible to get a new optimal solution with same algorithm. Table 4.1 includes the best six combinations among the first dataset.

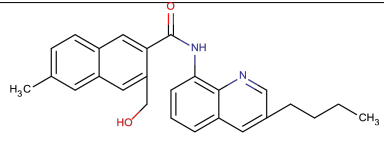
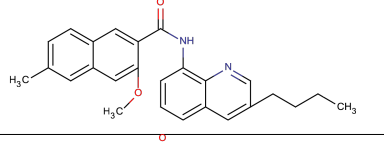
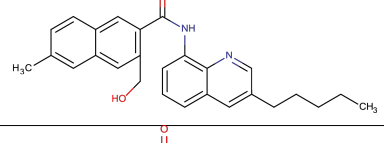
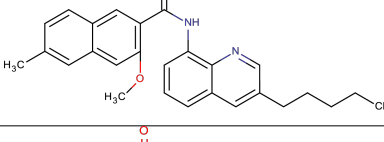
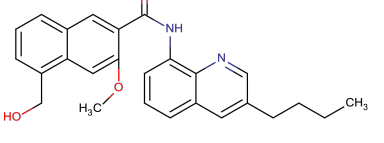
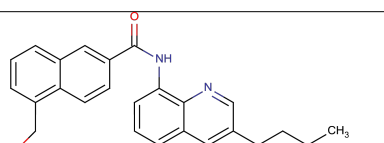
no	Drug structures	Docking energy(AD)	Binding energy (AD)	Docking energy(est.)	Binding energy(est.)
1		-9.06	-8.43	-9.72	-8.46
2		-9.06	-8.42	-9.61	-8.45
3		-9.31	-8.38	-9.43	-8.31
4		-9.31	-8.37	-9.31	-8.30
5		-9.42	-8.23	-9.20	-8.29
6		-9.42	-8.22	-9.16	-8.28

Table 4.1: AutoDock and estimated energies of optimal combinations

The algorithm was designed to choose the pareto optimal solutions which there would a decrease in docking energy without causing a simultaneous increase in binding

energy or vice versa. Pareto optimal algorithms do not always give a single solution and in our case the algorithm gives six optimal solutions. The trend for these possible solutions are given in the Figure 4.4.

