Predicting Drug Synergy Using Data Mining

by

MILAD HASSANI

Submitted to
the Graduate School of Engineering and Natural Sciences
in partial requirements for the degree of
Master of Science

SABANCI UNIVERSITY

August 2016

Predicting Drug Synergy Using Datamining

APPROVED BY:

Assoc. Prof. Dr. .....Kemal Kılıç.....
(Thesis Supervisor)

Assoc. Prof. Dr. ..... Abdullah Daşcı.....

Asst. Prof. Dr. .....Hilal Kazan.....

DATE OF APPROVAL: 10/Aug/2016

ABSTRACT

Predicting Drug Interaction Type Using Machine Learning

MILAD HASSANI

IE, M.Sc. Dissertation, August, 2016

Thesis Supervisor: Asst. Prof. Dr. Kemal Kılıç

Keywords: Data Mining, algorithms, drug interaction, synergy, antibiotics, antibiotic resistance

Antibiotic resistance has become an important health threat across the world during recent years. One of the solutions to reduce antibiotic resistance is to find ways in order to use efficient amounts of antibiotics in treatments. It has been seen that some antibiotics are synergistic, i.e, if they are administered together, they will boost the individual antibacterial and antifungal effects. Identification of synergistic antibiotics can be of significant assistance to medical practitioners in order to optimize the amount of antibiotics to be used. In this thesis we have conducted a set of analyses using data mining based approaches. Chemogenomic profiles and chemical properties of drugs have been utilized to predict synergy between them. Two datasets, E. Coli and yeast were used in order to perform the analysis. GRASP meta-heuristic algorithm was implemented on chemogenomic features in order to predict synregies which yielded in *0.94* accuracy and *0.82* Area Under ROC curve for E. Coli dataset. In order to further explore the chemogenomic features, we suggest a novel algorithm to predict synergy. This algorithm resulted in Area Under ROC curve and accuracy of *0.71* and *0.91*, respectively for E. Coli dataset. Next, two chemical features, *XLogP3* and *Q_PC-* were used to perform the analysis by employing decision trees and random forest classifiers. Our analysis indicate that *Q_PC-* chemical feature can be as discriminative as *XLogP3* which has been used in literature previously. Employing chemical features resulted in most accurate prediction among the implemented methods. In this thesis, details of the above-stated methods and algorithms will be presented.

ÖZET

MILAD HASSANI

Endüstri mühendisliği, Master Tezi, Ağostos, 2016

Tez Danışmanı: Doç. Dr. Kemal Kılıç

Son yıllarda antibiyotik direnci dünya çapında önemli bir sağlık tehdidi sırasında haline gelmiştir. Antibiyotik direncini azaltmak için çözümlerden biri tedavi sırasında antibiyotiklerin uygun miktarda kullanılmasının sağlanmasıdır. Bazı antibiyotiklerin sinerji oldukları, yani birlikte uygulandığı takdirde, tek başlarına sahip oldukları antibakteriyel ve antifungal etkilerin arttığı bilinmektedir. Antibiyotikler arasındaki sinerjinin belirlenmesi, çeşitli durumlarda kullanılacak olan antibiyotik miktarlarının optimizasyonuna katkı sağlayabileceğinden, tıbbi uygulamacılar için önemli bir fayda sağlayacaktır. Bu tezde veri madenciliği tabanlı yaklaşımlar kullanılarak antibiyotikler arası sinerjinin belirlenmesi problemine yönelik bir takım analizler yapılmıştır. Bu kapsamda ilaçların chemogenomic profilleri ve kimyasal özellikleri kullanılarak birbirleri arasındaki sinerjinin tahmin etmeye yardımcı olacak yeni yöntem ve yaklaşımlar geliştirilmiştir. Geliştirilen yöntem ve yaklaşımlar E. Coli ve maya verileri kullanılarak performansları karşılaştırılmıştır. Chemogenomic profillerin kullanılmasına dayanan yaklaşımlar arasında yer alan GRASP meta-sezgisel algoritması 0.94 doğruluk ve 0.82 AUC sonucuyla E.Coli veri kümesinde en iyi sonucu vermiştir. Öte yandan topoloji verilerinin de öğrenme sürecinde kullanılması amacıyla geliştirilen orijinal bir yaklaşımla gene E. Coli veri kümesinde sırasıyla 0.91 ve 0.71, doğruluk ve AUC sonucuna ulaşılmıştır. Öte yandan tez kapsamında ayrıca ilaçların kimyasal özellikleri verisinin kullanılması durumunda ne yapılabileceği üzerinde durulmuştur. Bu kapsamda, karar ağaçları ve karar ormanları gibi yöntemlerin de kullanılmasıyla yapılan çeşitli çalışmalar sonucunda iki özelliğin (XLogP3 ve Q_PC) sinerji ilişkisinin belirlenmesi kapsamında diğer 300 civarındaki kimyasal özelliğe göre daha çok yardımcı olabileceği belirlenmiştir. Her ne kadar XLogP3 bilimsel yazında zaten bu kapsamda etkili bir özellik olarak daha önceden belirlenmiş ise de, yapılmış olan analizler Q_PC'nin de XLogP3 kadar işe yarar olabileceğini göstermektedir. Bu tez de bu kapsamda geliştirilen yaklaşımlar, yöntemler ve bunların performanslarının belirlenmesine yönelik yapılan çeşitli analizlerin sonuçları sunulmaktadır.

Anahtar Kelimeler: Veri Madenciliği, Algoritmalar, İlaç Etkileşimi, Sinerjik İlişkiler, Antibiyotikler, Antibiyotik Direnci

Acknowledgements

I would like to thank

My advisers Dr. Kemal Kılıç for his support. He has been a role model for me both in personal and academic life.

Sabanci University Academic Support Program for funding my graduate education for two years.

My family for thier unending support from the beginning of my life.

Milad Hassani

Contents

List of Figures

List of Tables

Abbreviations

AUROC	**A**rea **U**nder **R**eceiver **O**perator **C**urve
AUPR	**A**rea **U**nder **P**recision **R**ecall
DM	**D**ata **M**ining
E. Coli	**E**scherichia **C**oli
GA	**G**enetic **A**lgorithm
GRASP	**G**reedy **R**andomized **A**daptive **S**earch **P**rocedure
KNN	**K**-**N**earest **N**eighbour
ML	**M**achine **L**earning
PCA	**P**rincipal **C**omponent **A**nalysis

Chapter 1

INTRODUCTION

1.1 Motivation

Humankind has always searched for solutions to the problems that takes place in his/her body. Early written evidence of such endeavors date 2500 BC where certain types of plants have been used to cure diseases in Egyptian civilization [1]. Each medical plant's usefulness to treat a category of illnesses had been identified by medical practitioners of this ancient civilization. Researchers in [1] suggest that these discoveries had been based on the physical properties of the plants. For example, Figure 1.1 depicts perforated leaves of St John's wort which suggests a benefit in healing perforated wounds. Similarly, in Ayruvedic medical tradition in India, Azadirachta Indica or neem leaves were used as a natural remedy for bacterial infections, antiseptic treatment, oral hygiene, parasite infection, fevers, and general infections [4]. Although traditional medical systems were effective in treating certaint types of infections and boosting general immune system, they were not able to provide solutions for all infections especially with global prevalence of infectious diseases in 18th and 19th century.

FIGURE 1.1: Ancient societies used St. John's wort to heal perforated wounds based on the shape of its leaves [1].

In 1928, Alexander Flemming, introduced the first chemical compound, Penicillin, which fights the bacteria and/or inhibits its growth. This discovery was effective in controlling infectious diseases which in turn led to a revolution in the medical sciences in 20th century. The term *antibioisis* means *against life* and was first introduced in a research paper published by William Roberts in 1874. Roberts observed and reported that the liquid in which the *Penicillium glaucum* existed could not be infected with bacteria [5].

Although antibiotics are not proven to be effective in battling diseases which are caused by viruses, bacteria based disorders can be effectively cured using this type of drugs. Infectious diseases such as tuberculosis, which was the cause for 25% of deaths during 19th century, have been prevented in developed countries [6].

Since the first introduction of antibiotics in medicine many new infections have been discovered and accordingly scientists have been constantly trying to find new antibiotics to defeat them. At the same time many lives have been saved and the use of antibiotics has significantly improved public health globally.

Antibiotic drug consumption has increased by 36% during 2000-2010 period. This increase has happened mainly in developing countries [7]. Although drug accessibility is a life saving opportunity for the individuals who were previously unable to provide these drugs for themselves, over-consumption of antibiotics results in antibiotic resistance which is currently considered as a major public health threat [7]. That is to say, antibiotic resistance is a serious problem facing global health that is caused by misuse and overuse of antibiotics.

When antibiotic resistance occurs, a microbe evolves such that it resists the chemical compounds which aimed to kill it. This resistance can be partial or total. In total resistance, presence of the drug makes no difference in the growth of the bacteria and in partial resistance, the drug kills only a percentage of the bacteria.

In order to prevent resistance issue, promoting rational use of antibiotics and educating communities about the effects and dangers of antibiotic resistance seems to be essential. Studying efficacy of drugs in cases where multiple number of antibiotics are required is one of the ways to help physicians optimize the amount and type of the antibiotic administered.

Researchers have previously studied the effects of using two drugs at the same time and the process is referred to as drug-drug interaction. Many of such research is based on purely experimental procedures, in which bacteria is grown in laboratory and a combination of drugs are introduced to it in order to observe the combined effects of drugs on bacteria's growth. However, these methods are expensive and time consuming, which limits the number of interactions that can be examined.

Employing data-mining (DM) can significantly assist in exploring drug-drug interactions. By applying these techniques it is possible to discover and study a greater number of drug-drug interactions. Although experimental data is needed in training phase of the DM procedures, implementing these methods brings the possibility of predicting drug interactions for which experimental data does not exist. That is to say, researchers can learn from existing data and predict those that are not available by using tools provided in DM. Despite the fact that there are several articles in literature that incorporate Machine Learning (ML) in this problem, there is still open research questions both in terms of the methodologies and their applications.

In this thesis, we have examined various approaches that employs DM techniques in order to predict the synergistic drug-drug interactions. Performance of the developed nethodologies are assessed in two datasets, namely, E. Coli and yeast. The rest of the thesis is organized as follows. In Chapter 2 we will introduce terminology and define the problem. Relevant literature will also be provided in that chapter. Next, one of the approaches that utilize topological information will be presented in Chapter 3 and the performance of the proposed approach will be evaluated in E. Coli and yeast datasets. Chapter 4 will follow the same structural flow in which non-topological approaches will be introduced and evaluated with the same datasets. In Chapter 5, we will introduce chemical features to the problem and present our proposed approaches. We will conclude the thesis in Chapter 6 with some insights we gained from the experimental analysis and further research topics that are worth to be examined.

