3D ANALYSIS OF THE BINDING SITES FOR PREDICTING BINDING AFFINITIES IN DRUG DESIGN

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ALI OSMAN ATAÇ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2014

Approval of the thesis:

3D ANALYSIS OF THE BINDING SITES FOR PREDICTING BINDING AFFINITIES IN DRUG DESIGN

submitted by **ALI OSMAN ATAÇ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Prof. Dr. Ferda Nur Alpaslan _____
Supervisor, **Computer Engineering Department, METU**

Prof. Dr. Erdem Büyükbingöl _____
Co-supervisor, **Department of Pharmacy., Ankara Uni.**

Examining Committee Members:

Prof. Dr. Sibel Süzen
Department of Pharmacy, Ankara Uni. _____

Prof. Dr. Ferda Nur Alpaslan
Computer Engineering Department, METU _____

Prof. Dr. Erdem Büyükbingöl
Department of Pharmacy, Ankara Uni. _____

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering Department, METU _____

Dr. Özlem Erdaş
Republic of Turkey Prime Ministry _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ALI OSMAN ATAÇ

Signature :

ABSTRACT

3D ANALYSIS OF THE BINDING SITES FOR PREDICTING BINDING AFFINITIES IN DRUG DESIGN

Ataç, Ali Osman

M.S., Department of Computer Engineering

Supervisor : Prof. Dr. Ferda Nur Alpaslan

Co-Supervisor : Prof. Dr. Erdem Büyükbingöl

September 2014, 52 pages

Understanding the interaction between drug molecules and proteins is one of the main challenges in drug design. Several tools have been developed recently to decrease the complexity of the process. Artificial intelligence and machine learning methods have promising results in predicting the affinities. Recently, accurate estimations have been performed by extracting the electrostatic potentials from images of the drug-protein binding sites which were generated by autodocking simulator. In this study, a new algorithm has been implemented, which is a modified version of CIFAP, to predict binding affinities of CheckPoint Kinase1 and Caspase3 inhibitors.

Keywords: Machine Learning, Protein - Drug Interaction, Affinity Prediction, Binding Site Analysis, 3D Data Model

ÖZ

İLAÇ DİZAYNİNDA AFİNİTELERİ TAHMİN ETMEK İÇİN BAĞLANMA ALANLARININ ÜÇ BOYUTLU ANALİZİ

Ataç, Ali Osman
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Ferda Nur Alpaslan
Ortak Tez Yöneticisi : Prof. Dr. Erdem Büyükbingöl

Eylül 2014 , 52 sayfa

İlaç ve protein molekülleri arasındaki etkileşimi anlamak, ilaç dizaynının en önemli zorluklarındandır. Son yıllarda ilaç tasarımını kolaylaştıran birçok metod ortaya konulmuştur. Yapay zeka ve makina öğrenimine dayalı yöntemler ile umut verici sonuçlar elde edilmektedir. Geçmiş etkileşimleri kullanarak, gerçeğe yakın tahminler yapılabilmektedir. Bu çalışmada da bir oto-kenetlenme simulatörü aracılığıyla elde edilen ilaç-protein etkileşim alanı görüntülerinden elektrostatic özellikler çıkarılmıştır. CIFAP (Compressed Images For Affinity Prediction) adı verilen veri modelleme tekniğinden esinlenilerek, yeni dinamik bir yöntemle modelleme yapılmıştır. Daha önce kullanılan 2 Boyutlu model, 3 Boyutlu dikdörtgenler prizmalarına dönüştürülmüştür. Daha sonra bu veri modeli, birçok makina öğrenimi tekniği ile birlikte kullanılarak, CHK1 inhibitörlerinin afinitesini tahmin etmekte kullanılmıştır.

Anahtar Kelimeler: Makina Öğrenimi, Protein - İlaç Etkileşimi, Bağlanma Yöresi Analizi, 3 Boyutlu Veri Modeli

To my family and people who are reading this page

ACKNOWLEDGMENTS

I wish to thank, first and foremost, my supervisor Professor Ferda Nur Alpaslan for her constant support, guidance and friendship. This thesis would not have been possible without the helps of my co-supervisor Erdem Büyükbingöl. He gave me the passion to go further in the study and always help with pharmaceutical content that I had to deal with first time in my life. It was a great honor to work with them for my master thesis work.

There are many more people who have supported me throughout this work. First of all, I would like to thank Doctor Özlem Erdaş who helped me understanding the basic idea of the CIFAP study and always give feedback to my ideas.

I am also grateful to Nisa Olgun who has given all my motivation to complete this study. She has observed my progress throughout this work and push me to my limits. I would like also to thank my friends and colleagues Göksel Uçtu and Onur Demir. We are a perfect team together against all the difficulties of the life.

It is with immense gratitude that I acknowledge the scholarship (2211) of Scientific and Technological Research Council of Turkey for my master study .

Lastly, sincerest thanks to each of my family members for supporting and believing in me all the way through my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

CIFAP	Compressed Images For Affinity Prediction
MIRS	Multiple Instance Regression Scoring
SVR	Support Vector Regression
BFS	Best First Search
PLSR	Partial Least Square Regression
LOOCV	Leave-one-out Cross Validation
RMSE	Root Mean Square Error
MLR	Multiple Linear Regression
IC50	Half maximal inhibitory concentration
CHK1	Checkpoint Kinase 1
CASP3	Caspase 3

CHAPTER 1

INTRODUCTION

1.1 Motivation

Throughout history, mankind has struggled to cure diseases. In the historical approach, studies were being conducted in laboratories and trial-and-error testing of drug candidates on animals [2]. However, today, a modern approach called "in silico drug discovery" placed in the literature. With the help of high performance computers, drug experiments could be performed in simulation environments. Although, this has eventually reduced the cost of the process, drug design is still a challenging task with lots of failures. A modern drug discovery pipeline includes several steps.

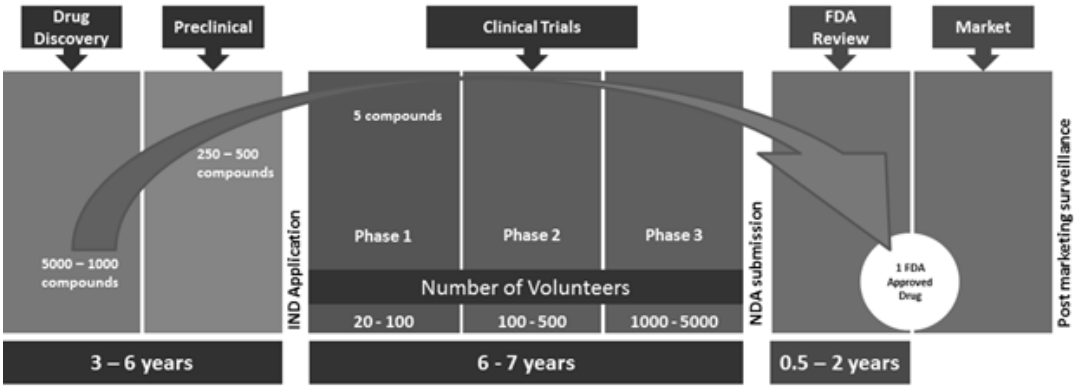


Figure 1.1: Drug discovery pipeline. The process starts with early drug discovery phase and the drug is placed on the market after FDA approval

In Figure - 1.1, a drug design process, from early stages to FDA (Food Drug and Administration) approval is demonstrated. It takes approximately 10-15 years and 1.2 billion $ for a new drug to place in the market. In the early target identification phase, 5000-10000 candidate compounds are tested, while only one of them probably could be approved.

It is quite important to reduce the time and cost of this process. As a result, many molecular docking software have been arisen, which help to understand how a candidate ligand interact with a target protein. In the last decade, more successful approaches using machine learning techniques have become popular in the drug design. Especially pharmaceutical companies extensively use machine learning algorithms to model, analyze, and predict biological results of a candidate drug in discovery process[3].

1.2 Background Information

Protein: Proteins are one of the most important molecules of the organisms. They have roles in every biological process by becoming a part of enzymes and hormones. A protein is a molecule which is composed of more than 50 amino acids. There are 20 types of amino acids. The types, sequence and length of amino acids changes the type of the protein. An amino acid molecule contains a central carbon (alfa carbon) which 4 other groups are attached to: Amine group (-NH2), carboxylic acid group(-COOH), hydrogen group(-H) and a side chain (-R). The side chain group is unique to each amino acid types. In a protein molecule, amino acids are combined with peptide bond[4] In this work, “target” and “receptor” are both used to refer a protein molecule.

Ligand: Ligand is a molecule of any size that binds to a protein to accomplish a biological activity. Changing the 3D shape and other chemical properties, a ligand can activate or stop an enzyme activity when it binds to the target protein.

X-ray crystallography: X-ray crystallography is a method to identify structure of molecule. In this method, first a purified molecule crystallized. Then beam of X-rays

are diffracted by the atoms in the molecule. By measuring the diffraction (angles and intensities), the positions of atoms and bonds are determined. Finally a 3D model of the molecule is generated[5].

Docking: Without having an X-ray crystallography of a protein – ligand compound, one may want to know the best interaction position of the complex. Docking is the method which predicts the best possible position of a ligand interacted with a protein. This helps estimating binding affinity by calculating the energy of the complex [6]

Binding Affinity: A ligand binds to its target by constructing some weak bonds such as hydrogen bonds, electrostatic attractions, Van der Waals bonds and hydrophobic forces. The combination of these bonds determines how strong the compound is, in other words the "binding affinity"[7].

Binding Site: The interaction between ligand and the receptor is like a "lock" and "key" relation. The surface of the target should fit the ligand. The part of the protein that ligand binds to is called the "binding site"[8].

1.3 Problem Definition

Estimation of binding affinity of an unknown complex is a hard challenge in drug design process. The scoring functions which are highly trusted in docking process is not very effective on calculating the actual binding affinity[9]. Therefore some new algorithms are needed to effectively predict unknown binding affinity.
Electrostatic potential values are one of the important descriptors which give information about binding affinities. Furthermore, binding sites of complexes are proven to be more informative than other areas of the complexes[1]. Even with binding site analysis, there is a large set of points that should be considered for electrostatic potential values. Thus, there is a need for a data model that is both efficient for time complexity and low on data-loss. In this study, we try to find a new method that composed of a data model, feature selection and prediction technique that predicts binding affinities of Checkpoint Kinase 1[10] and Caspace3 inhibitors[11] effectively, by modifying

the basic idea defined in the study called CIFAP[1]

1.4 Contribution

The main contribution intended by this study is to provide a 3D analysis of the binding site of the protein – ligand complexes by improving the idea proposed in the CIFAP study. Since a 3D analysis requires a more sophisticated data model, some different approaches to find an efficient data model have been implemented. Furthermore, problem of low number of instances and high number of features have been overcome by establishing a problem specific feature selection technique. This feature selection method is implemented with 2 different heuristic search methods that mainly aim to reduce number of possible feature subset combinations. It has been also shown that using a 3D model could have greater results than 2D analysis. The search and feature selection techniques used in the study could lead to better algorithms and results in further studies.