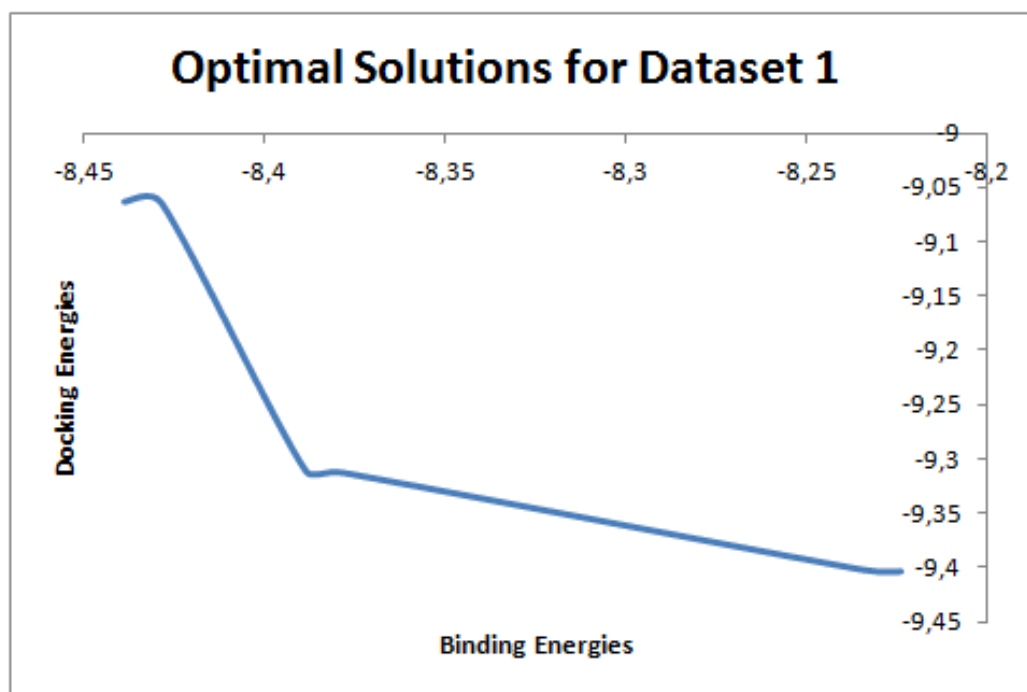


Figure 4.4: Optimal solutions for the first dataset

A decreasing trend going from top to bottom was obtained in terms of estimated docking and binding values. This trend proves that the algorithm provides a valid priority order for given dataset. Approximately 10% error between the Autodock and estimated results was observed for each drug combination.

In the first dataset, the dataset included all of the possible combinations for all positions and groups. Therefore, the training set was formed by deleting 88 combinations from the original dataset (including the results of original dataset). The optimal solution of the training set was given in the Figure 4.5.

The same trend is observed for both of the calculated and estimated energies in the training set. To explain, even the same decreasing trend between the second and the third data point for both calculated and estimated energies are similar. Although our algorithm has approximately 10 % error rate in terms of docking and binding

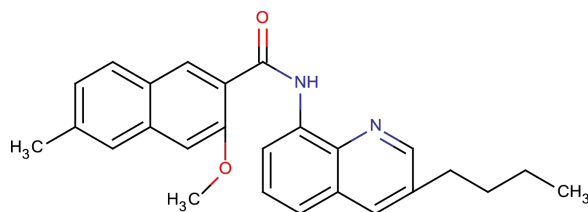


Figure 4.5: The optimal solution of the training set

energies, the overall decreasing trend is obtained. This observation is the proof of the accuracy for our algorithm.

4.2 Second Dataset

The second scaffold was determined from previous work done by Pelin Armutlu et al. as well. Two different positions R1 and R2 were selected. For R1 position, 24 different fragments were determined for the coordination with heme group of the CYP17 active site. For R2 position, 17 different fragments were used for the second scaffold to coordinate with the hydrophobic tail of the target protein. Docking and binding energies of all 446 combinations were calculated by AutoDock tool. According to the docking and binding energy datasets, R2 values for the estimation errors were tested for the second scaffold as well.

According to the R^2 value, the accuracy for the estimation error r_1 was 92.85%.

The accuracy for the estimation error r_2 was 92.96%. After determination of the accurate estimation error values, the design part was started. In the design part of the second dataset, the same design algorithm for the first dataset was used.

The optimal solution of the second dataset is given in the Figure 4.9.