Chapter 2

Problem Statement and Relevant Literature

In this chapter we will first introduce the basic terminology and the relevant literature of the synergistic drug-drug interaction prediction problem that we will focus in the thesis. Next, we will provide the details of the datasets that will be used in the analysis.

2.1 Drug-drug Interaction

The history of studying interactions among drugs can be traced back to traditional Chinese medicine and Indian medicine, Ayurveda. In traditional systems of medicine, practitioners combined herbal extracts and plants in order to increase the effects on a particular disease. These methods were developed by observing each individual empirically and proposing a solution based on the patient's particular characteristics [8]. One of the most popular fixed formulas in Ayurveda is "*Trikatu*." This mixture combines black pepper (*Piper Nigrum*), long pepper (*Piper Longum*), and ginger (*Zingiber Officinalis*) [9][10]. It has attracted the attention of phyto-medicine practitioners recently in an experiment and it is discovered that "*Trikatu*" increases the blood level of vasicine significantly, which is an antiasthmatic drug [10]. It appears that *Trikatu* mixture increases bio-availability of drugs by providing rapid absorption in garstointestinal tract or prevention of drug oxidization in its first passage through liver [10].

One of the disadvantages of using traditional systems of medicine is that geographical location and seasonality of the place of production for the medical plant is significantly influenced in their efficacy [8]. For example, researchers compared chemical compunds in two extracts of Brazilian Orchid Tree (*Bauhinia forficata*) from two different geographical areas and harvested in the same period of the year, and the results suggest that the extracts present different concentration of their main marker [8]. Despite their efficacy in dealing with chronic ailments, traditional medicine has not been proven to be successful in combating new infectious diseases and as a consequence, they are rarely practiced.

Drug interaction is investigated from two major points of view. Some try to investigate drug interaction type using pure biological experiments while some use computational methods and bioinformatics for prediction.

In pure biological studies each drug combination is applied to cultures of the biological entity and growth curves over a time period are reported. Growth conditions consist of multiple dosages for each one of the drugs. Growth curves indicate the number of grown cells at any specific time. Drug interaction type is then determined based on these growth graphs. E. Coli is one of the most researched bacterium in the literature. Some groups have also done experiments on yeast cultures [2].

As possible combination of drugs is extremely large, determining interaction types using pure biological methods for every combination is not possible in reality. In addition, sometimes experiments are repeated several times to prove their reproducibility, which is both expensive and time consuming. Computational methods bring the possibility of predicting drug interaction type using a reduced number of experiments. Using computational methods, the type of new drug interactions is predicted based on features extracted from the biological experiments.
Each of the above-stated ways to approach drug synergy prediction can be used for finding synergy in different types of cells.

Here we will introduce the basic terminology and review previous studies done in the field of drug synergy prediction in literature.

2.1.1 Measuring synergy

When two drugs are administered together they may interact in three different ways [11]:

- They might have expected efficacy of each drug as in the case that they are administered individually, which are called as *additive* drugs.
- They may have an efficacy higher than what it is expected from each one of the drugs; in this case these two drugs are said to be *synergistic*.
- One may inhibit the other one to function, thus, using two drugs together have less efficiency as using each one of them separately, which are called *antagonistic* drugs.

These three possible ways of interaction are depicted in Figure 2.1. X axis represents ratio between the dose of agent 1 when used in combination with agent 2, D_1, to the dose of agent 1 in isolation, $DI_{x,1}$, for the same biological effect. Axis Y represents the same ratio for agent 2 [8]. Concave line, i.e. when the curve moves towards the origin, indicates that the agents in the mixture are synergic, and when the opposite occurs (convex line) they present antagonism. In other words, the same biological effects of the agents in isolation are obtained at lower (or higher) doses of the mixture.

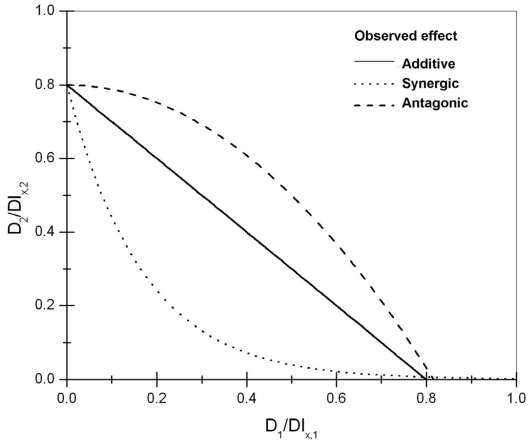

FIGURE 2.1: Drug interaction types of combined agents.

Synergistic interaction is of more interest as it allows treatment of the disease with less dosage of drugs. Administration of synergistic drugs results in reduction in

toxicity and side effects while the efficacy is increased or remains the same [11]. Combinatorial therapies that impact multiple targets inside the cell simultaneously are less prone to development of drug resistance, and thus they increase therapeutic efficacy [12]. Although in this thesis we study using a combinatorial therapeutic approach in treating infectious diseases, they are now standard in tackling other diseases as well, such as cancer, diabetes [12], AIDS, malaria and hypertension [8].

There are four main mechanisms that drug synergy might occur [2].

- Drugs impact separate targets to create a combination effect.
- One drug alters the ability of another to reach its target.
- The components bind separate sites on the same target to create a combination effect.
- Two drugs physically interact to make a new chemical entity.

In order to discover how two drugs interact, a target cell is cultured in a matrix which is called *dose matrix*, i.e., growth matrix [3]. Each plate has a different concentration of drug 1 and 2. For example, in Figure 2.2 concentrations of drug A and B are increased in eight steps and combination of both has been applied to the culture of cells. Each subplot in Figure 2.2 represents number of cells present in culture over a period. X axis in the subplots represent time and Y axis represents number of cells in culture. The type of interaction between the drugs are determined by the shape of Minimum Inhibitory Concentration (MIC) curve obtained from the experiments.

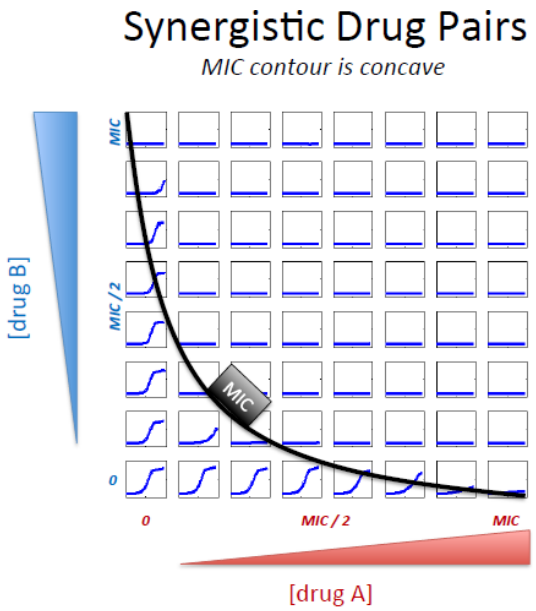

FIGURE 2.2: Drug interaction diagram when drug A and drug B have synergy[2]

MIC, is the *lowest* drug concentration that prevents visible growth of the target microorganism. In synergistic interaction connecting MIC concentrations in the dose matrix, results in a concave line.

When drug A and B have synergy, they kill the bacteria faster, in other words, in the first column of graphs in Figure 2.2, when there is no drug A in the solution, it takes seven units of drug B to kill the bacteria completely, but by introducing just one unit of drug B to the solution, it takes just 2 units of drug A to kill the bacteria completely. This method of representing drug interaction is called Loewe additivity model [13].

Figure 2.3 represents an example of real dose-matrix responses obtained in an experiment. Tunicamycin and Ciclosporine A are suggested to be synergistic in attacking *Saccharomyces cerevisiae* [3]. Each circle in Figure 2.3 depicts a colony of grown cells. Brighter spots implies more cells have survived to grow in comparison to darker spots. Since the MIC curve is concave, Tunicamycin and Cyclosporin A are synergistic.

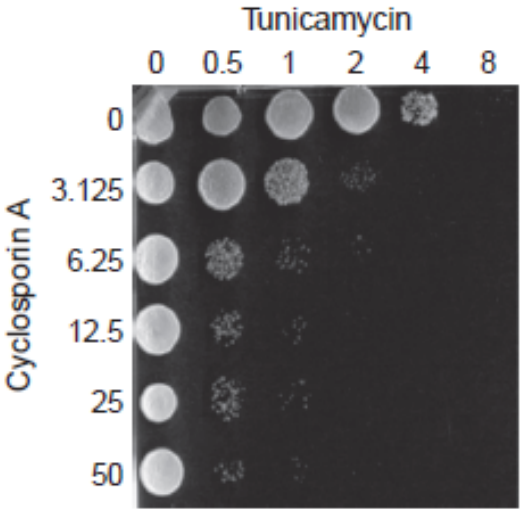

FIGURE 2.3: Dose matrix response of two antibiotic drugs which show synergy [3]

On the other hand when drugs are additive, they neither boost nor prohibit their effects, i.e., adding one more units of drug A to the growth environment kills the same number of bacteria cells as adding one unit of drug B. Loewe additivity model for additive drugs is represented in Figure 2.4.

By means of such experimental analysis, it is possible to identify the type of interaction of drugs in different target bacteria. However, it is costly due to the time required and money spent. Data-mining methods, on the other hand, use known interactions to predict unknowns which saves both in terms of money and time. In the literature, some research is available that uses chemogenomic profiles while some use chemical features as an alternative. Now we will introduce the literature that utilizes these two approaches.

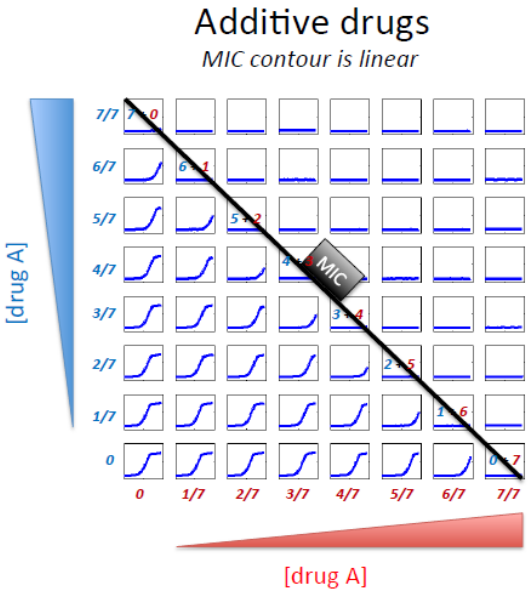

FIGURE 2.4: Loewe additivity model when drug A and drug B are additive [2]

2.1.2 Chemogenomic profiling

Researchers have developed different methods to predict whether two drugs are synergistic or not. One of these methods is to employ chemogenomic profiles of the drugs that interact. Chemogenomic profiles are an output of a research discipline known as chemogenomics. In this field, researchers find genomic response of a biological entity to chemical compounds [14]. When a chemical compound is introduced to a cell or a biological entity, it will target products of one, or more gene(s) which in turn inhibits that gene from expression. In other words, with chemogenomic profiling it is possible to find which gene is affected by a specific drug [14].