1.5 Organization of the Thesis

The thesis starts with the previous studies that try to estimate the binding affinity with different methods. The base study CIFAP[1] also summarized in this chapter. In the third chapter, data preparation techniques and algorithms are explained. This part contains ligand preparation, docking and obtaining grid cubes processes. Furthermore, the algorithms used in the feature selection and prediction phases were also discussed in this chapter. Both the dynamic 3D model and the prediction progresses are stated in the fourth chapter. At the very beginning of the chapter, a brief explanation about the idea is given and some of the challenges are defined. In next sections, different approaches to overcome these challenges are given. There are also 2 different heuristic functions to search the feature subset tree in this chapter. In the fifth chapter the experimental results of CHK1 and CASP3 complexes are represented. For each complex, firstly, the role of the target protein is explained. Then, the data preparation process is explained. Finally, the results are given with figures, tables and their explanations. By the way, the results for level based method and individual sub-cube

method are given in two subsections separately.

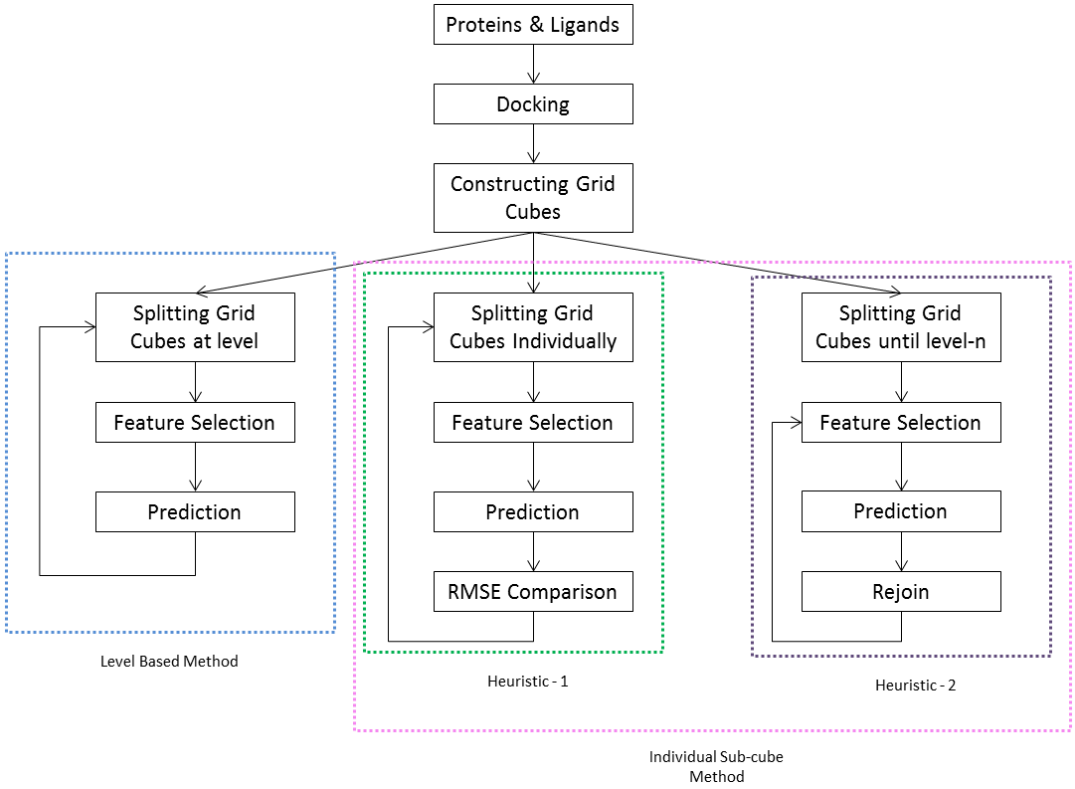


Figure 1.2: Flow chart for propsed methods

CHAPTER 2

RELATED WORKS

2.1 Multiple Instance Regression Scoring

Molecular Dynamics and Monte Carlo are the key methods to predict binding affinities of unknown protein - ligand complexes. However these methods have a high computational cost. Therefore various scoring functions have been arisen recently. These functions use force-field, knowledge based and empirical scoring functions. Recent studies have revealed that it is needed to calculate binding affinities in a different way, because these scoring functions alone do not predict binding affinities correctly enough. Teramoto and Kashima[12] come up with the idea of using different scoring functions and binding poses together with the previously known binding affinities of interactions. They created a new scoring function from the existing ones by using Multiple Instance Regression based scoring.

In this method, several scoring functions are used to score different docking poses of the complex. And these values are grouped together as "bag"s. Each bag has a binding affinity data which is obtained experimentally. Multiple Instance Regression is used with bags and binding affinity to train the model.

Multiple Instance Learning is supervised learning method, in which there are multiple instances for a specific label. One or more instances may contribute to the binding affinity, in other words, the observed classification may be effected by only one of the instances inside a bag[13].

In the experiments, 100 different protein-ligand complexes are used. Then, 100 different docking poses for each complex is calculated using AutoDock. In the next step, all of this poses are scored using 11 different scoring functions. Finally, Each bag(containing 1100 different score value) and its corresponding binding affinity is used for training. This new tuned scoring function (MIRS[12]) outperformed the conventional scoring functions in predicting the binding affinity

2.2 B2Bscore

In a recent work studied by Lui et al., they have proposed a new scoring function called B2BScore[14]. The main idea in this approach was to use B factor and B contacts which are two important physicochemical properties of ligand-protein complexes. B factor measures the changes of the atom positions in a protein and gives information about protein dynamics. B contacts are atom pairs which there are no other atoms between. B contacts involve very few unimportant contacts with the help of this property. In B2Bscore, B Factor values are integrated into B contacts in a vector representation which is obtained from protein – ligand interfaces.

Using different types of contacts vectors, 131 different descriptors were obtained to use with Random Forest Learning process.

B2Bscore are tested under two different data sets. First is an independent data set that may include any types of protein – ligand complexes. Second is leave-cluster-out cross-validation (LCOCV). In this type, none of the proteins in the test set has a high sequence similarity with any proteins in the training set. This is quite important to measure actual performance of the scoring function. Because designing drugs against new types of protein molecules is harder due to its low similarity with known protein – ligand complexes.

In the experiments, 26 different protein families are selected to use in the tests. Root Mean Squared Error and Pearson Correlation Coefficient are chosen as performance measures. On both tests (independent and LCOCV), B2BScore has a better perfor-

mance than well-known scoring functions used in RFScore, XScore, Autodock Vina, Autodock, and GOLD.

2.3 CIFAP

As stated by Li et al.[15], electrostatic potential plays an important role in protein – ligand binding affinity. In the study published by Erdas et al.[1], this feature is used in the data modeling method called Compressed Images For Affinity Prediction (CIFAP) [1]. In this approach, a candidate ligand docked into the related receptor and a grid cube containing electrostatic potential values of each point was generated from binding site. Each grid cube has 37x37x37 grid points. After extracting the electrostatic potentials of each grid point, 3D structure of the cube is compressed into 2D, by summing the values through X, Y and Z directions. Each 2D structure has 1369 features (electrostatic potential values).

Sequential Forward Selection method was used to reduce the number of features. Starting with an empty set, SFS try to find the best permutation in the feature set, meaning the subset of features leading to the lowest Root Mean Squared Error. Multiple Linear Regression was the prediction method in feature selection. Running 100 times with 10-folds cross-validation, 25 features were selected from X vectors, 27 from Y vectors and 31 from Z vectors are selected.

For the prediction phase, Support Vector Regression (SVR) and Adaptive Neuro-Fuzzy Inference System (ANFIS) methods were selected. In the tests, 57 different complexes of Checkpoint Kinase 1 proteins were used. pIC50 values are selected as the binding affinity measure, which is obtained by $-log_{10}IC_{50}x10^{-}99$ function where IC50 is the half maximal inhibitory concentration.

The experimental results showed a RMSE value of 0.56 for SVR and 1.11 for the ANFIS method which are validated with leave-one-out cross-validation

CHAPTER 3

DATA PREPERATION AND ALGORITHMS

3.1 Ligand Preparation and Docking

This study is an improvement over 2D CIFAP algorithm, so the ligand preparation and docking methods are exactly same with the previous study. First, x-ray crystallography of protein in complex with a ligand that is probably the ligand with the highest affinityis obtained from Protein Data Bank. To make it ready for the docking process, the ligand is discarded from the binding site of the receptor.

Secondly, all ligands are drawn and minimized using MM2 Force Field[16] function of HyperChem 5.1[17], then saved in PDB Format with Discovery Studio Visualizer v.1.7[18]. Each ligand finally docked to the target protein using AutoDock Vina v.1.1.2.[19].

3.2 Obtaining Grid Cubes

This is the part where the molecular descriptor of our method, namely electrostatic potential maps, for each complex is obtained. A grid cube is a cubic frame in which it has a number of sub-cubes that contain electrostatic potential of their coordinates. Our grid cubes are placed on the binding site of the receptor which is thought to have enough electrostatic information about the complex. A grid cube has 37 x 37 x 37 points each has a distance of 0.5 Å from the others. In our study, we will call each point as a sub-cube in order to split from each other or join them together. Then each

sub-cube has a dimensions of 0.5 Å x 0.5 Å x 0.5 Å and there are 50653 sub-cubes in a grid cube. The center coordinates of a grid cube for a receptor is determined by averaging the center coordinates of the ligands. Grid cubes are generated by the AutoGrid4 module of AutoDock v4.2 [20]. The final output is the electrostatic potential values of each sub-cube written in a file in ASCII format.

3.3 Attribute (Feature) Selection

In an application that uses many attributes (50653) but less instances (57 for CHK1, 35 CASP3), it is an important issue to select a attribute selection method. In this work, WEKA[21] 3.6.9 machine learning tool is used for both feature selection and prediction phases. In WEKA, a feature selection method is defined with its evaluator, classifier and search algorithm.