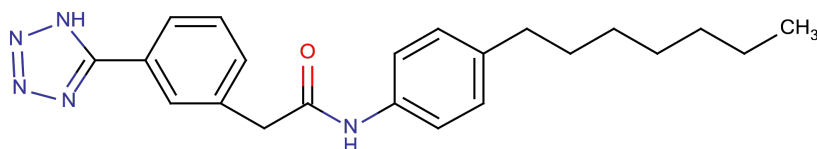


Figure 4.8: Optimal solution for the second dataset

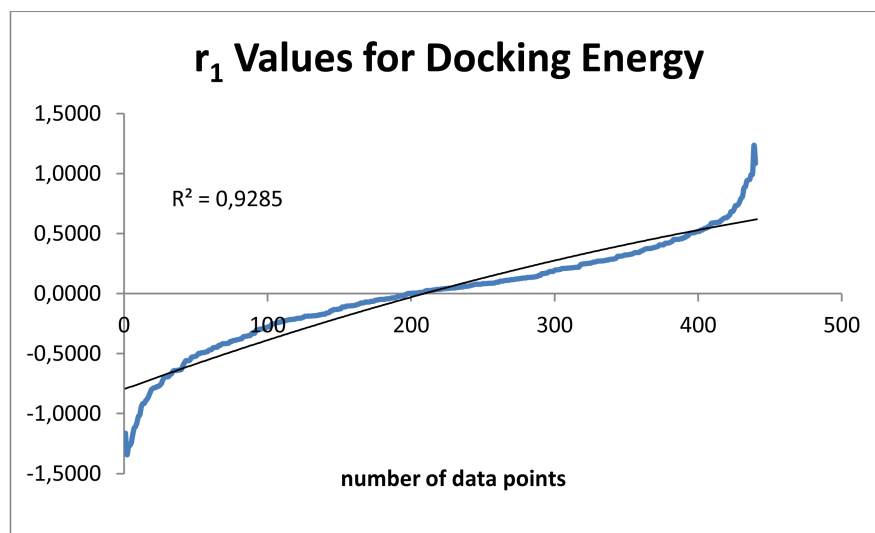


Figure 4.6: The R^2 value of the estimation error r_1 for the docking energy of the second dataset

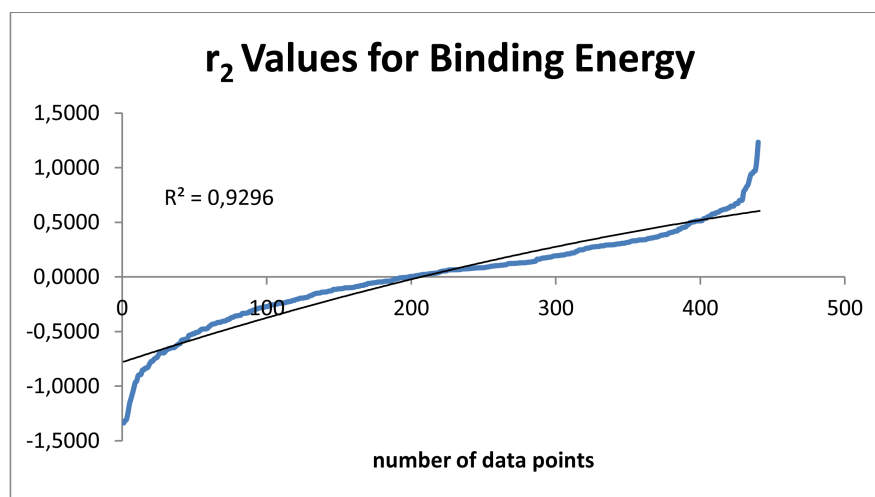


Figure 4.7: The R^2 value of the estimation error r_2 for the binding energy of the second dataset

According to the docking and binding energy values, the most negative energy values belong to the combination with 5-phenyl-1H-1,2,3,4-tetrazole in R1 position and heptylbenzene group in the R2 position for the second scaffold.

To provide a priority order among 408 combinations, an algorithm written in MATLAB is applied. The priority order in the given dataset is formed as given in

the Table 4.2.

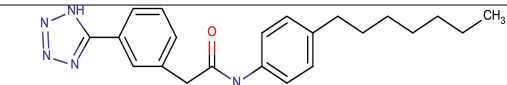
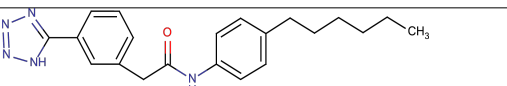
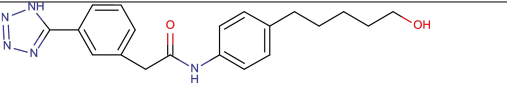
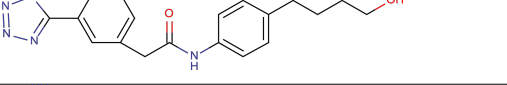
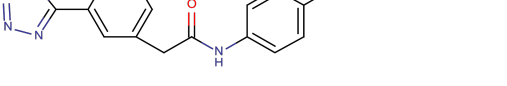
no	Drug structures	Docking en-ergy(AD)	Binding energy (AD)	Docking en-ergy(est.)	Binding en-ergy(est.)
1		-10.72	-7.46	-10.71	-7.56
2		-10.44	-7.50	-10.44	-7.50
3		-10.16	-7.55	-10.40	-7.49
4		-10.15	-7.57	-10.15	-7.47
5		-9.66	-7.82	-10.14	-7.42

Table 4.2: AutoDock and estimated energies of first ten combinations

The algorithm is the same as the first dataset and was designed to choose the pareto optimal solutions which there would a decrease in docking energy without causing a simultaneous increase in binding energy or vice versa. Pareto optimal algorithms do not always give a single solution and in our case the algorithm gives five optimal solutions. The trend for these possible solutions are given in the Figure 4.10.