In order to access the chemogenomic profile of a drug, a gene is removed from bacterium's genome and a *stress* is introduced to that phenotype, then that phenotype is arrayed on agar plates[1]. Number of bacteria cells which are present on the agar plate after applying the stress over a period of time is estimated. Based on this estimation, which is done by multiplying size of the colony by approximate number of cells per unit of area for E. Coli, a score is calculated. This score is proportional to expression level of the cells which lack the related gene. This score

[1] An agar plate is a dish which contains a growth medium and is used to culture microorganisms. Other compounds such as antibiotics might also be added to the plate [15].

is calculated for all gene deletions. The combined vector is the chemogenomic profile of that particular type of cells.

Chemogenomic profiles have been used in drug synergy prediction in [3]. Authors use hyperergeometric probability distribution and calculated a similarity measure between chemogenomic profiles of two drugs involved in the interaction. If there are x genes that are common in both profiles, similarity measure is the probability of obtaining x or more genes that overlap between the two by chance. Similarity measure can be considered as a p-value derived from hyper-geometric distribution [3].

Note that, the approaches used in [3] does not utilize the topological features of the drug-drug interaction network. In this thesis we will also develop an algorithm which utilizes the topology information. Furthermore, we will introduce alternative approaches that use genetic algorithms (GA) and Greedy Randomized Adaptive Search Procedure (GRASP) which will not use topology information. The algorithms that are based on the former approach will be presented in Chapter 3 and the algorithms that are based on the other approach will be introduced in Chapter 4.

2.1.3 Utilizing Chemical Features

Another approach in predicting drug synergy by data-mining methods is utilizing the chemical features of the drugs. In [2] authors have found that synergy of drugs based on their lipophilicity (LogP) values. LogP of a drug is by means of the octanolwater partition coefficient which is the relative solubility of a compound in octanol over the solubility in water.

$$logP = log(\frac{solute\ in\ octanol}{solute\ in\ water}) \tag{2.1}$$

A major advantage of this method is that logP of all compounds are known with high precision and at no cost.

In [2] authors have defined *synergicity* of a drug as the ratio of number of synergies of that drug with other drugs to number of experiments done on that drug. For example, out of 24 drugs that Pantamidine was tested against, 12 of them were synergistic and 12 non-synergistic which sets synergicity score of Pentamidine as

0.5. Synergicity of a drug can be considered as a measure of willingness of a drug to have synergy with others in a particular drug set.

Figure 2.5 (left) indicates that lipophilicity (XLogP3) is significantly correlated with synergicity (Spearman correlation = 0.51, p-value = 0.0036). Diameter of each circle corresponds with number of experiments done on that drug. Figure 2.5 (right) is histograms of XLogP3 distribution for non-synergistic (black) and synergistic (red) pairs. Two drugs with highest number of experiments (Pentamidine and Terbinafine) are more likely to have synergy with lipophilic drugs [2].

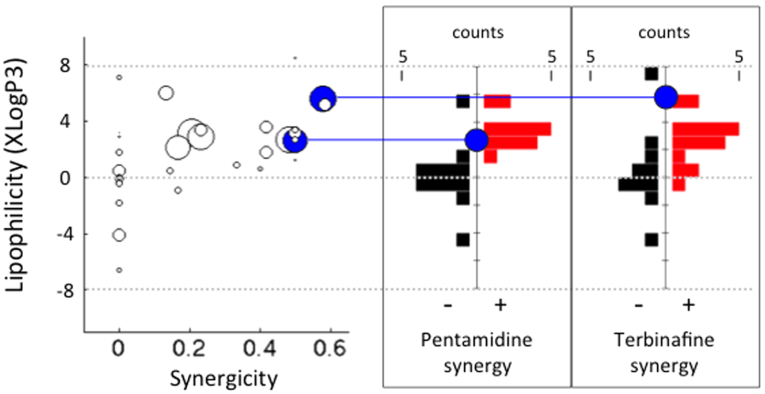

FIGURE 2.5: (Left) Drug synergicity scores vs. drug lipophilicity for 31 drugs. (right) is histograms of XLogP3 distribution for non-synergistic and synergistic pairs

Since correlation between XLogP and synergicity is positive, it can be inferred that if a drug has a higher lipophilicity it is more probable to have synergy with other drugs, therefore, XlogP can be a good measure that cn be used in predicting synergy among drugs. We can also observe in Figure 2.5 (right) that the drugs with highest number of synergy experiments tend to show synergy with drugs that have higher XLogP values which repeats the previous assumption.

Using chemogenomic dataset two approaches are developed; methods based on topological data and non-topological methods. In experimental analysis that are conducted both E. Coli and yeast are used. Next, we will summarize the datasets that are used in the analysis.

2.2 Datasets

When two drugs are used together three different possibilities arise. They might be antagonistic, additive or synergistic. In this thesis we aim to predict synergy of the drugs and we do not consider predicting neither additive nor antagonistic interactions. We approach the problem using two different different approaches and propose various algorithms for both. One of the approaches utilizes chemogenomic profiles of the drugs involved in an interaction and the other uses chemical properties of the drugs for the same purpose. We have used chemogenomic profiles that have been discovered for E. Coli [17] and yeast [18] respectively.

In each of these researches, a series of experiments have been conducted on E. Coli and yeast using different drugs in order to determine their chemogenomic profiles. For each drug, different dosages have been applied and *drug-gene* scores have been recorded in the relative chemogenomic profiles.

In order to select one of available profiles for a drug, the profile which has the highest number of expressed genes was selected. An *expressed gene* is the one which its drug-gene score is higher than a threshold in the profile, i.e. number of grown cells which lack the gene passes a threshold when that drug has been present in the cell culture. That is to say, by means of a threshold, the levels (scores) are binarized. The threshold that is used to binarize profiles is set by the domain experts.

Bacteria's growth is prohibited not only by introducing drugs to it but also by putting it under various physiologically relevant stresses. As we are going to analyze the effects of drugs on E. Coli bacteria, we only consider chemogenomic profiles when only a drug used as a stress factor.

Recall that MIC contour was used in order to determine the interaction type between the drugs. The MIC curve is summarized by a metric refered to as *alpha score* [20]. That is to say, alpha scores are extracted from dose matrices, explained in Section 2.1, to indicate number of grown cells of the target cells (E. Coli or yeast) when two drugs have been present in the culture. A low alpha score in an interaction implies low number of grown cells, thus, we conclude that involved drugs have effectively prohibited the growth of target cells, similarly, if the alpha score is more than a threshold, which is again provided by experts, we can label

the interaction as antagonistic. If the alpha score of an interaction is lower than the threshold we consider the interaction as synergistic.

2.2.1 E. Coli Dataset

In [17] authors have compiled chemogenomic profiles of E. Coli in 324 conditions for 3979 gene deletions extracted from Keio single-gene deletion library [19]. Out of 324 conditions, 209 belongs to drugs including antibiotic, antimicrobial and antifungal drugs [17].

Conditions to which gene deletions were introduced consists of 114 unique conditions in different varietes, e.g. a drug is used as an stress in different dosages or other physiological stresses have been applied.

The dataset of E. Coli contains $\binom{25}{2} = 300$ alpha scores of interactions between 25 drugs. In order to be able to validate interaction types which are predicted, a subset of drugs are selected which we both have validation data and chemogenomic profiles. After removing drugs that were not common between train and validation sets, 19 drugs and $\binom{19}{2} = 171$ interactions among them were remained. Validation dataset was extracted from [20].

2.2.2 Yeast Dataset

Chemogenomic profiles of yeast [18] were used in the analysis. In this dataset 726 conditions were applied on yeast's 5985 gene deletions to find its chemogenomic profiles. Interaction data available in supplementary materials of [2] were used to compile required dataset for yeast. As the interaction data was in binary numbers, i.e, synergy is represented by *true* and *false* otherwise, we used the same data and did not apply a threshold on alpha scores. This dataset contains 175 interactions among 33 drugs. Two different subsets were extracted to be used in validation phase of the proposed algorithms. The first contains 165 pairs of 31 drugs and the second includes 67 pairs of 21 drugs. Validation sets were selected according to the type of features used for training. First set was used to validate the methods based on chemical features, that will be discussed in Section 2.2.3, and the second is used for the chemogenomic profile based prediction approaches.

As chemical property that we use in proposed methods for yeast does not exist for two drugs, i.e. *Lithium* and *Cisplatin*, we removed these two drugs and their relevant interactions leaving the validation set with 31 drugs and 165 interactions among them. Chemogenomic profiles of 22 drugs, with 68 interactions, exist in yeast's validation set. One drug in the dataset, *Clozapine*, has only one interaction. We removed that interaction and drug in order to increase reliability of the methods.

2.2.3 Chemical features

In the second method of drug interaction type prediction, chemical properties of drugs are used in order to predict their interaction type. Our dataset consists of 328 chemical features for 31 drugs and 165 drug interactions. In this dataset interaction data is not complete and instead of $\binom{31}{2} = 465$ interactions, we have a subset of 165. Our aim in this part is to investigate the possibility of predicting synergy type using just one of the chemical features of the drugs in that interaction. The advantage of using this method instead of using experimentally obtained features is that chemical features are calculated mathematically with a predefined formula which makes them significantly more reliable in comparison to the features based on experiments. This fact helps us remove uncertainty caused by the experimental procedure and improve the accuracy of predictions.

A set of analysis that had been done using XLogP3 has been published in [2]. Here we first aim at reproducing the results found in [2] and secondly finding features which did not have correlation with XLogP3 and result in reasonable Area Under ROC and Area Under Precision Recall curve.

Chapter 3

Topology-Based Algorithm

3.1 Introduction

In this chapter a novel algorithm is proposed in order to predict drug synergy based on synergy networks. A synergy network is a graph that depicts synergy relationships among a set of drugs, i.e., nodes of the graph represent drugs, and existence of an edge indicates that connected nodes are synergistic. Similarly, antagonistic network can also be developed by considering edges when connected drugs are antagonistic.