3.3.1 Evaluator (Classifier Subset Evaluator)

An evaluator is the algorithm that determines how attribute subsets are evaluated. In other words, an evaluator chooses if a subset of features will be placed on the final set of attributes or not. There are several evaluator algorithms in WEKA tool[21]. However, the classifier subset evaluator gives the best results for our specific problem. A classifier subset evaluator estimates the worth of a set of attributes by using a classifier. The classifier trains the model with the given subset and instances and the evaluator adds the attributes to the final result if the error is less than some value.

3.3.2 Classifier (Linear Regression)

This is the method which evaluator is used to estimate the worth of the subset. In our problem, Linear Regression is chosen as a classifier for feature selection. Linear Regression is a linear model where linear predictor functions are used. In this regression technique, a dependent variable (predictant) is modeled with more than one explanatory variables (predictors). The relationship between predictant and predictors are defined with following linear function:is the error term and is the dependent variable.

3.3.3 Search Algorithm (Best First Search)

The search algorithm is an important factor for attribute selection, because it is not applicable to estimate the worth of all possible feature subsets. A search algorithm determines which way the search should go. Some of the feature combinations may not be used to train the model by eliminating them. In this study, Best First Search (BFS) has a better overall performance than other search algorithms. Thus, only the results of BFS are given in this report. Best First Search searches the space of attributes with greedy hill climbing, augmented with a backtracking facility. In our case, BFS starts with an empty set of attributes and adds features to the set while it progresses.

Algorithm 1 Best First Search

Input: S = Empty Feature Subset
OPEN = Empty "Feature Subset – Non-improving Nodes Count Tuple" List
CLOSED = Empty "Feature Subset – Non-improving Nodes Count Tuple" List
H = Heuristic Function
N_{max} = Maximum Number of Non-improving Nodes
S_{best} = Null – Feature Subset that has Best Heuristic Function Value
H_{best} = ∞ - Best Heuristic Function Value
Output: S_{best} = Feature Subset that has Best Heuristic Function Value
Add (S,0) to OPEN list
repeat
Get subset (X,n) that has the best score H(X)
Remove X from OPEN put it to CLOSED
if H(X) < H_{best} **then**
$S_{best} \leftarrow X$
$H_{best} \leftarrow H(X)$
end if
if (n+1) < Nmax **then**
Add children of X to OPEN as (Xm, n+1)
end if
until OPEN empty

3.4 Prediction

In this study, SMOreg (Support Vector Regression implementation of WEKA) and PLSClassifier (Partial Least Squares implementation of WEKA) are selected as prediction algorithms.

3.4.1 Support Vector Regression

Support vector regression tries to find a model in the form:

$$f(x) =< w \cdot x > \tag{3.1}$$

where w and x are weight and the input data vector, respectively. In the equation, b is the bias. To deal with noisy data, Vapnik defined a new loss function called ϵ – sensitive loss function. Using this, the algorithm tries to minimize the regression function:

$$\Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i^- + \xi_i^+) \tag{3.2}$$

and the ϵ – sensitive loss function is defined as follows:

$$L_\epsilon(y) = \begin{cases} 0, & \text{if} |f(x) - y| < \epsilon \\ |f(x) - y| - \epsilon, & \text{o}therwise \end{cases} \tag{3.3}$$

The solution to minimize the loss function is calculated by equation-3.3 using Lagrange multipliers α, α^*:

$$\alpha, \alpha^* = argmin\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i + \alpha_i^*)(\alpha_j + \alpha_j^*)\langle x_i, x_j\rangle - \sum_{i=1}^{l}(\alpha_i + \alpha_i^*)y_i - \sum_{i=1}^{l}(\alpha_i + \alpha_i^*)\epsilon \tag{3.4}$$

with constraints:

$$0 \leq \alpha, \alpha^* \leq C, i = 1, \ldots, l \tag{3.5}$$

$$\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) = 0 \tag{3.6}$$

Solving equation 3.4 with constraints 3.5 and 3.6 the lagrange multipliers, weight and bias of regression function is determined.

$$w = \sum_{i=1}^{l} (\alpha - \alpha_i^*) x_i \tag{3.7}$$

$$b = -\frac{1}{2} \langle w, (x_r + x_s) \rangle \tag{3.8}$$

Using the Karush-Kuhn-Tucker (KKT) conditions, it can be concluded support vectors are points where exactly one of the lagrange multipliers is greater than zero. Then, the problem turns into minimization of L_1 loss function where $\epsilon = 0$:

$$min \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \beta_i \beta_j \langle x_i, x_j \rangle - \sum_{j=1}^{l} \beta_i y_i \tag{3.9}$$

with constraints:

$$-C \leq \beta_i \leq C, i = 1, \ldots, l \tag{3.10}$$

$$\sum_{i=1}^{l} \beta_i = 0 \tag{3.11}$$

And finally, regression function becomes:

$$w = \sum_{i=1}^{l} \beta_i x_i \tag{3.12}$$

$$b = -\frac{1}{2} \langle w, (x_r + x_s) \rangle \tag{3.13}$$

3.4.2 Partial Least Squares Regression

Partial Least Squares Regression is an extension of multiple linear regression. In most data-analysis problems, predictions using a linear relationship between dependent and independent variables are enough. Although there are several extensions for MLR (such as discriminant analysis, principal components regression, canonical correlation) to analyze more complex relationships, all of them have restrictions. PLSR removes these restrictions which are:

- Factors underlying X and Y variables cannot be extracted from matrices that have both X and Y variablesi.
- The number of independent variables cannot exceed number of instances. where X and Y are independent and dependent variables respectively.

In this study, an implementation of the PLSR called "PLSClassifier" which is one of the regression models offered by WEKA is used.

3.4.3 Prediction Performance Measures

In the tests throughout this study, the major performance measures are the Root Mean Squared Error(RMSE) and Correlation Coefficient (R^2). RMSE is calculated by averaging the squares of all errors (the difference between the actual and the predicted values) and taking the square root of the value.
To determine if a regression method is predictive or not, R^2 is a choice to measure this. The average of squares of errors is divided by the standard deviation of the variables. One minus this value gives the R^2. A value of 1 for R^2 means a perfect fit with the regression model while 0 is no correlation between actual and predicted data.

CHAPTER 4

DYNAMIC PREDICTION

In this study, we have tried to find a better prediction method for our specific problem. Previous study, 2D CIFAP, has great results; however it can be improved with a more dynamic data model. Its basic strategy which we have previously mentioned is to sum up the electrostatic potential of the points through X, Y and Z directions. However, this may not be the best data model since it always group (sum up) the same points.

4.1 Main Idea

As an improvement for the previous data model, the idea in this study is to group points (sub-cubes) dynamically. Instead of grouping 37 sub-cubes for each point on the surface, it may be better to group a sub-cube with its neighbors in different dimensions. It looks more suitable to consider a 3D region because we will not be compelled with a rule to sum up through same direction for each point. However, this makes problem harder because there are 50653 sub-cubes and each has 6 neighbors in 3 different directions and even we do not know how many sub-cubes we should join together.

To solve these issues, it is a suitable idea to use a top-down approach. At the very beginning, the algorithm starts with the whole grid cube as one piece that has an electrostatic potential in it. The electrostatic potential of the grid cube is sum of all electrostatic potential values of the points in the grid cube. In the second step, the grid cube is divided into two pieces, let's say g_1 and g_2. The division tries to create two

sub-cubes with equal sizes to make the problem easier. It "tries" because sometimes a binary division on a direction with an odd length could result different sized sub-cubes. Now, g_1 and g_2 has non overlapping parts of the starting grid cube and each has an electrostatic potential. Again, each potential value is calculated by summing up all the electrostatic potential values of the points in the sub-cube.

After applying the division method many times, a number of 3D electrostatic potential grid cubes are obtained. Then using the potential values of the sub-cubes as features, the model is trained and the unknown instances are predicted. However, there are two challenges in this method to consider. The first one is that in each step, which direction the method should split the cubes from and the second is when it should stop dividing.

4.2 Dividing Direction

The first challenge to overcome is to find a solution to dividing direction. There are three possibilities considered in this study:

1. Select division that creates two sub-cubes with highest electrostatic potential difference
2. Select division that has a higher accuracy of predicting results (smaller RMSE)
3. Try all combinations

The first option runs as follows: Divide the grid cube and calculate total charge on each division. This is done by summing up all the values on each sub-cube. Then, get the difference of the potential values of the sub-cubes. Finally, divide other directions as well and compare 3 values and select the division that maximizes the difference. With this method, similar charged points grouped together and information loss could be minimized.
Second possible solution to dividing direction uses the idea of prediction ability of the sub-cubes. In this option, the potential values of sub-cubes (obtained by dividing the grid cube) are used as features to learn the model and a Root Mean Squared Error

is calculated. Dividing direction that gives the smallest error is selected.

The final choice is trying all the combinations. However, in terms of computational and memory cost, this option is not applicable. For example, if the algorithms splits the cubes for 12 level there will be $6^{12} = 2.176.782.336$ possible combinations to train and predict.

4.3 Dividing Limit

The algorithm starts with a grid cube and generates many sub-cubes that is used as features for the training. If it runs enough, at the end it creates 37 x 37 x 37 sub-cubes. This is the maximum number of features that can be generated from the starting grid cube. However, we need optimal sub-cube sizes that will generate a better learning performance with a small RMSE value. In this study, we have investigated two different methods for the problem of dividing limit.

4.3.1 Level Based Method

Starting with a grid cube that contains the whole binding site and electrostatic potentials, splitting operation creates new sub-cubes of lower levels. Here, level is defined as the distance to the root node which we have started dividing. In level based method, in each iteration, the cubes of the lowest level are split at the same time.

Since there is no idea of the optimal sub-cube size, we have applied a dynamic learning model. As each iteration ends, the sub-cubes of new level are used as feature subset to learn the model. Then, an error value is calculated by predicting the unknown instances. After dividing the grid cube until none of the sub-cubes can be divided any more, the level which we obtained the minimum RMSE value is the optimal stop point.