A decreasing trend was obtained from top to bottom. The error of estimated values between the calculated values of docking and binding energies are about 10%. This small error rate is a proof that our algorithm definitely provides a priority order among given dataset.

A training set is formed for the second dataset as well. The number of combinations was decreased to 350 by deleting 96 combinations randomly. The optimal solution of the training set was the same with the optimal solution of the original set for the second scaffold.

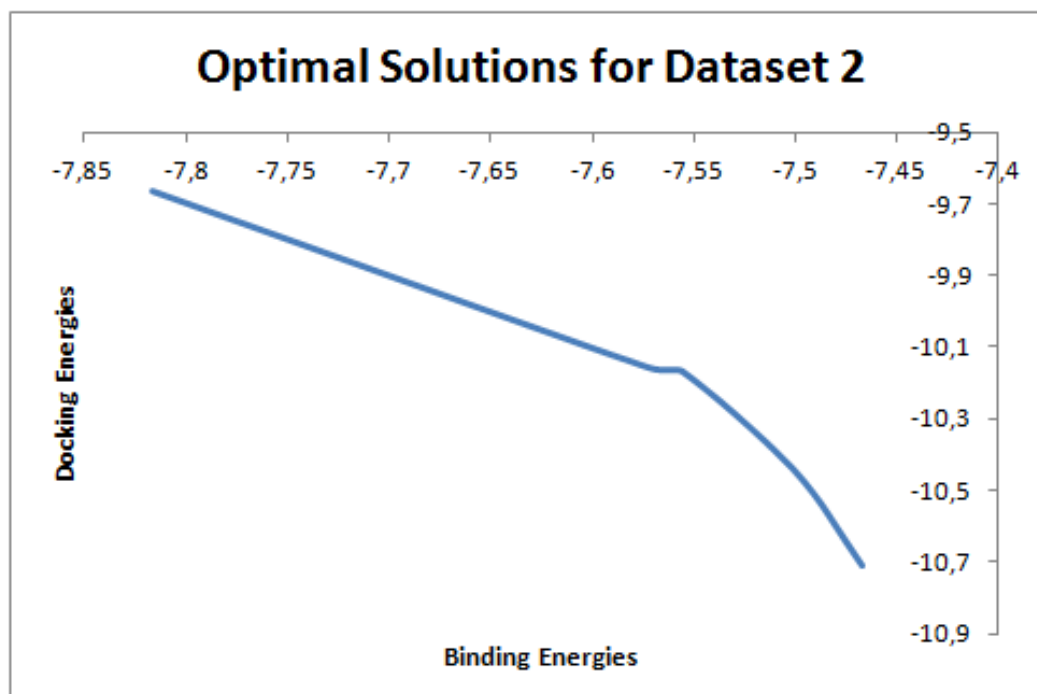


Figure 4.9: AutoDock and estimated energies of optimal combinations

To compare the trend between the docking energy and binding energy, calculated and estimated docking and binding energies are observed in the Figure 4.11.