Synergistic and antagonistic pairs are determined by setting a threshold, extracted from expert knowledge, on the alpha scores. If the score is less than synergistic threshold we consider that interaction to be synergistic knowing the fact that when alpha score is low, two drugs have effectively kill target cells (E. Coli or yeast). If alpha score is higher than antagonistic threshold the pair is assumed to be antagonistic. Synergy and antagony networks contain valuable information that can be used to predict synergy. If a subset of drugs, i.e. a sub-graph, in the network are fully connected and a common property can be found among them, employing these properties may help us predicting synergistic pairs.

In this chapter we have utilized an algorithm to predict synergy using these common features. The following sections will provide details of this algorithm.

TABLE 3.1: Binarization thresholds used in validation sets.

	Antagony network	Synergy network	# of synergies
E. Coli	3	-0.5	20
Yeast	1	-0.5	17

TABLE 3.2: Maximal cliques in E. Coli synergy network and their nodes

Clique #	1	2	3	4	5	6	7	8	9	10	11
Nodes	1,7	2,7	7,14	3,5,7,16	3,6,7,16	7,13,16	7,16,19	8,16	10,11	13,18	18,19

3.2 Synergy and Antagony Networks

Synergy and antagony networks, as described in introduction section, depicts the topology in which drugs interact. Thresholds used for binarizing alpha scores are presented in Table 3.1 for the two datasets used in in the analysis.

Synergy and antagony networks of E. Coli dataset are represented in Figures 3.1 and 3.2 respectively. Recall that the existing dataset had 19 drugs for E. Coli, on the other hand, chemogenomic profiles of 21 drugs are available for the second yeast dataset. The synergistic and antagonistic networks of yeast are depicted in Figures 3.3 and 3.4 respectively.

The proposed topology based algorithm utilizes maximal cliques that are available in the networks. Maximal cliques are complete sub-graphs which form a complete connectivity inside the subset. In other words, cliques are complete sub-graphs of a network.

Bron-Kerbosch algorithm [21] is used to determine all maximal cliques in both synergy and antagony graphs of E. Coli and yeast.

E. Coli and yeast synergy networks consist of 11 maximal cliques listed in Tables 3.2 and 3.4 respectively. Note that we used Leave One Out strategy in the analysis. According to this strategy every time a drug is removed from the dataset and used as the test drug in order to evaluate the performance of the algorithm, the structure of the graph will change, which consequently modifies number of maximal cliques. For example if drug #7 is used as test drug in E. Coli dataset, there will be eight cliques which are represented in Table 3.3.

TABLE 3.3: Maximal cliques in E. Coli dataset when drug 7 is left out as test drug.

Clique #	1	2	3	4	5	6	7	8
Nodes	10,11	3,5,16	3,6,16	8,16	13,16	16,19	13,18	18,19

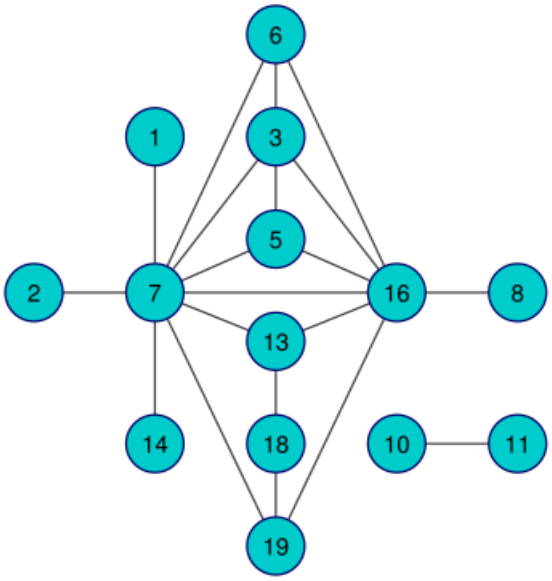

FIGURE 3.1: Synergy network for E. Coli dataset. Connected nodes represent synergistic pairs.

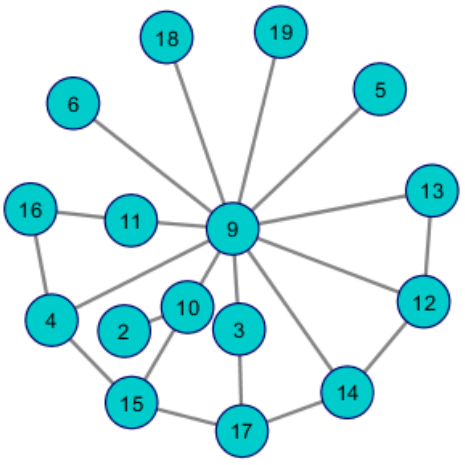

FIGURE 3.2: Antagony network for E. Coli dataset. Connected nodes represent antagonistic pairs.

Yeast's synergy and antagony network is provided in 3.3 and 3.4

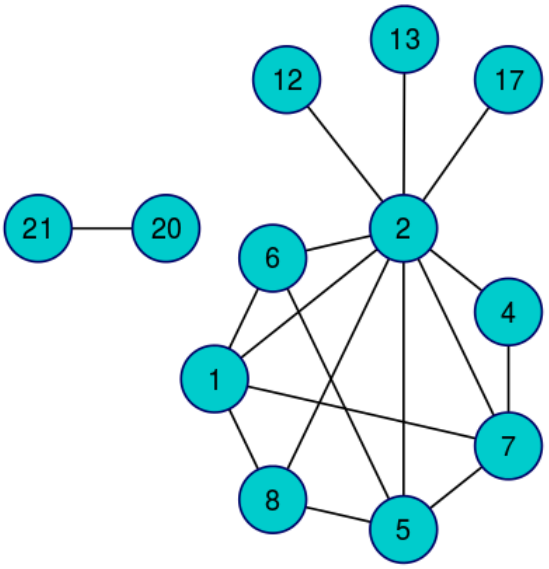

FIGURE 3.3: Synergy network for yeast dataset. Connected nodes represent synergistic pairs.

TABLE 3.4: Maximal cliques in yeast synergy network and corresponding nodes

Clique #	1	2	3	4	5	6	7	8	9	10	11
Nodes	1,2,6	1,2,7	1,2,8	2,4,7	2,5,6	2,5,7	2,5,8	2,12	2,13	2,17	20,21

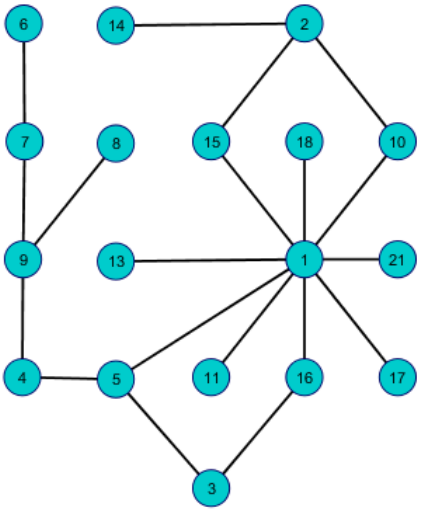

FIGURE 3.4: Antagony network for yeast dataset. Connected nodes represent antagonistic pairs.

3.3 Topological Algorithm

Our proposed method is structured in two phases. In the first phase we make a set of possible synergistic pairs which we call *"suggested synergies"* and in the second

phase, suggested synergies are filtered to remove less probable predictions. The following sections present the steps of each phase in the algorithm.

3.3.1 Finding Possible Synergies

A threshold, extracted from expert knowledge, was used to convert drug chemogenomic profiles into binary vectors. In the converting process, if a drug-gene score is over this defined threshold, binary value will be *true* and *false* otherwise. Threshold of -0.5 was used for E. Coli and -1 for yeast. Leave-One-Out cross validation concept was performed by leaving one drug and its relevant data as test data and using other drugs as training set.

Binary chemogenomic profiles were used as an input for Algorithm 1. D_{t_i} and D_C parameters used in Algorithm 1 are explained in Section 3.3.1.3.

Data: binarized chemogenomic profiles
Result: *"suggested synergies"* set
for *all drugs in dataset* **do**
    select drug as "*test drug*";
    remove test drug from the network;
    find maximal cliques;
    for *each maximal clique* **do**
        extract *discriminative features* set;
        calculate distance matrix;
    end
    if $D_{t_i} < D_C$ **then**
        add i_{th} drug in clique and test drug to *"suggested synergies"*;
    end
end

Algorithm 1: Topological based synergy prediction - creating *suggested synergies* set

3.3.1.1 Discriminative Features

A subset of features(gene deletions) were selected for each clique which we refer to as "*discriminative features*". This subset is used in order to calculate distance matrices. Recall that features have binary values after the binarization process.

Three possible cases exist for a feature value inside and outside of a clique,

- A feature value can be *true* in clique drugs' profiles and *false* for others.
- It can be *false* for clique drugs and *true* for others.
- Feature's value can be the same in clique and non-clique drugs, i.e. both *true* or both *false*.

Features which fit in first and second cases are possibly discriminative in finding synergies. Based on this observation we suggest the following: If a feature has the same value, e.g. *true*, among *all* clique drugs and is in lower s percentile of the other binary value, i.e. *false* we include them in "*discriminative features*". For example, value of feature f is *true* for both drug $d1$ and $d2$ which are members of clique c in E. Coli dataset and in only *one* of the remaining 17 drugs f value is *true*, therefore, we include this feature to discriminate features related to clique c. This process is repeated for each maximal clique as indicated in Algorithm 1.

As structure of network changes for each test drug, feature selection was done for each test drug network separately. We performed the same procedure for synergy and antagony networks of each dataset. Two cases were considered for both antagony and synergy network, therefore, total number of extracted feature subsets equals to $\sum_D N_{C_D} * 4$. Where N_{C_D} is number of cliques for corresponding drug D.

3.3.1.2 Distance Matrices

Distance matrices were calculated as a means to form classification. A *distance matrix* contains pairwise euclidean distances of drug chemogenomic profiles based on the subset of features determined. That is to say, only discriminative features are used to generate the distance matrices. For example when d_t is selected as test drug in E. Coli dataset, features selected for clique c earlier in the algorithm is represented as a vector which is a subset of drugs' chemogenomic profiles. This *sub-profile* is extracted for all 19 drugs and euclidean distance between each possible pair of these sub-profiles is calculated to form distance matrix of clique c. Distance matrices of all cliques are formed and stored as distance matrices for each test drug d_t.