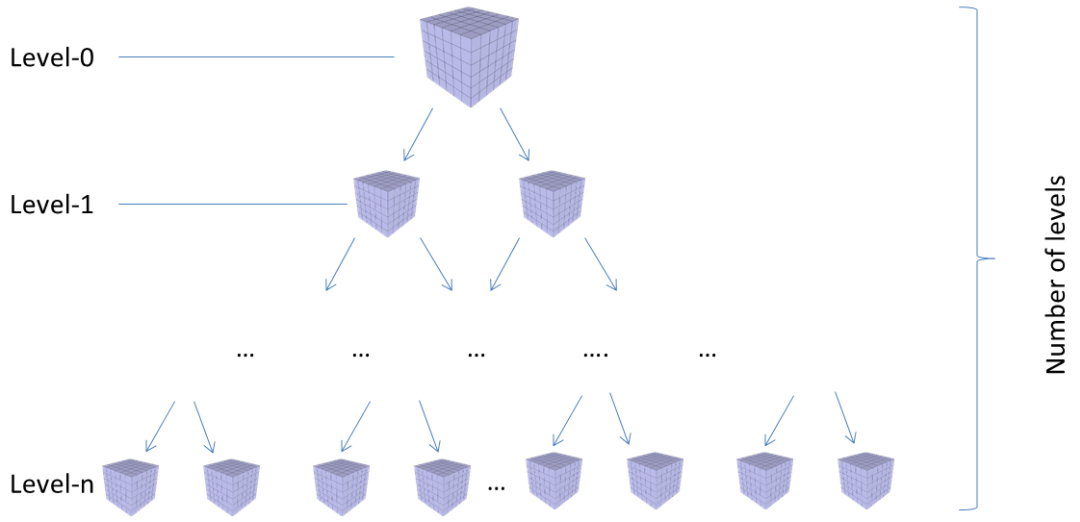


Figure 4.1: An illustration of level based method. Starting with the root grid cube, in each level, the leaf nodes are split into two. Final level has 2^n sub-cubes where n is the number of levels

4.3.2 Individual Sub-cube Method

Dividing sub-cubes on the same level at each iteration and calculating RMSE is a good solution in terms of complexity. However, this reduces the chance of finding important areas in the grid cube. Having sub-cubes with different sizes could give the flexibility to the algorithm in finding important or similar areas on the binding site effectively. Considering this factor, the first idea that comes is a naïve approach, which is diving each cube in the feature set and calculating RMSE in each iteration. Unfortunately, this requires computing the learning algorithm $3x10^{89}$ times and this looks impossible with today's technology. In order to find when to stop dividing the sub-cubes, a heuristic search method is needed. This way, some of the sub-cubes are stopped before reaching the leaf nodes and this could possibly reduce number of combinations.

4.3.2.1 Heuristic Search – 1

This heuristic method has 3 rules to reduce the number of possible combinations by eliminating unpromising splits.

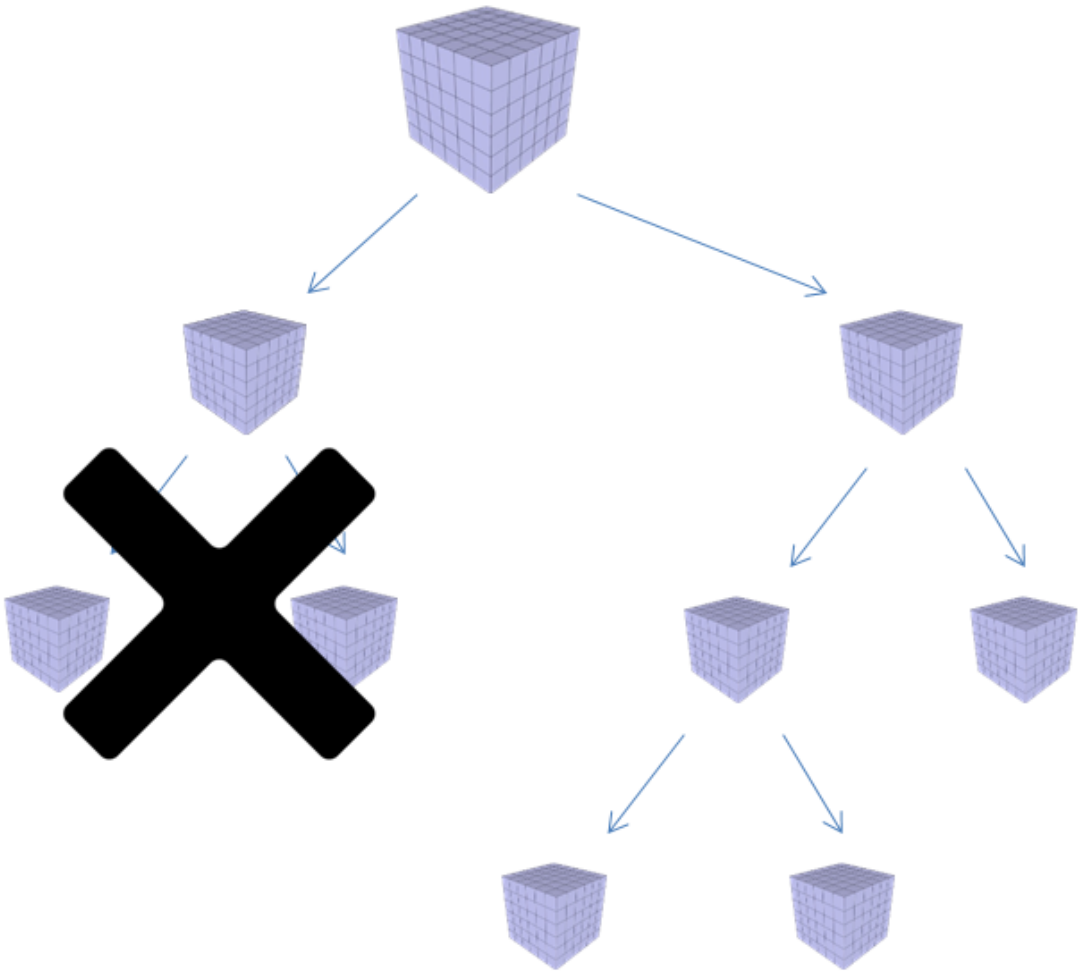


Figure 4.2: A possible situation in heuristic search - 1. The branch on the left side is eliminated because it is an unpromising branch. Other splits satisfy the conditions defined in the algorithm.

Algorithm runs as follows:

Before dividing a sub-cube, calculate the RMSE. After dividing the sub-cube, learn the model again and calculate a new RMSE.

Case – 1: If new error is larger than previous, this is an unpromising branch. Rollback the divide operation.

Case – 2: If new RMSE is equal to the previous, increment a counter for that branch. If the counter is larger than a predefined value, rollback the divide operation, else continue.

Case – 3: If new RMSE is smaller than the previous, this is a promising branch. Con-

tinue splitting.

Case 2 actually count the non improving nodes which is formulated as

$$0 < rmse_2 - rmse_1 < \epsilon \tag{4.1}$$

where $rmse_1$ is the previous error and $rmse_2$ is the current error value. In the algorithm, n_m is defined as maximum number of non improving nodes. If the counter is larger than this value that branch is no more expanded.

4.3.2.2 Heuristic Search – 2

This method uses a dynamic feature selection and prediction to find the best feature subset. First, the grid cube is divided until level-n. After having 2^n sub-cubes, the feature selection method is applied. Feature selection again uses Classifier Subset Evaluator as evaluator, Linear Regression as classifier and Best First Search as search algorithm. Then the model is trained with selected features and an RMSE, let's say e_1, value is calculated. In the second part, unselected features (sub-cubes) of the first part are joined again. Now, there are fewer features than the first part and same feature selection algorithm is applied again. Newly selected features are used in training and a new RMSE, namely e_2 is calculated. If e_1 is greater than e_2, this means the algorithm should keep the selected nodes of the first part and join others again. If e_2 is greater than e_1, then unselected nodes of second part is joined.

This method runs like a backtracking algorithm. First the entire sub-cube tree is obtained until a predefined level. Then, by joining some of the sub-cubes, an optimal subset of sub-cubes is reached. Figure - 4.3 shows a possible situation in heuristic search - 2

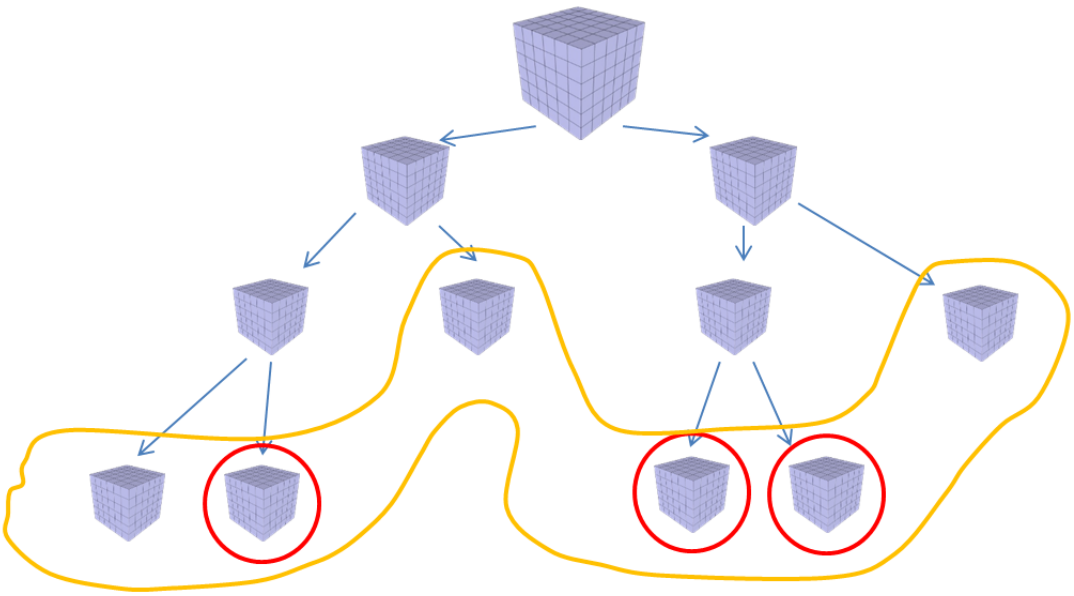


Figure 4.3: A snapshot from heuristic search - 2. The grid cubes with red circles are the ones selected by the feature selection algorithm. After selection, model is trained and an RMSE value is calculated. Then unselected nodes are joined. Sub-cubes encircled with orange line is the feature set of the next prediction.

CHAPTER 5

EXPERIMENTS

5.1 Checkpoint Kinase 1 experiments

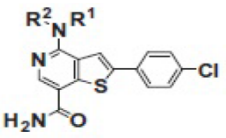
In this part of the chapter, target protein Checkpoint Kinase 1 is explained with its properties and duties in the cell. Then, the experimental results related to it are given with tables and graphs.