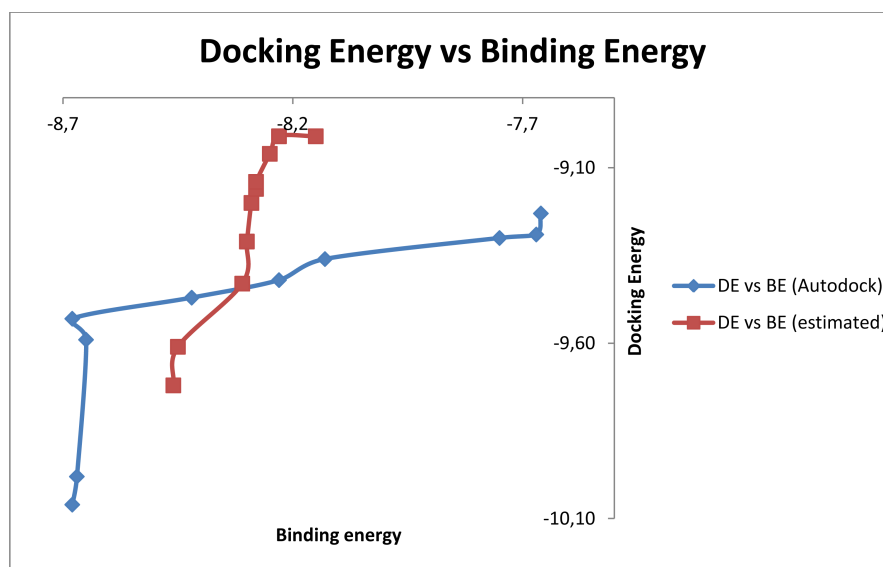


Figure 4.10: The trend between calculated and estimated docking and binding energies for the second dataset

The same trend is observed for both of the calculated and estimated docking and binding energies for the second dataset as well. Even the same decreasing trend between the first and the second data point for both calculated and estimated energies are similar. Although our algorithm has approximately 10 % error rate in terms of docking and binding energies, the overall trend is obtained by our method. This observation is the proof of the accuracy for our algorithm as well.

Chapter 5

CONCLUSION AND FUTURE WORK

This thesis has the motivation of developing a computational method to be used for fragment-based drug design based on the idea of choosing the best fragments on the given scaffold positions and finding out the best combination which has the minimum docking and binding energies. The idea is built on the prerequisite that the scaffold must be previously determined with binding positions of the fragments for the specific target and a fragment library is available.

In this study, the target protein was Cytochrome P450 (CYP17) that is the key enzyme in androgen synthesis and in the treatment of the prostate cancer. The aim was to decrease androgen levels in the cells and therefore prevent the progression of the prostate cancer with deactivating the CYP17 protein by designing drug with fragment-based drug design method. Two different scaffolds were selected from the previous work done [46]. In the first scaffold, there were two different positions on the scaffold and three different fragments for each position. In addition, in the second scaffold, there were two different positions for the fragments. The available fragment library was better so the second prerequisite was successfully implemented to the second scaffold. The number of available fragments was 25 for the first position and 17 for the second position. Docking and binding energies of all possible combinations were calculated by computational tool AutoDock.

Even docking and binding energies of all possible fragment combinations were calculated, it should be emphasized that the synthesis of some of that combinations would be impractical. Therefore, the synthesizability of the compounds may be checked before the dataset formation. If there is any combination that is difficult to synthesize, this combination may be eliminated from the dataset by adding a

constraint to the optimization model.

The main part of the project, i.e. the mathematical programming algorithm that creates energetically minimized structure comprising the energy values of all combinations and weight coefficients of each fragment was attempted. In addition to the energy values, IC₅₀ values can be experimentally found out and to provide a co-work between experimental and computational studies, IC₅₀ values may be added to the objective and constraints. To provide the best fit and have the minimal sum of the least square error from a given set of data, the least square regression model is formed in our model. As an alternative and try to find better fit as a future work, polynomial fit may be used. In our case, there is approximately 10% error rate for the optimal solution and polynomial fit or other curve fitting methods may help decreasing this error rate.

In the design phase, the model chooses the fragment that has the minimum weight coefficient, which means has smaller estimation error for each position for both docking and binding energies. This is a multi objective optimization problem and in the objective there is a minimization model for the both docking and the binding energies and multiple criteria decision making is applied. In the case of experimental IC₅₀ value consideration, a new coefficient should be defined and added to the objective of the design part as well. An algorithm that automatically calculates the error rates of estimated data points for both docking and binding energies can be designed as a future work. Even a visual tool can serve providing an automatic comparison of estimated and calculated data points.