Distance matrices are square with N_D rows and columns, where N_D indicates number of drugs in the dataset. Test drug is included in the process of calculating distances as there is a need to find distance between test drug and the drugs in the clique drugs. Inclusion of test drug in matrices does not bring circularity as distance values are not used in training part of the algorithm and is only used later in order to find synergies of the test drug.

We calculate a matrix for each member of discriminative features set, therefore, again a total of $\sum_D N_{C_D} * 4$ distance matrices exist.

3.3.1.3 Suggested Synergies

We first calculate average distance, D_C, between each clique drug and other non-clique drugs using the distance matrices of all cliques. Non-clique drugs set includes all drugs except the ones in the corresponding clique and the test drug.

If test drug's distance with any of clique drugs is less than D_C, we consider test drug and that specific clique drug as a possible synergistic pair. M will be an average of $N_D - N_{DC} - 1$ numbers where N_{DC} is number of drugs in clique.

Distance matrices of both cases discussed in section 3.3.1.2 were used to obtain a set of synergy suggestions. Using selected features, clique drugs have an average profile distance, D_C, among themselves and test drug has a distance, D_{t_i}, with i_{th} member of drugs in a clique. If $D_{t_i} < D_C$ we suggest that test drug and i_{th} clique drug are possibly synergistic and include them into *"suggested synergies"* set. After suggested synergies are dtermined, a classification algorithm is used to make the final decision. Next, we will present the details of the classification algorithm used for this purpose.

3.3.2 Classification

Data: suggested synergies
Result: predicted synergies
for *all drugs in dataset* **do**
    select subset W of *suggested synergies*;
    calculate scores for members of W;
    select high score members as final predictions
end
Algorithm 2: Topological based synergy prediction - predicting synergistic pairs

A subset, W, of most frequent suggestions for each test drug was extracted from *suggested synergies* set. Subset W has a size, p, which is an indicator of connectivity of a drug in the network,

$$p = \lfloor avg(synergicity) * (N_D - 1) \rfloor$$

largest integer before(floor) average synergycity score in the network multiplied by $N_D - 1$ which is number of drugs excluding test drug. . For example if drug #7 is selected as the test drug, average synergicity equals to 0.35 and $p = 6$.

Three scores were extracted from distance matrices to identify candidate synergistic drug(s) from W.

- *Synergicity* is number of synergies a candidate drug shows in all experiments that has been done on it.
- *Difference score for synergistic network*
- *Difference score for antagonistic network*

When two drugs are included in *suggested synergies* set,

$$\frac{D_C - D_t}{mean(distance\ matrix)}$$

is added to a separate vector which we call as *"difference vector"*. Average *"difference vector"* values of items in W were calculated to form *Difference score* for each member of W set. This vector was also created using distance matrices obtained from antagonistic network. A higher difference score in synergy network and a

lower score in antagony network, indicates that test drug is more affine to have synergy with the drug in W, thus, when combining these two scores, antagonistic scores will have a negative coefficient.

Synergy Index is a normalized weighted sum of three scores. Average synergy index of all drugs in W added to a percentage of their standard deviation was used as a threshold to determine final synergies. Percentage of used standard deviation is different in yeast and E. Coli datasets. In E. Coli dataset 100% and in yeast dataset 75% of of standard deviation was added to the average synergy index. Percentages of standard deviation was determined empirically.

3.4 Results for E. Coli

Employing this method on E. Coli dataset resulted in 566 members of *suggested synergy* set for synergy network and 759 for antagonistic network. Suggested synergies has 215 unique members for synergy network and 260 for the antagony network. Size of W set varies between 8-9 suggestions for this dataset. Synergy index was calculated using weights equal to 0.5 for synergicity and synergy difference scores and -0.5 for antagony difference score. Table 3.6 represents confusion matrix after applying the method on original network. A confusion matrix is a table that represents performance of an algorithm. Each column of the matrix indicates the instances in a predicted class while each row represents actual instances. Table 3.5 represents the elements of a confusion matrix.

- true positives (TP): These are the cases in which the classifier correctly predicts interaction to be synergistic, and it is synergistic in reality.
- true negatives (TN): These are the cases in which the classifier correctly predicts interaction to be non-synergistic, and it is non-synergistic in reality.
- false positives (FP): The classifier predicts synergy but the interaction is non-synergistic in reality.
- false negatives (FN): The classifier predicts non-synergy but the interaction is synergistic in reality.

TABLE 3.5: Structure of a confusion matrix.

	Predicted(P)	Predicted(N)
Actual(P)	TP	FN
Actual(N)	FP	TN

TABLE 3.6: Confusion matrix after network based classification on E. Coli dataset.

147	4
11	9

Base accuracy rate is defined as

$$Base\ accuracy\ rate = \frac{Size\ of\ major\ class}{Total\ number\ of\ instances}$$

Major class is the output class of the classifier which has the highest instances in data set. Base accuracy rate indicates accuracy value in case that the classifier outputs the major class for every input In this dataset 20/171 interactions belong to the positive, i.e., synergistic class which makes base accuracy rate equal to 20/171 Area Under ROC curve for this result and accuracy are 0.71 and 0.91, respectively, in which accuracy is more than base rate of 0.88.

As an alternative approach, the method used in [2] was implemented in order to verify the results. We conducted the analysis on node-shuffled and edge-shuffled networks for 1000 iterations. Edge shuffled network has the same number of nodes as the original network but the links between nodes are randomly distributed. Node-shuffled network on the other hand has the same topology of edges but node labels are shuffled. Since more information is lost in edge shuffled network we expect when we shuffle edges, performance of the classifier be lower than the node-shuffled network. At the same time original network's classification performance should be better than node-shuffled's as node label information is lost in this network. Figure 3.6 represents the results of 1000 randomizations for both node and edge-shuffled networks. AU-ROC of node shuffled was more than the original network in 42/1000 of the trials. As there is no randomization for the original network we performed the analysis once.

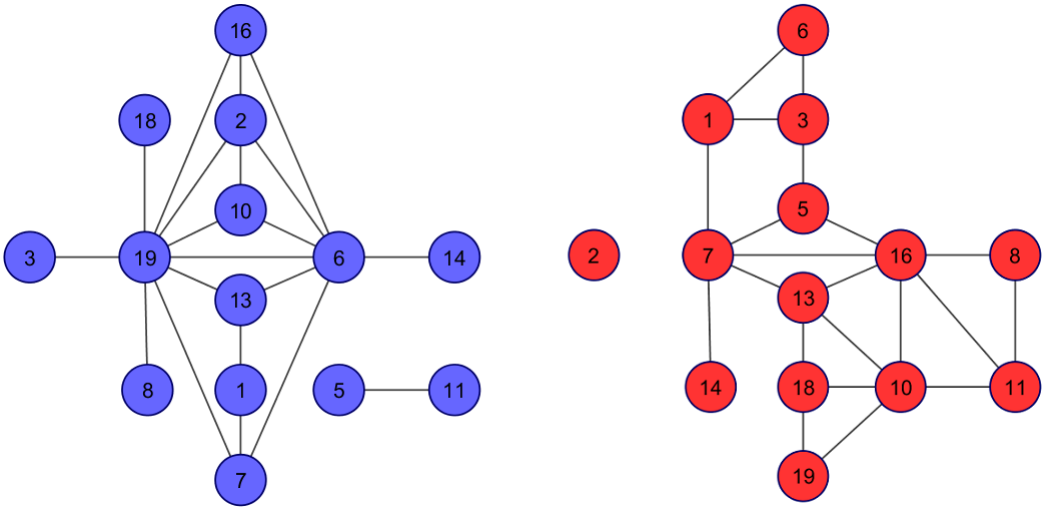

FIGURE 3.5: Area Under ROC curve for of 1000 iterations of the method on edge-shuffled (blue), node-shuffled (green) and original network(black). The real counts value for original graph is one.

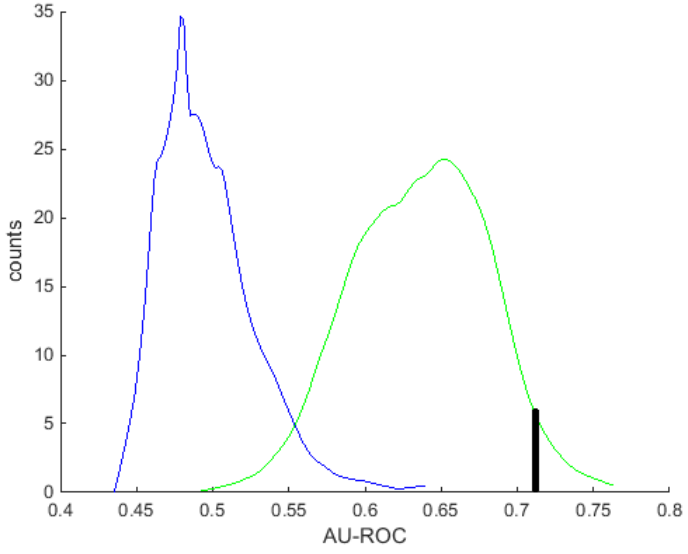

FIGURE 3.6: Area Under ROC curve for of 1000 iterations of the method on edge-shuffled (blue), node-shuffled (green) and original network(black). The real counts value for original graph is one.

3.5 Results for yeast

Employing this method on yeast dataset resulted in 556 members of *suggested synergy* set for synergy network and 773 for antagonistic network. Suggested synergies has 179 unique members for synergy network and 260 for the antagony network. Size of W set varies between 8-9 suggestions for this dataset, except for the time that drug # 2 is selected as test drug. Synergy index was calculated

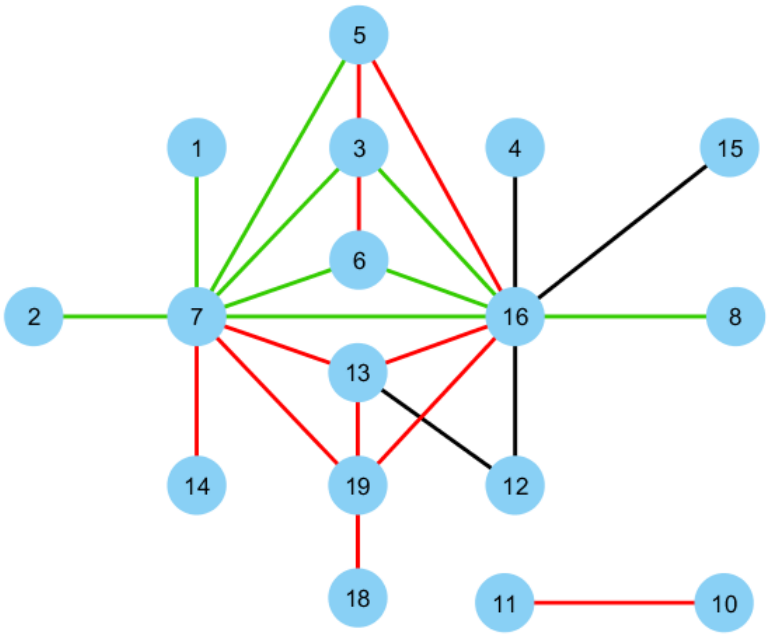

FIGURE 3.7: Predicted synergy network for E. Coli. Nine synergies were predicted correctly (green), 11 synergies could not be predicted (red), and four non-synergies were predicted as synergy (black)

TABLE 3.7: Confusion matrix after network based classification on yeast dataset.