5.1.1 Checkpoint Kinase 1

Checkpoint Kinase 1 is a Serine/threonine-specific protein kinase in humans, which plays an important role at cell cycle control[22]. An healthy cell cycle is arrested at G1 and G2/M checkpoints. The major approach used in cancer therapy is DNA damaging cancerous cells. Many of the DNA damaging agents uses p53 dependency of G1 phase and Chk1 dependency of G2 phase of the cancerous cells. Since 50% of the cancerous cells are p53 deficient, CHK1 inhibition especially plays an important role on these cells. Inhibition of CHK1 increases the cell sensitivity to DNA damage agents. This either kills the cancerous cell or the damaged DNA is repaired. Although none of the Chk1 inhibitors passed the phase III clinical trials, they are still being developed and tested for cancer threapy[23]. Other than being targeted in cancer therapy, Chk1 proteins also play an important role on DNA repair processes, gene transcription, embryo development, cellular responses to HIV infection and somatic cell viability[24].

5.1.2 Preparing Initial Data

The initial grid cubes of 57 CHK1 complex that has 37 x 37 x 37 grid points are taken from CIFAP study to compare results effectively. In CIFAP, the X-ray crystallograpy of CHK1 in complex with compound70 is obtained from the Protein Data Bank (PDB ID: 3PA3). The center coordinates of the binding site which is (20, -3, 11) is used as the center of the grid cube that covers all the ligands with minimum volume. The compressing part is not used in this study while obtaining the grid cubes from CIFAP because this study uses grid cubes as 3D structures. The inhibitors are retrieved from the study of Zhao et al[10]. Figure - 5.1, 5.2, 5.3 show the 57 ligands that are docked to Chk1. The feature selection process will be explained in the prediction part because of the dynamic approach of the algorithm.



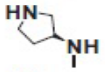
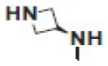
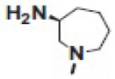
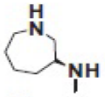
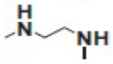
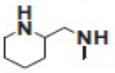
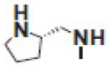
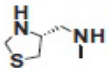
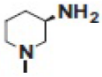
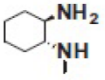
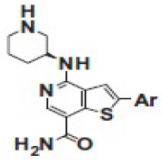
Compound	R^1	R^2	CHK1 IC_{50} (nM)
1	—	—	2
2a	H		3
2b	H		14
2c	H		707
2d			746
2e	H		3
2f	H		7
2g	H		292
2h	H		197
2i	H		937
2j			148
2k	H		211
2l	H		27,143

Figure 5.1: SAR at 4-position of thienopyridine[1]



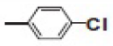
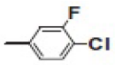
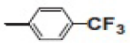
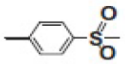
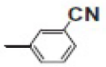
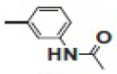
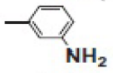
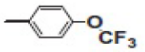
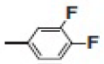
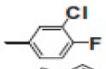
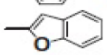
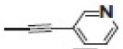
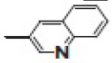
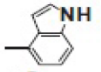
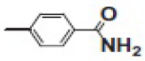
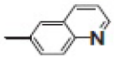
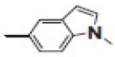
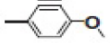
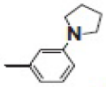
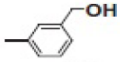
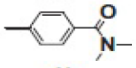
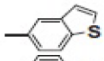
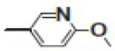
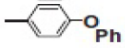
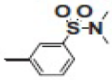
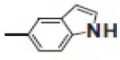
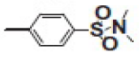
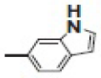
Compound	Ar	IC_{50} (nM)	Compound	Ar	IC_{50} (nM)
2a		3	39		9
19		24	40		5
20		5	41		2
21		4	42		7
22		2	43		5
23		2	44		9
24		3	45		60
25		2	46		4
26		5	47		2
27		4	48		67
28		21	49		34
29		5	50		1
30		19	51		5
31		3	52		8
32		3	53		3
33		3	54		1
34		4	55		3
35		4	56		20
36		60	57		16
37		16	58	Br	56
38		9	59	H	717

Figure 5.2: SAR at 2-position of thienopyridine[1]

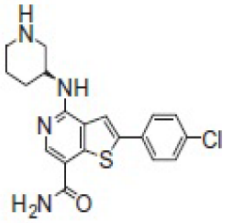
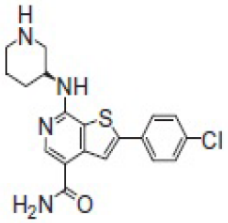
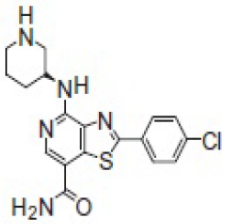
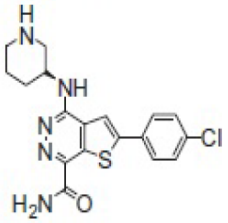
Compound	Structure	CHK1 IC_{50} (nM)
2a		3
60		4582
69		9
70		1

Figure 5.3: SAR of core modification of thienopyridine[1]

5.1.3 Prediction

5.1.3.1 Level Based Method Results

As it is explained in the fourth chapter, level based method divides the grid cube until the determined level. When the tree is constructed, the electrostatic potential of the leaf nodes are the features of the prediction process. It is appropriate to experiment as many levels as possible, because there is no information about the optimal size of a sub-cube. For each level, the first operation after the leaf nodes are constructed is to apply the feature selection. Using Classifier Subset Evaluator with Linear Regression and Best First Search of the WEKA Tool[21], attributes are selected. These attributes are the inputs for SVR and PLSR learning methods with the corresponding pIC50 value of the complex.

In Table 5.1, different levels, number of features, sub-cube sizes and learning results are given. The RMSE values are calculated with leave-one-out cross validation. As it is seen, the best results are obtained in level – 10. The results of SVR are better than PLSR in this level. When compared with CIFAP, the RMSE of 0.489 of 3D analysis is lower than SVR error of CIFAP. For the overall performance, PLSR performance (RMSE = 0.411) of CIFAP over performed the results of this method. The error values after level – 11 becomes larger due to high number of features, thus they are ignored.
In Table 5.2, the predicted results of both PLSR and SVR of each complex is given. The values belong to level-10 in which the best results (lowest error) are observed. Figure - 5.4 shows the scatter plot for SVR predictions for CHK1 complexes. Similarly, Figure - 5.5 shows PLSR predictions for 57 complexes. As it can be observed, SVR results are closer to the "perfect prediction" line.

Random subsampling test results are given in Table - 5.3 and 5.4 for SVR and PLSR respectively. The results are parallel with the leave-one-out cross validation for "test" sets. For the test sets, the correlation coefficient is also greater for SVR. The training set results shows PLSR is better when all the instances are used for both training and test phases, which is not a situation in the real world problems where there might be no observed value for an instance. The tables also shows the highest correlated 3

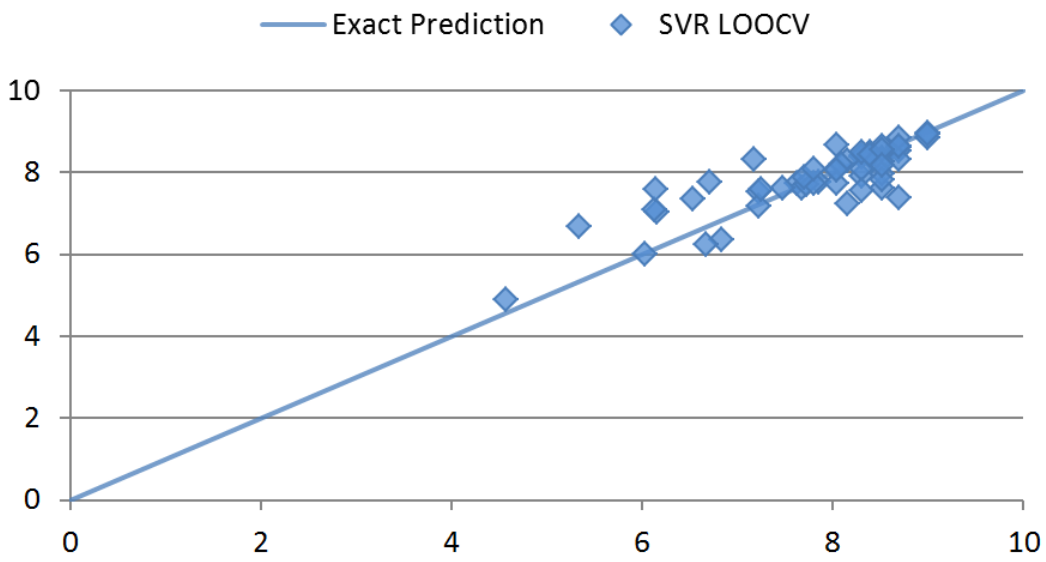


Figure 5.4: SVR prediction and actual binding affinity (pIC50) values of CHK1 complexes with leave-one-out cross validation.

subsamplings.

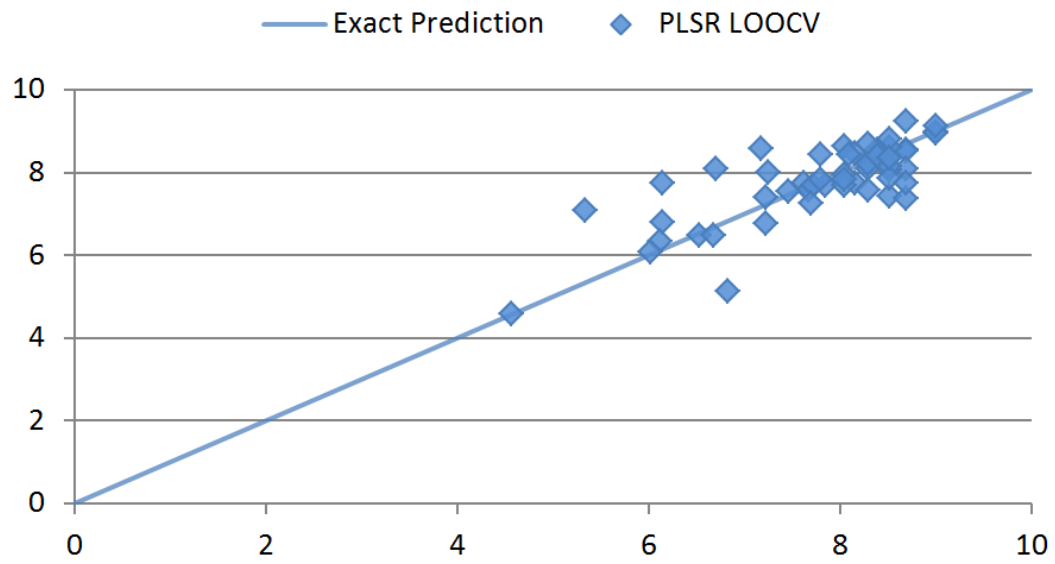


Figure 5.5: PLSR prediction and actual binding affinity (pIC50) values of CHK1 complexes with leave-one-out cross validation.