There were two main contributions of this work. The first one was to provide a priority order among given dataset that includes all the possible combinations and make easier the experimental work by decreasing the number of compounds to be tested. To provide priority order, an algorithm in MATLAB was applied to the final optimal solution in the design part. The number of most probable drug compounds in the subset could be determined depending on the size of the original dataset.

The second contribution is to decrease the cost and time that is necessary to approve a drug among a huge dataset. Our method decreases the number of compounds

to be tried experimentally which saves cost and time certainly. As a future work, if the error rate in the optimal values could be decreased, this method would be most efficient in terms of cost and time considerations.

As a result, as this work is one of the first studies on the computational fragment based drug design that helps and shortens the experimental procedure; this thesis could serve as an appropriate base for further studies within this area.

BIBLIOGRAPHY

- [1] Olson, A., Goodsell, D., Morris, G., Huey, R.: Autodock user guide: Automated docking of flexible ligands to receptors, version 2.4. scripps research institute, department of molecular biology (1995)
- [2] Michielan, L., Moro, S.: Pharmaceutical perspectives of nonlinear qsar strategies. *Journal of chemical information and modeling* **50**(6) (2010) 961–978
- [3] Smith, D., Waterbeemd, H.: Pharmacokinetics and metabolism in early drug discovery. *Current opinion in chemical biology* **3**(4) (1999) 373–378
- [4] Clark, D., Pickett, S.: Computational methods for the prediction of drug-likeness. *Drug Discovery Today* **5**(2) (2000) 49–58
- [5] Freire, E.: A thermodynamic guide to affinity optimization of drug candidates. *Proteomics and Protein-Protein Interactions* (2005) 291–307
- [6] Lipinski, C.: Drug-like properties and the causes of poor solubility and poor permeability. *Journal of pharmacological and toxicological methods* **44**(1) (2000) 235–249
- [7] Veber, D., Johnson, S., Cheng, H., Smith, B., Ward, K., Kopple, K.: Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **45**(12) (2002) 2615–2623
- [8] Varnek, A., Tropsha, A.: Chemoinformatics approaches to virtual screening. Royal Society of Chemistry (2008)
- [9] Gardel, M., Valentine, M., Crocker, J., Bausch, A., Weitz, D.: Microrheology of entangled f-actin solutions. *Physical review letters* **91**(15) (2003) 158302

-
- [10] Erlanson, D., McDowell, R., O'Brien, T., et al.: Fragment-based drug discovery. *Journal of medicinal chemistry* **47**(14) (2004) 3463–3482
- [11] Pozio, E.: Epidemiology and control prospects of foodborne parasitic zoonoses in the european union. *Parassitologia* **50**(1/2) (2008) 17
- [12] Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L., Pähler, A.: Computational toxicology in drug development. *Drug discovery today* **13**(7-8) (2008) 303–310
- [13] Papa, E., Villa, F., Gramatica, P.: Statistically validated qsars, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in pimephales p romelas (fathead minnow). *Journal of chemical information and modeling* **45**(5) (2005) 1256–1266
- [14] Castillo-Garit, J., Marrero-Ponce, Y., Escobar, J., Torrens, F., Rotondo, R.: A novel approach to predict aquatic toxicity from molecular structure. *Chemosphere* **73**(3) (2008) 415–427
- [15] Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Oberg, T., Dao, P., Cherkasov, A., Tetko, I.: Combinatorial qsar modeling of chemical toxicants tested against tetrahymena pyriformis. *Journal of chemical information and modeling* **48**(4) (2008) 766–784
- [16] Jorissen, R., Gilson, M.: Virtual screening of molecular databases using a support vector machine. *Journal of chemical information and modeling* **45**(3) (2005) 549–561
- [17] Pal, M.: Multinomial logistic regression-based feature selection for hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation* **14**(1) (2012) 214–220