43	7
9	8

using weights equal to 0.5 for synergicity and synergy difference scores and -0.5 for antagony difference score. Table 3.7 represents confusion matrix after applying the method on original network.

In this dataset 17/67 interactions belong to the positive, i.e., synergistic class which makes base accuracy rate equal to $\frac{50}{67} = 0.75$ Area Under ROC curve for this result and accuracy are 0.66 and 0.76, respectively, in which accuracy is more than base rate of 0.75.

Table 3.7 represents confusion matrix after applying this methon on yeast dataset. Figure 3.8 represents the results of 1000 randomizations for both node and edge-shuffled networks. AU-ROC of node shuffled was more than the original network in 581/1000 of the trials. As there is no randomization for the original network we performed the analysis once.

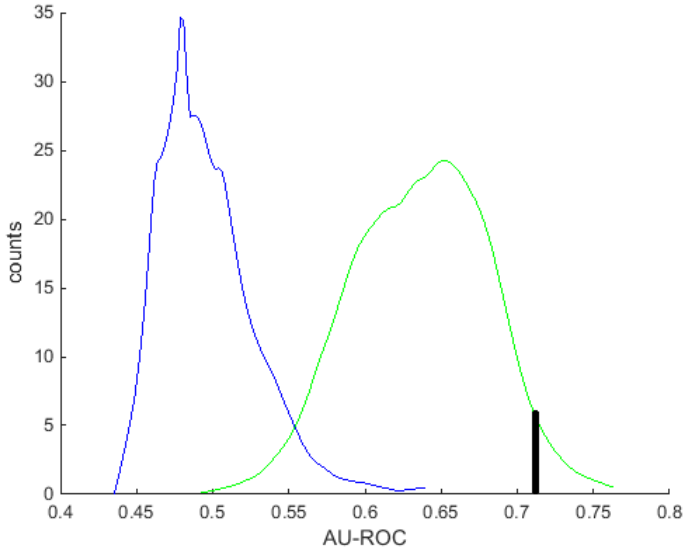

FIGURE 3.8: Area Under ROC curve for of 1000 iterations of the method on edge-shuffled (blue), node-shuffled (green) and original network(black). The real counts value for original graph is one.

Chapter 4

Non-topological Methods

4.1 Introduction

In this chapter of the thesis we have used two meta-heauristic algorithms for feature reduction and prediction of synergistic pairs. A brief review of the used algorithms will be provided in Section 4.2, results of employing such methods are represented in Section 4.3. Genetic Algorithm was only used to find a subset of available features which provide a better objective function. Details of the objective function will be provided in Section 4.2. GRASP algorithm was uswd only on E. Coli dataset. In order to predict interaction types using the feature subset found by GRASP, decision tree and KNN classifiers were used.

4.2 Algorithms

4.2.1 Genetic Algorithm

Genetic algorithms are a subset of a larger class of meta-heuristic algorithms called *evolutionary algorithms*. In these algprithm natural selection is inspired from biological evolution. Biological functions such as reproduction, mutation and selection is common in this type of methods. GA is especially beneficial in solving the problems in which the solution space is significantly large and fitness landscape is complex, i.e., there are many local optimal solutions. Mutation process in GA allows the algorithm to avoid remaining in the local optima and look for more or

global optimal solutions. A solution is usually called an *individual* and a subset of the solution space that algorithm preforms on is called a *population* or a *pool* of solutions. Fitness function is designed in such a way that represents the value of objective function of an individual. A crossover operator produces next population by incorporating more than one parent solutions. Children solution are referred to as *offsprings* in GA terminology. There are different types of performing a crossover on an individual. In this thesis we have used two-point crossover procedure in which children solution is created by two parts of first parent and one part from the second one, as depicted in Figure 4.1.

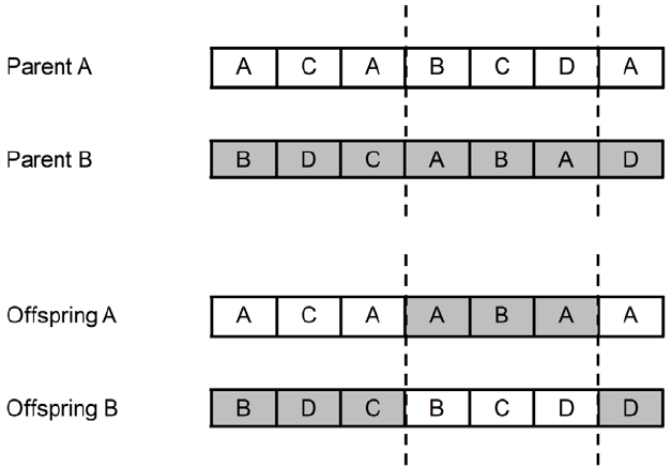

FIGURE 4.1: Two point crossover used in this analysis.

Locations of crossover are determined randomly in each iteration of GA. A mutation operator simulates the biological mutation phenomenon by randomly changing some parts of the offspring individuals, Figure 4.2 depicts a simple example of mutation in two points of a solution.

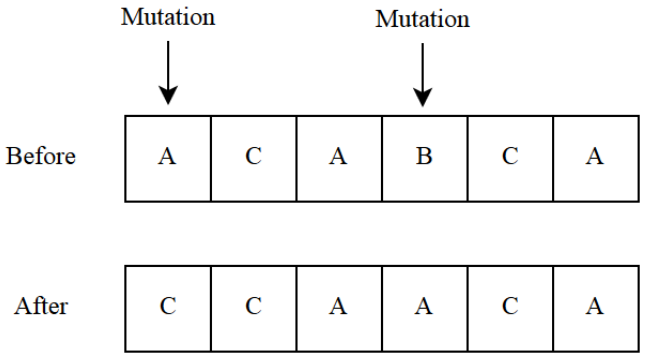

FIGURE 4.2: Mutation randomly changes some part of the solution.

We define selected subset of a chemogenomic profile which is defined by a solution as *"sub-profile"*. In this study, our objective function is defined as correlation between sub-profiles and alpha scores of the drug interactions in both E. Coli and yeast chemogenomic datasets.
A genetic algorithm with two-point crossover, one percent mutation was used in 10000 iterations. Probability of doing a crossover and a mutation was 0.5 in each iteration and when a mutation was going to happen 1% of solution pool was mutated. Algorithm 3 represents steps taken in this algorithm.

Data: chemogenomic profiles
Result: sub-profile
create initial pool;
for *number of iterations* **do**
    calculate fitness function for all pool solutions;
    find best and worst fitness solutions;
    crossover on best solutions $->$ new solutions;
    exchange worst reslts with new solutions ;
    mutate the pool ;
end

Algorithm 3: Genetic algorithm for sub-profile selection

4.2.2 GRASP

GRASP is a commonly used meta-heuristic algorithm in optimization problems. GRASP algorithm consists of two phases. First is to create a feasible solution, and the second phase in which local optimum is searched in neighborhood of the feasible solution found in the first phase. Each iteration of the algorithm provides a solution to the problem at hand. Since GRASP selects the best possible solutions, is categorized as a greedy algorithm. In every iteration of the algorithm a random solution of non-best solutions is added into a selected set of solutions with highest value of objective function. This set is then used for further steps of searching the

feasible region. A generic GRASP procedure is represented in Algorithm 4.

create initial solution;
for *GRASP stopping criterion is not satisfied* **do**
    construct greedy randomized solution ;
    local search;
    update solution ;
end
return best found solutions

Algorithm 4: Generic GRASP algorithm procedure [22].

In order to select a subset of features that can help us predict synergistic interactions, we have implemented GRASP on E. Coli and yeast datasets. In the first phase of the method, we select a subset of length two from available features, feature i and j. For each drug interaction, k, we select ith and jth element in the chemogenomic profiles of the drugs involved in the interaction to calculate the correlation between these two sub-profiles. We defined *correlation vector*s as pairwise correlation between sub-profiles of the drugs in an interaction. Figure 4.3 represents depicts the process to calculate correlation vectors. In Figure 4.3 drugs x and y are two drugs in interaction k.

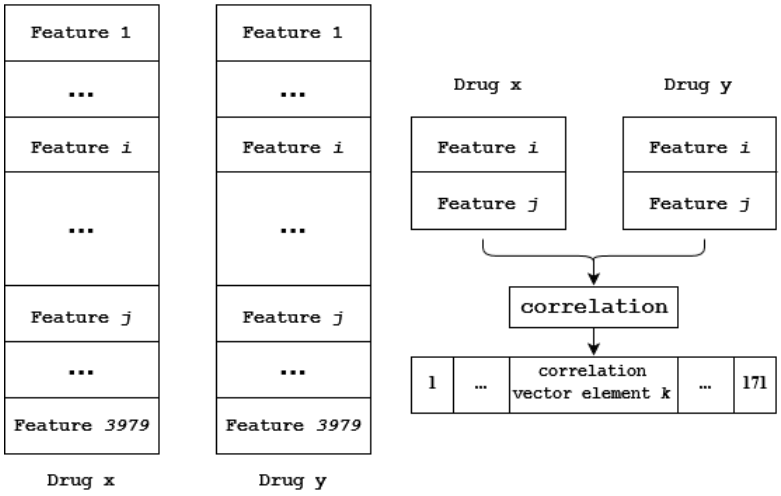

FIGURE 4.3: Element k in the correlation vector is a result of correlation between elements i and j in feature vectors x and y

After calculating all elements of the correlation vector, q, we calculate the correlation between alpha scores and q which is the value of fitness function for GRASP.

Fitness values vector and q have the same length equal to number of the interactions.
Fitness value indicates similarity between alpha scores and feature vector using only ith and jth feature instead of all. Similarly we can calculate fitness values for a subset of more than two features.

In the first part of the method, We find the fitness value for all possible couples of features, with length of $\binom{3979}{2} = 7914231$. As we expect features which represent alpha scores better will be more successful in the classification process, we select a subset of n feature couples with highest fitness values.