Table 5.1: Level based method RMSE value comparison for different levels of CHK1 complexes. RMSE values are calculated with Leave-one-out cross validation.

Level No.	Features Before Feature Selection	Average Sub-cube Size	Features After Feature Selection	SVR RMSE	PLSR RMSE
Level - 4	16	3166	7	0.736	0.737
Level - 5	32	1583	9	0.836	0.781
Level - 6	64	791	11	0.751	0.756
Level - 7	128	395	4	0.71	0.72
Level - 8	256	198	11	0.79	0.713
Level - 9	512	99	5	0.694	0.717
Level - 10	1024	49	20	0.489	0.618
Level - 11	2048	25	20	0.569	0.698

Table 5.2: Level based method prediction results of level-10 compared to actual binding affinity (pIC50) values of the CHK1 complexes. The inhibitors of the CHK1 are ligands that are published in Zhao et al. Leave-one-out cross validation is used in the tests.

No.	Actual	SVR	PLSR	No.	Actual	SVR	PLSR
1	8.699	8.517	8.519	35	8.398	8.388	8.463
2a	8.523	7.634	7.394	36	7.222	7.524	7.378
2b	7.854	7.756	7.666	37	7.796	7.729	7.807
2c	6.151	7.036	6.778	38	8.046	8.667	8.612
2d	6.127	7.094	6.314	39	8.046	8.081	7.935
2e	8.523	8.658	8.336	40	8.301	8.092	8.174
2f	8.155	7.234	7.71	41	8.699	8.848	9.21
2g	6.535	7.346	6.462	42	8.155	8.286	8.447
2h	6.706	7.757	8.079	43	8.301	8.488	8.656
2i	6.028	5.999	6.053	44	8.046	7.747	7.672
2j	6.83	6.357	5.108	45	7.222	7.188	6.756
2k	6.676	6.251	6.464	46	8.398	8.436	8.379
2l	4.566	4.893	4.549	47	8.699	8.657	7.73
19	7.62	7.737	7.715	48	7.174	8.337	8.543
20	8.301	8.381	8.206	49	7.469	7.628	7.506
21	8.398	8.499	8.462	50	9	8.86	8.965
22	8.699	8.613	8.486	51	8.301	7.562	7.543
23	8.699	7.392	7.338	52	8.097	8.2	8.408
24	8.523	8.366	8.585	53	8.523	8.555	8.783
25	8.699	8.333	8.06	54	9	8.97	8.938
26	8.301	8.403	8.302	55	8.523	8.178	8.317
27	8.398	8.445	8.524	56	7.699	7.896	7.217
28	7.678	7.611	7.539	57	7.796	8.082	8.398
29	8.301	7.916	8.059	58	7.252	7.577	7.986
30	7.721	7.719	7.67	59	6.145	7.584	7.733
31	8.523	8.263	8.024	60	5.339	6.673	7.069
32	8.523	7.985	8.117	69	8.046	8.056	7.817
33	8.523	7.836	7.831	70	9	8.928	9.098
34	8.398	8.416	8.473	RMSE		0.489	0.618

Table 5.3: Random subsampling test results of CHK1 complexes. SVR is used for regression. 1000 random samples are tested. Random-1, Random-2, Random-3 are the highest correlated test sets with binding affinity values.

SVR	test		train	
	RMSE	R2	RMSE	R2
Random – 1	0.189	0.989	0.471	0.847
Random – 2	0.257	0.983	0.489	0.843
Random – 3	0.294	0.982	0.458	0.850
Average	0.576	0.763	0.423	0.893

Table 5.4: Random subsampling test results of CHK1 complexes. PLSR is used for regression. 1000 random samples are tested. Random-1, Random-2, Random-3 are the highest correlated test sets with binding affinity values.

PLSR	test		train	
	RMSE	R2	RMSE	R2
Random – 1	0.187	0.990	0.451	0.863
Random – 2	0.195	0.987	0.452	0.860
Random – 3	0.207	0.985	0.447	0.869
Average	0.704	0.716	0.383	0.912

5.1.3.2 Individual Sub-cube Results

In this part, the results of the individual sub-cube method which is explained in the previous chapter is given. The positive effect of individual sub-cube method on the error values can be understood from Table - 5.5 from exhaustive search columns. However, the runtime gets bigger and bigger in each level for. It possibly takes a month or more for level 6 to find a result. Anyway, it is shown that this method could find results with lower RMSE than level based method.

"Heuristic -1" columns of Table - 5.5 represents the heuristic - 1 results. As it is shown it finds same subcube configuration with exhaustive search for level - 4 and level - 5 . The runtime is also reduced effectively with this approach. However, the error values become stable after level - 5. This is probably because search method stucks at a local minimum.

The effect of the first heuristic is not very promising, however this approach could somehow give the following idea: In the search space of grid sub-cube configurations, there is a sub-cube configuration with a lower error value than the RMSE of level based approach. And this can be found with a heuristic function that completes in a reasonable runtime. Table - 5.6 shows the results of heuristic - 2. This method finds at least as low error as the level based method. However, the results are only better in levels 5 and 8. This search method runs very fast, eliminate most of the feature subsets. It is again demonstrated that there is a better feature subset with a lower RMSE than level based method.

Table 5.5: The error and runtime in seconds of CHK1 complexes based on the predictions of SVR technique with heuristic search - 1. The runtime values are for informative purposes and highly dependent on the hardware running on. On the tests, heuristic - 1 search is used with $\epsilon = 0.08$ and $m_n = 2$

SVR	Level Based RMSE	Individual Subcube			
		Exhaustive Search		Heuristic - 1	
		RMSE	Run Time (sec)	RMSE	Run Time (sec)
Level – 4	0.736	0.711	123	0.711	10
Level – 5	0.836	0.699	13900	0.699	28
Level – 6	0.751	?	?	0.699	65
Level – 7	0.71	?	?	0.699	118
Level – 8	0.79	?	?	0.699	215
Level – 9	0.694	?	?	0.699	390
Level – 10	0.489	?	?	0.699	797

Table 5.6: The error and runtime in seconds of CHK1 complexes based on the predictions of SVR technique with heuristic search - 2. The runtime values are for informative purposes and highly dependent on the hardware running on.

SVR	Level Based RMSE	Individual Subcube			
		Exhaustive Search		Heuristic - 2	
		RMSE	Run Time (sec)	RMSE	Run Time (sec)
Level – 4	0.736	0.711	123	0.736	3
Level – 5	0.836	0.699	13900	0.755	5
Level – 6	0.751	?	?	0.751	6
Level – 7	0.71	?	?	0.71	8
Level – 8	0.79	?	?	0.71	12
Level – 9	0.694	?	?	0.694	18
Level – 10	0.489	?	?	0.489	43

5.2 Caspace 3 experiments

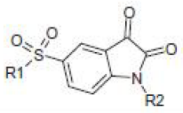
In this part, same experiments are applied with Caspase3 inhibitors

5.2.1 Caspace 3

Caspase-3 is a caspase protein which is encoded by CASP3 gene. it is a member of the cysteine-aspartic acid protease. Working together with other caspace proteins, it plays an important role at cell apostosis, in other words programmed cell deaths.

5.2.2 Preparing Initial Data

The same approach in CHK1 is used for CASP3 experiments too. The grid cubes that contain electrostatic potentials of grid points are obtained from CIFAP study. In that study, the CASP3 complex with compound1 is gathered from Protein Data Bank (PDB ID:1GFW). The center coordinates of the CASP3 grid cubes are located at (39,36,27). The 35 inhibitors are extracted from the study of Wang et. al. In addtion the pIC50 values are adopted from the study of Hasegawa et. al



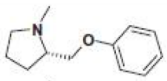
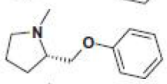
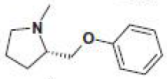
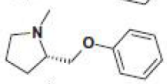
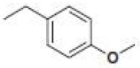
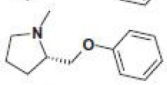
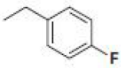
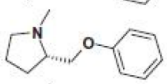
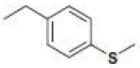
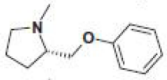
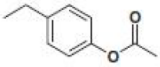
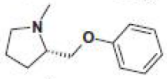
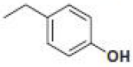
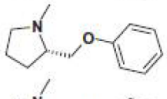
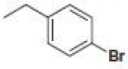
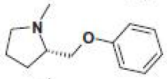
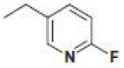
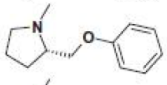
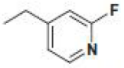
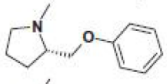
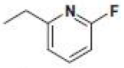
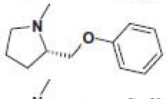
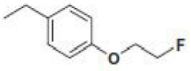
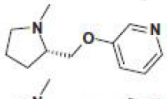
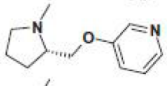
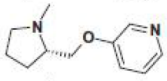
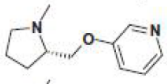
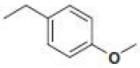
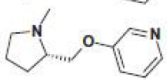
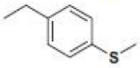
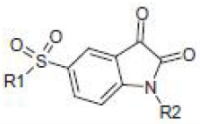
No.	R1	R2	pIC_{50}
1		$-CH_3$	6.92
2		–H	6..62
3			7.91
4			7.84
5			7.92
6			7.91
7			7.92
8			7.87
9			8.01
10			7.99
11			7.67
12			8.04
13			8.01
14		–H	7.23
15		$-CH_3$	7.63
16			8.28
17			8.41
18			8.36

Figure 5.6: Structures and pIC50 values of 1 - 18 ligands of Caspase3 [1]



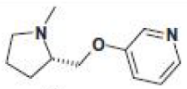
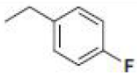
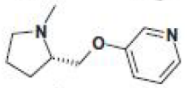
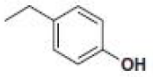
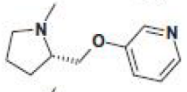
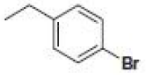
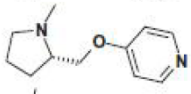
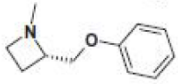
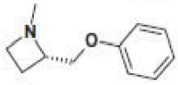
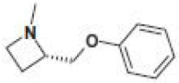
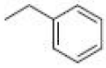
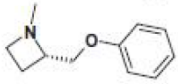
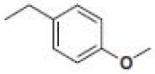
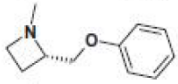
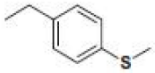
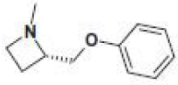
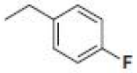
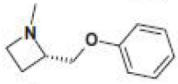
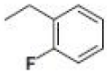
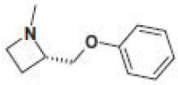
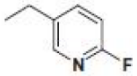
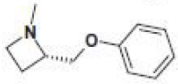
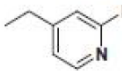
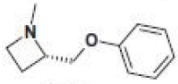
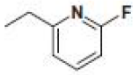
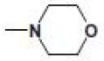
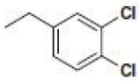
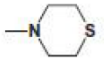
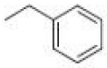
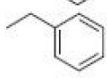
No.	R1	R2	pIC_{50}
19			8.08
20			8.41
21			8.44
22		–H	7.69
23		–H	6.54
24		$-CH_3$	7.04
25			8.01
26			8.08
27			7.95
28			8.06
29			8.03
30			7.96
31			7.53
32			8.24
33			5.84
34			5.99
35			6.94

Figure 5.7: Structures and pIC50 values of 19-35 ligands of Caspase3 [1]

5.2.3 Prediction

5.2.3.1 Level Based Method Results

Prediction phase of CASP3 binding affinities are exactly the same with CHK1 complexes. For several levels, namely levels between 4-11, the grid cube is divide in a binary fashion which creates 2^n features where n is the number of levels. Then feature selection algorithm is applied to the features. The final feature set is used to train the SVR and PLSR methods.