- [18] Norinder, U.: Support vector machine models in drug design: applications to drug transport processes and qsar using simplex optimisations and variable selection. *Neurocomputing* **55**(1) (2003) 337–346
- [19] Frydenberg, M., Wijesinha, S.: Diagnosing prostate cancer: What gps need to know. *Australian family physician* **36**(5) (2007) 345–347
- [20] Jorgensen, W.: The many roles of computation in drug discovery. *Science's STKE* **303**(5665) (2004) 1813
- [21] Fejzo, J., Lepre, C., Peng, J., Bemis, G., Murcko, M., Moore, J., et al.: The shapes strategy: an nmr-based approach for lead generation in drug discovery. *Chemistry & biology* **6**(10) (1999) 755–769
- [22] Lewell, X., Judd, D., Watson, S., Hann, M.: Recap retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of chemical information and computer sciences* **38**(3) (1998) 511–522
- [23] Hajduk, P., Bures, M., Praestgaard, J., Fesik, S.: Privileged molecules for protein binding identified from nmr-based screening. *Journal of medicinal chemistry* **43**(18) (2000) 3443–3447
- [24] Jacoby, E., Davies, J., Blommers, M.: Design of small molecule libraries for nmr screening and other applications in drug discovery. *Current Topics in Medicinal Chemistry* **3**(1) (2003) 11–23
- [25] Lajiness, M., Vieth, M., Erickson, J., et al.: Molecular properties that influence oral drug-like behavior. *CURRENT OPINION IN DRUG DISCOVERY AND DEVELOPMENT* **7**(4) (2004) 470–477
- [26] Schneider, G., Böhm, H.: Virtual screening and fast automated docking methods. *Drug Discovery Today* **7**(1) (2002) 64–70

- [27] Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A., Cambillau, C.: Atomic resolution (1.0 Å) crystal structure of fusarium solani cutinase: stereochemical analysis. *Journal of molecular biology* **268**(4) (1997) 779–799
- [28] Eisen, M., Wiley, D., Karplus, M., Hubbard, R.: Hook: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins: Structure, Function, and Bioinformatics* **19**(3) (1994) 199–221
- [29] Wang, R., Gao, Y., Lai, L.: Ligbuilder: a multi-purpose program for structure-based drug design. *Journal of molecular modeling* **6**(7) (2000) 498–516
- [30] Levitt, D., Banaszak, L.: Pocket: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics* **10**(4) (1992) 229–234
- [31] Pearlman, D., Murcko, M.: Concerts: dynamic connection of fragments as an approach to de novo ligand design. *Journal of medicinal chemistry* **39**(8) (1996) 1651–1663
- [32] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter* **11**(1) (2009) 10–18
- [33] Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Thun, M.: Cancer statistics, 2007. *CA: a cancer journal for clinicians* **57**(1) (2007) 43–66
- [34] Latil, A., Azzouzi, R., Cancel, G., Guillaume, E., Cochran-Priollet, B., Berthon, P., Cussenot, O.: Prostate carcinoma risk and allelic variants of genes involved in androgen biosynthesis and metabolism pathways. *Cancer* **92**(5) (2001) 1130–1137