A random couple out of n is selected in the beginning of the second part to start an *experiment*. In each iteration of the experiment one feature is added to the subset of selected features. Starting from the first feature we check whether it has been already selected or not. If it has not been used before, we concatenate that feature to the subset of selected features and objective values are calculated using new set of selected features. We select k features which yielded best objective values and due to the nature of GRASP algorithm we add one feature which has not been among k best objective values. Adding features is continued until a predefined limit, F, is reached in an experiment. E experiments will be done and the best result is reported. Figure 4.4 represents the steps of each experiment.

4.2.3 KNN classifier

K-Nearest Neighbors algorithm is an instance-based classification method, i.e., it stores all available cases and classifies new cases based on a similarity measure ,e.g., distance functions [23] Classification using an instance-based classifier is performed by locating the nearest neighbor in instance space and labeling the unknown instance with the same class label as that of the located neighbor in training data. A disadvantage in using this approach its sensitivity to the noise in training data.

In order to achieve more robust models, location of k, where k ¿ 1, neighbours can be considered and majority vote can decide the outcome of the class labelling. A higher value of k results in a smoother, less locally sensitive, function. The nearest neighbour classifier can be regarded as a special case of the more general k-nearest neighbours classifier, hereafter referred to as a kNN classifier.

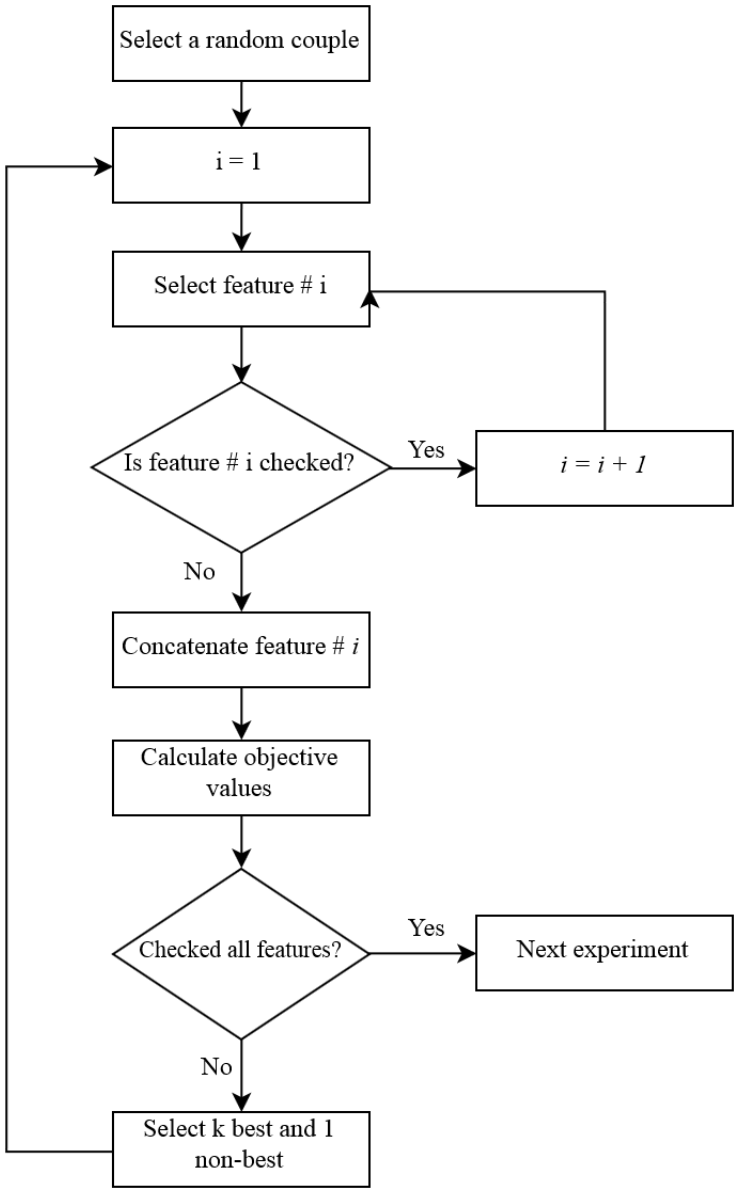

FIGURE 4.4: Flowchart of experiments done using GRASP algorithm in second part of feature selection method

4.3 Results

4.3.1 GA Results

In case of E. Coli, objective value was -0.24 before performing GA and the relationship between correlation of chemo-genomic profiles and alpha scores was as depicted in Figure 4.5

Fitness function value was decreased to -0.58 by applying GA. Fitness function for

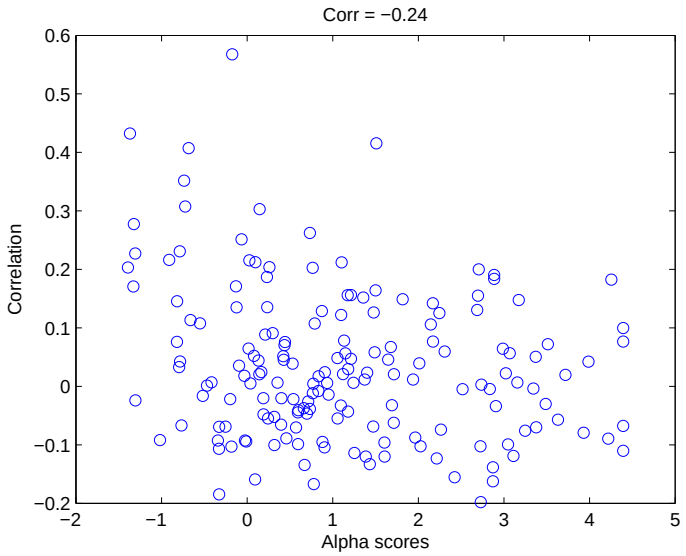

FIGURE 4.5: Correlation of chemo-genomic profiles versus alpha scores of drug interactions before applying genetic algorithm in E. Coli dataset

each iteration is depicted in Figure 4.6. Relationship between correlation vector and alpha scores after GA in E. Coli dataset is represented in 4.7.

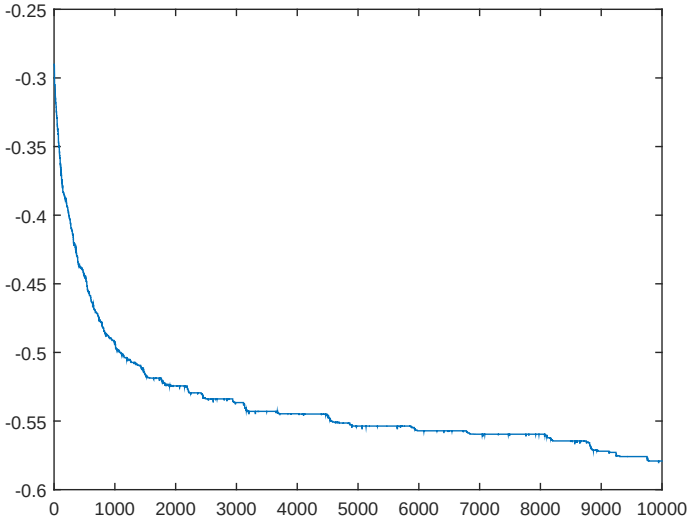

FIGURE 4.6: Fitness function in each iteration of GA in E. Coli dataset

Size of selected sub-profile is around 1600 features. By applying genetic algorithms we could select a subset of features by which the value of objective function has been minimized.

For yeast before applying genetic algorithm, using 67 out of 231 possible drug interactions, value of objective function was -0.15. A genetic algorithm with two

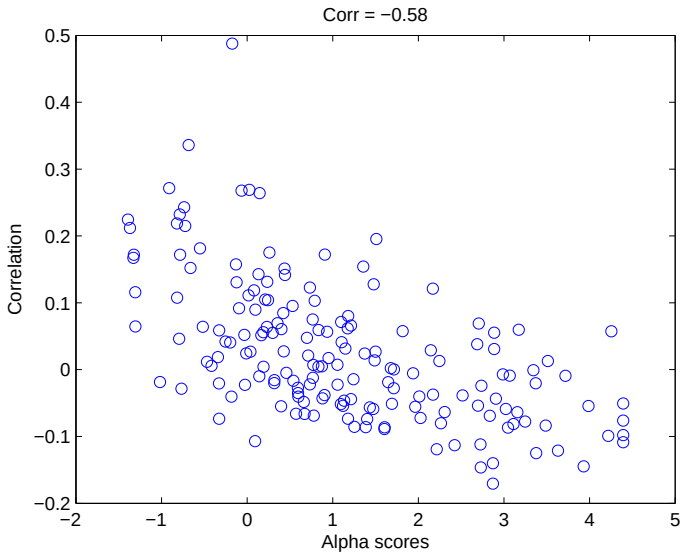

FIGURE 4.7: Relationship between alpha scores and sub-profiles after GA in E. Coli dataset

point crossover function and 1% mutation was performed for 20000 and the maximum absolute value of objective function in each iteration were as depicted in Figure 4.8

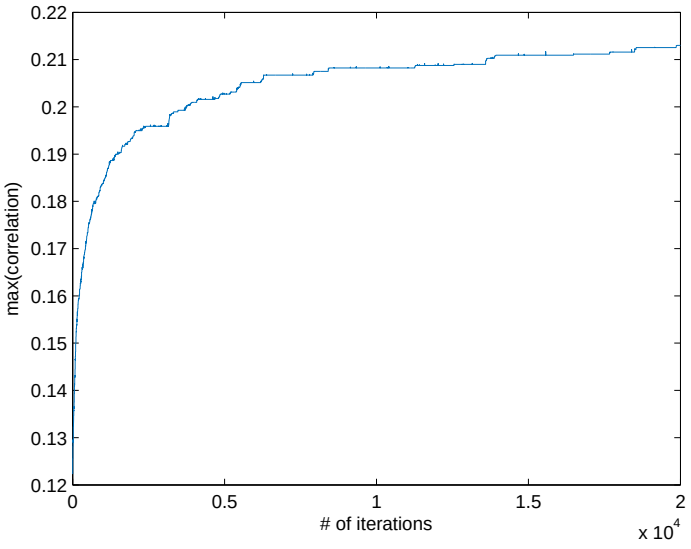

FIGURE 4.8: Value of objective function in each iteration of genetic algorithm

TABLE 4.1: Parameters used for GRASP algorithm

Number of experiments	20
Number of selected features	32
Number of best features selected	5
n	100

TABLE 4.2: Parameters used for GRASP algorithm

Number of experiments	10
Number of selected features	20
Number of best features selected	5
n	100

4.3.2 GRASP results

Figure 4.9 represent the final output of feature reduction procedure using the parameters stated in Table 4.1 for the model.