In Table - 5.7, leave-one-out cross validation test results of SVR and PLSR are represented in the same table. The results are quite parallel with CHK1 results. SVR has the best results at level - 10 where the error is as small as 0.102. For PLSR, there is an interesting situation for level - 9, where the error is 0.001. This is a very small value for a prediction where the validation method is leave-one-out cross validation. This result can be thought as an overfitting situation, because random subsampling error of PLSR for test sets in level - 9 is fairly high.

When compared to CASP3 results of CIFAP, the error values for both SVR and PLSR at level - 10 with LOOCV is slightly lower. The errors are 0.102 and 0.107 for level based method of SVR and PLSR respectively, while CIFAP has RMSE values of 0.110 and 0.150 for SVR and PLSR respectively.

Table - 5.8 shows the output values of each prediction for SVR and PLSR. In addition, calculated root mean square errors also can be found in this table. They seem close to the perfect prediction line. Similarly, figure - 5.9 shows the same plot for PLSR predictions. In figure 5.8, the scatter plot of SVR predictions are displayed.

Table 5.7: Level based method RMSE value comparison for different levels of CASP3 complexes. RMSE values are calculated with Leave-one-out cross validation.

Level No.	Features Before Feature Selection	Average Sub-cube Size	Features After Feature Selection	SVR RMSE	PLSR RMSE
level - 4	16	3166	7	0.658	0.782
level - 5	32	1583	3	0.464	0.488
level - 6	64	791	16	0.25	0.247
level - 7	128	395	21	0.266	0.199
level - 8	256	198	14	0.146	0.154
level - 9	512	99	33	0.114	0.001
level - 10	1024	49	14	0.102	0.107
level - 11	2048	25	14	0.144	0.227

Table 5.8: Level based method prediction results of level-10 compared to actual binding affinity (pIC50) values of the CASP3 complexes. The inhibitors of the CASP3 are ligands that are published in Wang et. al. Leave-one-out cross validation is used in the tests.

No.	pIC50	SVR	PLS	No.	pIC50	SVR	PLS
1	6.92	6.802	6.856	19	8.08	8.065	8.03
2	6.62	6.783	6.766	20	8.41	8.304	8.294
3	7.91	7.925	7.934	21	8.44	8.268	8.265
4	7.84	7.755	7.769	22	7.69	7.673	7.685
5	7.92	7.916	7.915	23	6.54	6.757	6.769
6	7.91	7.912	7.912	24	7.04	6.793	6.785
7	7.92	7.954	7.92	25	8.01	7.958	7.974
8	7.87	7.898	7.918	26	8.08	7.921	7.919
9	8.01	7.924	7.914	27	7.95	8.01	8.061
10	7.99	7.984	8.006	28	8.06	8.049	8.059
11	7.67	7.681	7.706	29	8.03	8.104	7.987
12	8.04	8.143	8.295	30	7.96	7.963	7.998
13	8.01	7.982	7.999	31	7.53	7.632	7.494
14	7.23	7.331	7.308	32	8.24	8.058	8.154
15	7.63	7.468	7.507	33	5.84	5.811	5.823
16	8.28	8.416	8.459	34	5.99	6.013	6.029
17	8.41	8.461	8.511	35	6.94	7.03	7.031
18	8.36	8.371	8.408	RMSE		0.102	0.107

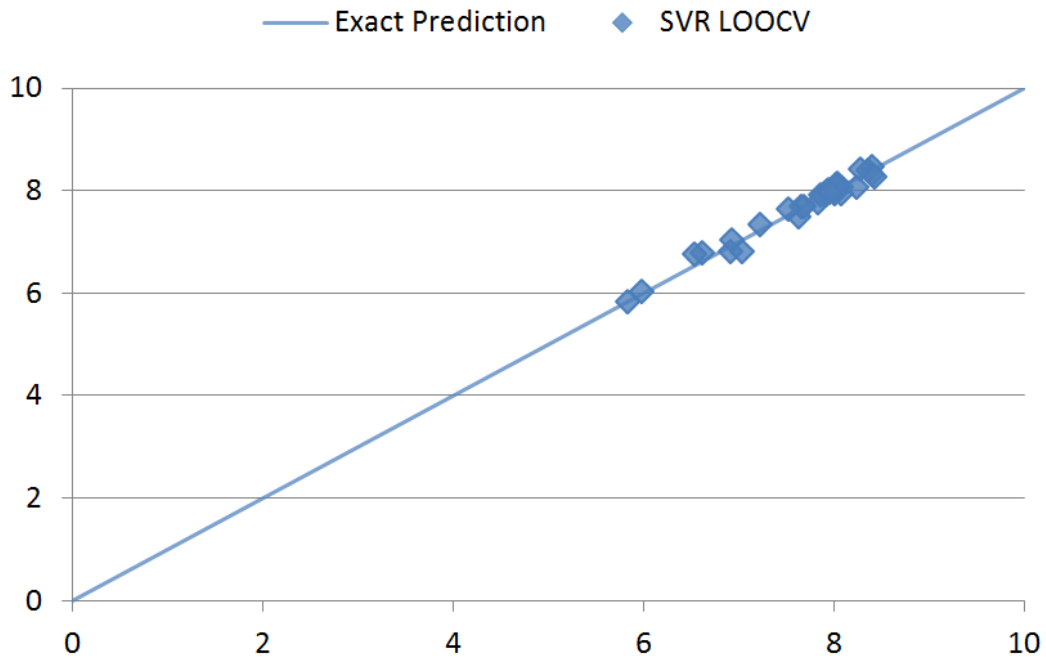


Figure 5.8: SVR prediction and actual binding affinity (pIC50) values of CASP3 complexes with leave-one-out cross validation.

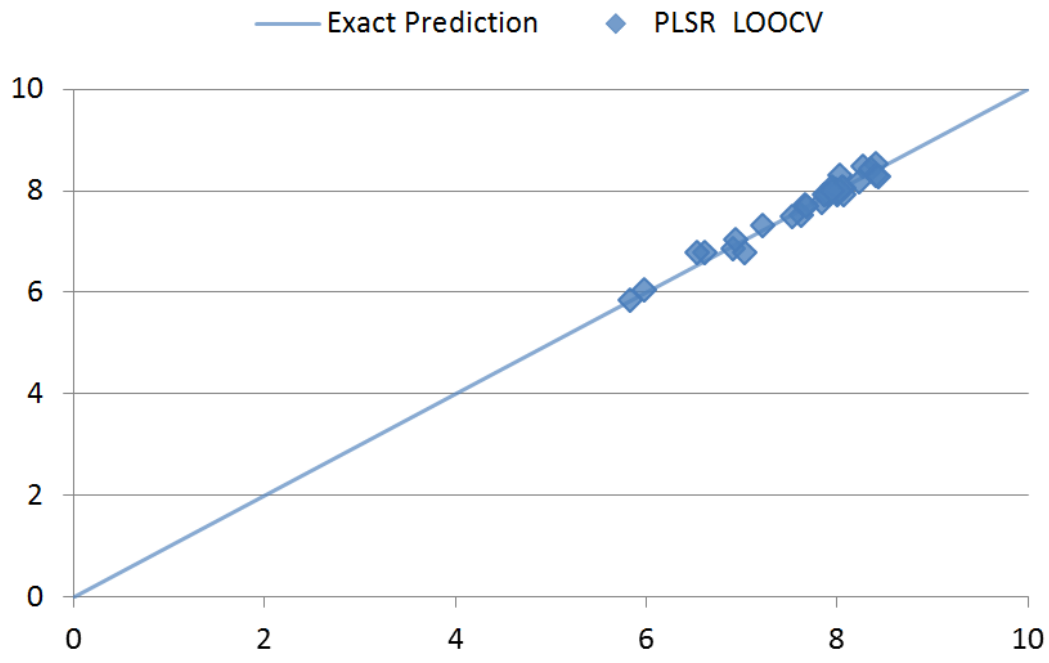


Figure 5.9: PLSR prediction and actual binding affinity (pIC50) values of CASP3 complexes with leave-one-out cross validation.

Random subsampling test results are also given in Table - 5.9 and Table5.10. The tables show the RMSE and R2 values of highest correlated 3 subsamples and the average of 1000 subsamples. The test results has higher RMSE values than leave-one-out cross validation errors. This is a usual case in validation tests because in random subsampling training sets are smaller than the ones of LOOCV. These results also supports the idea that SVR better performed than PLSR for CASP3 inhibitors. CIFAP has slightly lower RMSE value for CASP3 comlexes with 0.162 RMSE value for test samples predicted with PLSR. However, the correlation coefficient of 0.904 looks lower than the results of this study. SVR has value of 0.939 as correlation coefficient.

Table 5.9: Random subsampling test results of CASP3 complexes. SVR is used for regression. 1000 random samples are tested. Random-1, Random-2, Random-3 are the highest correlated test sets with binding affinity values.

SVR	test		train	
	RMSE	R2	RMSE	R2
Random – 1	0.160	1.000	0.067	0.995
Random – 2	0.066	1.000	0.075	0.992
Random – 3	0.154	0.999	0.052	0.996
Average	0.166	0.939	0.073	0.994

Table 5.10: Random subsampling test results of CASP3 complexes. PLSR is used for regression. 1000 random samples are tested. Random-1, Random-2, Random-3 are the highest correlated test sets with binding affinity values.