- [35] Thompson, I., Ankerst, D.: Prostate-specific antigen in the early detection of prostate cancer. *Canadian Medical Association Journal* **176**(13) (2007) 1853–1858
- [36] Laughton, C., Neidle, S., Zvelebil, M., Sternberg, M.: A molecular model for the enzyme cytochrome p450; sub ζ 17 α /sub ζ , a major target for the chemotherapy of prostatic cancer. *Biochemical and biophysical research communications* **171**(3) (1990) 1160–1167
- [37] Geller, D.H., Auchus, R.J., Mendonça, B.B., Miller, W.L.: The genetic and functional basis of isolated 17, 20-lyase deficiency. *Nature genetics* **17**(2) (1997) 201–205
- [38] Qiao, J., Hu, R.M., Peng, Y.D., Song, H.D., Peng, Y.W., Gao, G.F., Hao, J.H., Hu, N.Y., Xu, M.Y., Chen, J.L.: A complex heterozygous mutation of his373leu and asp487-ser488-phe489 deletion in human cytochrome p450c17 causes 17 α -hydroxylase/17, 20-lyase deficiency in three chinese sisters. *Molecular and cellular endocrinology* **201**(1) (2003) 189–195
- [39] Stigliano, A., Gandini, O., Cerquetti, L., Gazzaniga, P., Misiti, S., Monti, S., Gradilone, A., Falasca, P., Poggi, M., Brunetti, E., et al.: Increased metastatic lymph node 64 and cyp17 expression are associated with high stage prostate cancer. *Journal of endocrinology* **194**(1) (2007) 55–61
- [40] Clement, O.O., Freeman, C.M., Hartmann, R.W., Handratta, V.D., Vasaitis, T.S., Brodie, A.M., Njar, V.C.: Three dimensional pharmacophore modeling of human cyp17 inhibitors. potential agents for prostate cancer therapy. *Journal of medicinal chemistry* **46**(12) (2003) 2345–2351
- [41] Attard, G., Yap, T., Reid, A., Parker, C., Barrett, M., Raynaud, F., Dowsett, M., Dearnaley, D., Lee, G., De Bono, J.: Phase i study of continuous oral dosing of an irreversible cyp17 inhibitor, abiraterone, in castration refractory prostate cancer

- (crpc) patients incorporating the evaluation of androgens and steroid metabolites in plasma and tumor. *J Clin Oncol* **25** (2007) 5063
- [42] Nnane, I.P., Kato, K., Liu, Y., Long, B.J., Lu, Q., Wang, X., Ling, Y.z., Brodie, A.: Inhibition of androgen synthesis in human testicular and prostatic microsomes and in male rats by novel steroidal compounds. *Endocrinology* **140**(6) (1999) 2891–2897
- [43] Arlt, W., Neogi, P., Gross, C., Miller, W.L.: Cinnamic acid based thiazolidinediones inhibit human p450c17 and 3beta-hydroxysteroid dehydrogenase and improve insulin sensitivity independent of ppargamma agonist activity. *Journal of molecular endocrinology* **32**(2) (2004) 425–436
- [44] Handratta, V.D., Vasaitis, T.S., Njar, V.C., Gediya, L.K., Kataria, R., Chopra, P., Newman, D., Farquhar, R., Guo, Z., Qiu, Y., 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* **48**(8) (2005) 2972–2984
- [45] Pinto-Bazurco Mendieta, M.A., Negri, M., Jagusch, C., Hille, U.E., Müller-Vieira, U., Schmidt, D., Hansen, K., Hartmann, R.W.: Synthesis, biological evaluation and molecular modelling studies of novel acd-and abd-ring steroidomimetics as inhibitors of cyp17. *Bioorganic & medicinal chemistry letters* **18**(1) (2008) 267–273
- [46] Armutlu, P., Ozdemir, M., Ozdas, S., Kavakli, I., Turkay, M.: Discovery of novel cyp17 inhibitors for the treatment of prostate cancer with structure-based drug design. *Letters in Drug Design & Discovery* **6**(5) (2009) 337–344
- [47] Shetty, M.: *Nonlinear programming*. Wiley Online Library (1993)
- [48] Mettlin, C.: Recent developments in the epidemiology of prostate cancer. *European Journal of Cancer* **33**(3) (1997) 340–347

- [49] Haidar, S., Hartmann, R.: 5.7 enzyme inhibitor examples for the treatment of prostate tumor. *Enzymes and their inhibition: drug development* **44** (2005) 241
- [50] Pinto-Bazurco Mendieta, M.: Heme-iron complexing biphenyl and naphthalene derivatives as CYP17 inhibitors for the treatment of prostate cancer: design, synthesis and evaluation. PhD thesis, Universitätsbibliothek (2009)
- [51] Avriel, M.: *Nonlinear programming: analysis and methods*. Dover Pubns (2003)