PLSR	test		train	
	RMSE	R2	RMSE	R2
Random – 1	0.153	1.000	0.050	0.997
Random – 2	0.132	1.000	0.060	0.996
Random – 3	0.175	1.000	0.075	0.990
Average	0.170	0.938	0.066	0.995

5.2.3.2 Individual Sub-cube Results

In this part, the results of individual sub-cube method results for CASP3 complexes are given and discussed Similar to the CHK1 results, for CASP3 exhaustive search finds better results for level 4 and level -5. However, it was not possible to run this search method for level - 6 with a runtime smaller than 2 days. However, it can be supposed that this method would find better results than level based method if it could. The results of exhaustive search can be found in Table - 5.11 from exhaustive search columns.

"Heuristic -1" columns of Table - 5.11 represents the heuristic - 1 results. This time, heuristic search method could find better results than level based method until level - 6. However, it again stucks at a local minimum. The branches that are eliminated while searching the feature subset is the main cause of this local minimum stuck.

Heuristic - 2 results have lower RMSE values only for level - 4. The runtimes are quite low again as it is in CHK1 results of heuristic search - 2. Other levels have equal RMSE values to level based RMSE values. The results again show the heuristic search methods are not very efficient. But these could be the start points for another research because it may be possible to find a sub-cube configuration, i.e. feature subset, that has a lower RMSE than level based method in a reasonable runtime.

Table 5.11: The error and runtime in seconds of CASP3 complexes based on the predictions of SVR technique with heuristic search - 1. The runtime values are for informative purposes and highly dependent on the hardware running on. On the tests, heuristic - 1 search is used with $\epsilon = 0.13$ and $m_n = 2$

SVR	Level Based RMSE	Individual Subcube			
		Exhaustive Search		Heuristic - 1	
		RMSE	Run Time (sec)	RMSE	Run Time (sec)
Level – 4	0.658	0.575	105	0.575	19
Level – 5	0.464	0.314	11231	0.314	643
Level – 6	0.25	?	?	0.192	2320
Level – 7	0.266	?	?	0.192	3210
Level – 8	0.146	?	?	0.192	3890
Level – 9	0.114	?	?	0.192	4800
Level – 10	0.102	?	?	0.192	5210

Table 5.12: The error and runtime in seconds of CASP3 complexes based on the predictions of SVR technique with heuristic search - 2. The runtime values are for informative purposes and highly dependent on the hardware running on.

SVR	Level Based RMSE	Individual Subcube			
		Exhaustive Search		Heuristic - 2	
		RMSE	Run Time (sec)	RMSE	Run Time (sec)
Level – 4	0.658	0.575	123	0.613	2
Level – 5	0.464	0.314	13900	0.464	3
Level – 6	0.25	?	?	0.25	4
Level – 7	0.266	?	?	0.266	6
Level – 8	0.146	?	?	0.146	8
Level – 9	0.114	?	?	0.114	81
Level – 10	0.102	?	?	0.102	95

CHAPTER 6

CONCLUSIONS AND DISCUSSIONS

Predicting the binding affinity of unknown protein - ligand interactions is a hard challenge to solve. To reduce the complexity of drug discovery process, in silico methods provides intelligent and effective solutions for this problem. In this study, we also try to find a 3D analysis method to analyze the binding site of the complexes. The molecular descriptor used in the study is electrostatic potential. It is already proven to be useful to analyze electrostatic potential values of binding sites for affinity prediction in the study of Erdas et al.[1]

In the scope of this study, a new data model which is modified version of 2D CIFAP[1], a dynamic feature selection and prediction techniques are proposed. The dynamic feature selection algorithm is applied using classifier subset evaluator, multiple linear regression and best first search techniques of WEKA Tool. They are called dynamic because they are applied in different phases of data modeling. 2 distinct prediction method used in this study are support vector regression and partial least square regression.

The initial data which is the grid cubes that contain the electrostatic potential values of 50653 grid points of binding site of the complexes are obtained from the study CIFAP. This way, it is possible to compare the 2D data model of CIFAP with 3D data model proposed in this study.

The experiments are applied in 2 subsections called level based method and individual sub-cube method. For level based method in affinity prediction of CHK1 complexes,

SVR method has the best results for both LOOCV and random subsampling tests. SVR has RMSE values of 0.489 and 0.576 for LOOCV and random subsampling test subsets respectively. PLSR poorly performs for CHK1 tests with RMSE values of 0.618 and 0.704 when compared to SVR. The correlation coefficients for random subsampling test sets support the idea that SVR overperformed PLSR for the data model proposed. The correlation coefficient for SVR is 0.763 where it is 0.716 for PLSR.

For CASP3 inhibitors results look similar. Only for level - 9 in LOOCV, PLSR has an error of 0.001 which can be concluded as an overfitting situation by considering PLSR performance for random subsamples. It has an RMSE of 0.501 and R^2 of 0.712 for random subsample test sets which can be thought as a poor performance for CASP3. If we ignore the level - 9 performance of PLSR, a similar situtaion of CHK1 can be noted in LOOCV tests. Both SVR and PLSR find lowest errors at level - 10. For SVR the RMSE is 0.102 and for PLSR it is 0.107. Both performances are better than CIFAP results. CIFAP has an RMSE of 0.110 with PLSR in LOOCV tests. Random subsampling results also proven that SVR is preferable to PLSR with this data model.

The individual sub-cube method actually finds good results for exhaustive search, but it is impossible to search the space of feature subsets without a heuristic due to the high number of possible combinations. The result could not be calculated after level 6 for both CHK1 and CASP3 complexes. The heuristic search methods that are defined for this problem actually do not performed well. But they have proven that it may be possible to find better solutions without searching the whole space of feature subsets.

The heuristics may also show that the problem of estimating binding affinities using electrostatic grid cubes is a different type of problem that includes its own properties. The electrostatic potential values of two points in a grid cube could effect the overall affinity by creating a force between each other. The points may be grouped together to construct a larger volume and this could form a larger force. This research actually tries to find the important areas in the binding site that are not compelled to have equal sizes. Instead, there may be some volumes that are larger because the points

all together have a meaning in the resulting binding affinity. In contrast, some of the points may have smaller regions that have less atoms but affect the binding affinity of the whole complex too much.

The Individual sub-cube method exactly intended to consider grid cube points in this manner. However, more efficient and intelligent search algorithms are needed here to find the optimal configuration. Future studies may target finding a heuristic to achieve better results. It is possible to find a heuristic even by using the coordinates of ligands, coordinates of individual atoms etc.

REFERENCES

[1] Ozlem Erdas, Cenk A. Andac, A. Selen Gurkan-Alp, Ferda Nur Alpaslan, and Erdem Buyukbingol. Compressed images for affinity prediction (cifap): a study on predicting binding affinities for checkpoint kinase 1 protein inhibitors. *Journal of Chemometrics*, 27(6):155–164, 2013.

[2] V. Srinivasa Rao and K. Srinivas. Modern drug discovery process: An in silico approach. *Journal of Bioinformatics and Sequence Analysis*, 2(5):71–74, 2011.

[3] Nikil Wale. Modern drug discovery process: An in silico approach. *Drug Development Research*, 72(1):112–119, 2011.

[4] Rosette M. Roat-Malone. *Bioinorganic Chemistry: A Short Course*. John Wiley and Sons, Inc, 2007.

[5] K. Lundstrom. Structural genomics on membrane proteins: Mini review. *Combinatorial Chemistry and High Throughput Screening*, 7(5):431–439, 2004.

[6] Thomas Lengauer and Matthias Rarey. *Computational methods for biomolecular docking*. GMD - Forschungszentrum Informationstechnik, 1996.

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

[8] Bahaa Saleh, editor. *Introduction to Subsurface Imaging*. Cambridge University Press, 2011.

[9] Eva Stjernschantz and Chris Oostenbrink. Improved ligand-protein binding affinity predictions using multiple binding modes. *Biophysical Journal*, 98:2682–2691, 2010.

[10] Zhao L., Zhang Y., Dai C., Guzi T., Wiswell D., Seghezzi W., Parry D., Fischmann T., and Siddiqui M.A. Design, synthesis and sar of thienopyridines as potent chk1 inhibitors. *Bioorganic and Medicinal Chemistry Letters*, 20:7216–7221, 2010.

[11] Q. Wang, R.H. Mach, and D.E. Reichert. Docking and 3d-qsar studies on isatin sulfonamide analogues as caspase-3 inhibitors. *J. Chem. Inf. Model.*, 49(8):1963–1973, 2009.

[12] Reiji Teramoto and Hisashi Kashima. Prediction of protein-ligand binding affinities using multiple instance learning. *Journal of Molecular Graphics and Modelling*, 29:492–497, 2013.

[13] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[14] Qian Liu, Chee Keong Kwoh, and Jinyan Li. Binding affinity prediction for protein-ligand complexes based on beta contacts and b factor. *Journal of Chemical Information and Modeling*, 53:3076–3085, 2013.

[15] S. Li, L. Xi, C. Wang, J. Li, B. Lei, H. Liu, , and X. Yao. A novel method for protein-ligand binding affinity prediction and the related descriptors exploration. *Journal of Computational Chemistry*, 30:900–909, 2009.

[16] D. N. Sathyanarayana. *Vibrational Spectroscopy: Theory And Applications*. New Age International, 2004.

[17] *HyperChem 5.1*. Hypercube Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA, `http://www.hyper.com`.

[18] *Discovery Studio Visualizer v.1.7*. Accelrys Software Inc., San Diego, `http://www.accelrys.com/`.

[19] O. Trott and A. J. Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry*, 31:455–461, 2010.

[20] Hwangseo Park, Jinuk Lee, and Sangyoub Lee. Critical assessment of the automated autodock as a new docking tool for virtual screening. *Proteins: Structure, Function and Bioinformatics*, 65:549–554, 2006.

[21] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2006.

[22] Mallikarjun Patil, Navjotsingh Pabla, and Zheng Dong. Checkpoint kinase 1 in dna damage response and cell cycle regulation. *SIGKDD Explorations*, 70(21):4009–4021, 2013.

[23] Z.F. Tao and N.H. Lin. Chk1 inhibitors for novel cancer treatment. *Anti-Cancer Agents Med. Chem.*, 4:377–388, 2006.

[24] Youwei Zhang and Tony Hunter. Roles of chk1 in cell biology and cancer therapy. *International Journal of Cancer*, 134(5):1013–23, 2013.