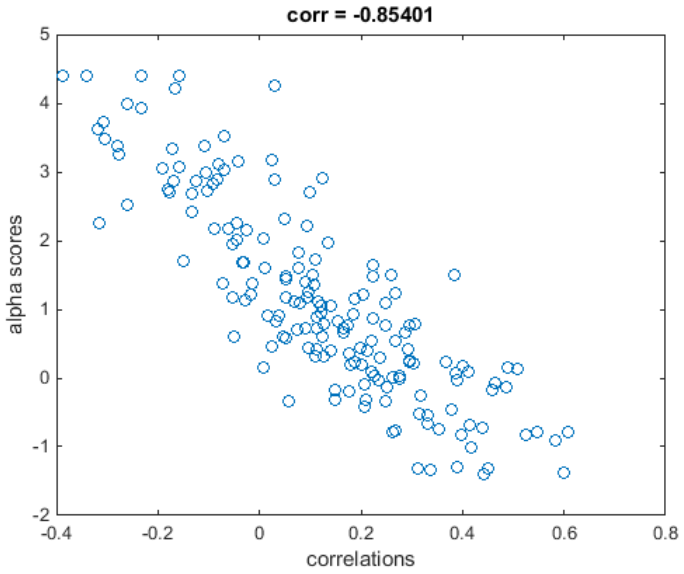

FIGURE 4.9: Objective value has decreased by using the features selected by GRASP with parameters stated in Table **??**.

We tried another set of parameters but fitness value was not improved, Table 4.2 represents used parameters and Figure 4.10 depicts the relationship between correlation vector and alpha scores.

In order to verify that using selected features yields in a better classification accuracy, we predicted interaction types using desicion tree and KNN classifiers. These

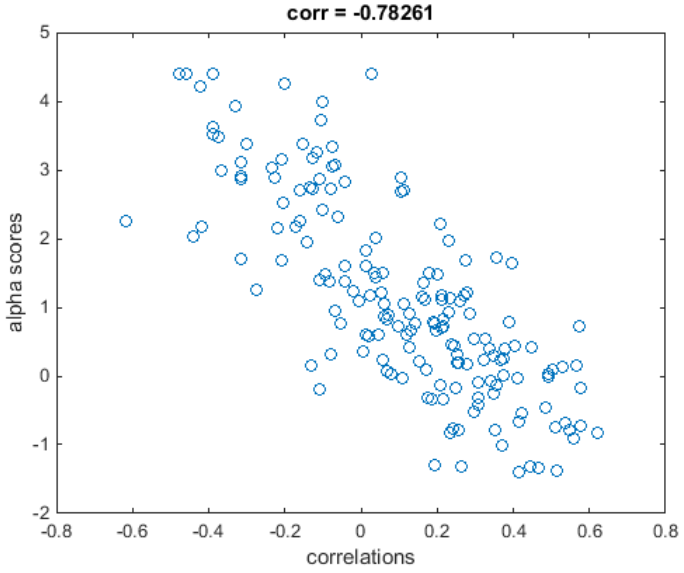

FIGURE 4.10: Objective value has decreased by using the features selected by GRASP with parameters stated in Table 4.2

TABLE 4.3: Confusion matrix after using

148	3
7	13

two common classifiers voted for the type of each of 165 interactions, i.e. if *both* of them predicted the interaction to be synergistic, the final output is synergy. PR Tools [24] was used as a means to perform the classification, 5-fold cross validation and voting procedure.

Selected features of experiments were sorted and the subset which leads to the best objective function was selected, i.e., the same feature set related to Figure 4.9. Since value of objective function is not strictly increasing inside the selected subset, a *reduced subset* of features were selected. Reduced subset contains members of the original subset from the first member to the one that leads to the best fitness function. Since in our analysis the best fitness function was the last member of the feature subset, reduced subset contained all 32 selected features.

Correlation vector was calculated using selected features and was introduced to the voting classifiers. Binarized alpha-scores with threshold -0.5, which is extraced fron expert knowledge, was used as the labels for learning. 4.3 represents the confusion marix of our analysis. Area Under ROC equal to 0.82 and Area Under PR curve of 0.87.

Chapter 5

Utilizing Chemical Features

5.1 Introduction

Chemical features of drugs participating in an interaction can be used as a means to predict synergy. These features are based on chemical formulations and can be calculated mathematically which makes them significantly more reliable in comparison to the features based on experiments. It has been seen that if a feature is correlated significantly with synergicity, it can be a leading to more accurate predictions. Drug lipophilicity is an example of such features and has been used in [2]. Lipophilic drugs have a higher tendency to be synergistic with other drugs [2]. In this chapter we reproduced results found in [2] with a different classifier in first step and selected a new feature, Q_PC-, from a subset which consisted of 327 drug descriptors. We found that although Q_PC- is not neither correlated with lipophilicity nor synergicity, using it will yield in good classification results. Decision tree and random forests were used in order to perform the classification. Details of the methods used will be discussed in the following sections.

5.2 Classifiers

In this section a brief introduction to the used classification methods will be provided.

5.2.1 Decision tree

Decision tree classifiers use graphs in tree format to model possible outcomes of a datum based on attributes it possesses . Each internal node in the tree represents a test on the data attribute and based on the criterion which is calculated using the value of that attribute, a branch is made. When all data attributes are considered or a stop condition is met, a label(an outcome) is assigned to the datum. Final nodes of a decision tree, i.e. assigned class labels, are called *leaves* of the tree. If nodes of the tree are traced from root to the leaves, a *decision rule* can be generated. Decision rules verbally express the rules generated by the decision tree algorithm. Figure 5.1 depicts an example of a trained decision tree on *"play tennis"* dataset. This dataset consists of weather conditions and labels which state whether it is suitable to play tennis or not. Leaves of this tree represent the decision that is suggested by the algorithm. Based on this decision tree We can generate a decision rule which states *"if outlook is sunny and humidity is normal, it is suitable to play tennis today"*. If a node is closer to the root of a tree, it contains more valuable information than the one that is further.

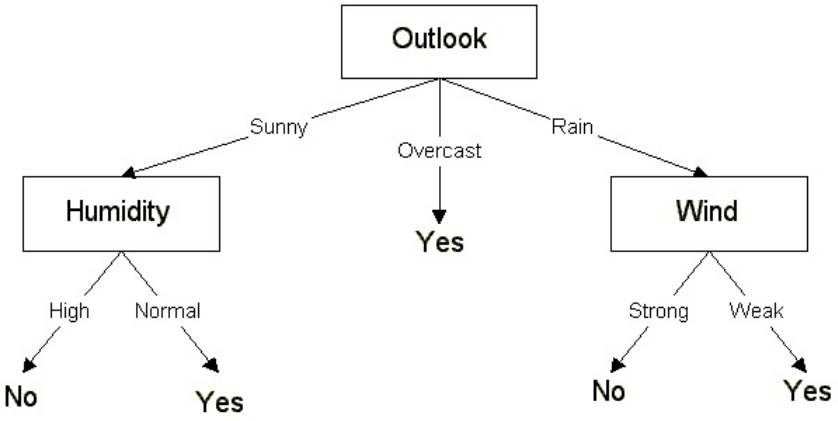

FIGURE 5.1: An example of a trained decision tree for "*play tennis*" dataset

In the analysis that we have done in this chapter, nodes of the trained decision tree consists of the chemical features of first and second drug in a test interaction and the output is based on the chemical attributes are they synergistic or not. MATLAB's default implementation of decision trees has been used to train the models.

5.2.2 Random Forest

A *random forest* uses an an ensemble of decision trees in training phase. Median of outputs from these individual trees are reported as the final label random forest has generated. Each tree in the ensemble is trained by using a subset of input variables, i.e., if there are V variables in input data, a subset v of these variables is selected randomly with replacement to branch the nodes of the trees in an ensemble [25]. Using random forests can avoid over-fitting that is a common problem in decision tree algorithm. They also perform better in unbalanced datasets [25]. In an unbalanced dataset, majority of training set consists of data from one class, thus, the classifier has a tendency to label test data as the majority class. Performance on unbalanced datasets is an important factor in selecting classification methods. In our analysis we have used the random forest implemented in WEKA toolkit [26].

5.3 Chemical features

5.3.1 Lipophilicity feature

Lipophilicity, XLogP3, of drugs has a significantly high correlation with drug synergicities, Figure 5.2. The method used in [2] to predict synergy in yeast dataset was replicated with random forest classifier of WEKA toolkit [26] in MATLAB. The feature vector introduced to the random forest classifier was XLogP3 value of first drug concatenated with XLogP3 value of the second drug and vice versa, i.e. we added values of second drug concatenated with first to the bottom of the previous vector. As there were 165 interactions available in yeast dataset, total number of rows of (observations) in the feature vector was $2 * 165 = 330$, Figure 5.3. WEKAs random forest was used in MATLAB with the following settings

- Unlimited depth of trees
- 2 features to be used in random selection
- 100 as the number of trees to be trained

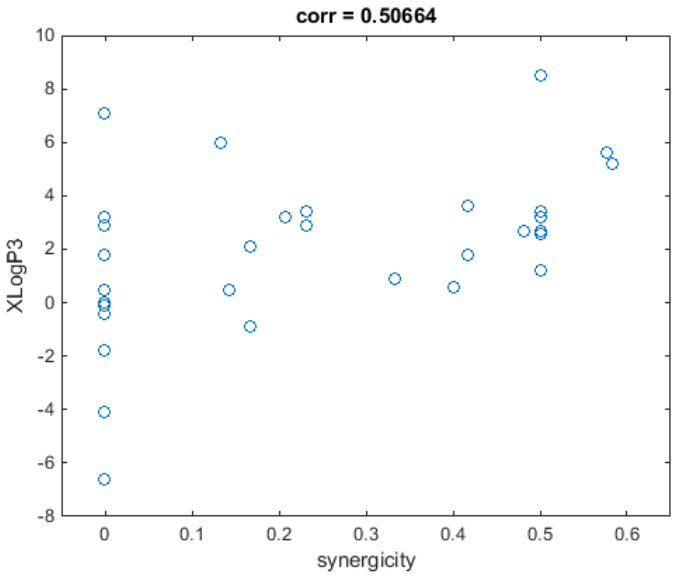

FIGURE 5.2: XLogP3 has a significantly high correlation with synergicity, Spearman r = 0.51 , p = 0.0036

XLogP3 Drug #1	XLogP3 Drug #2
XLogP3 Drug #1	XLogP3 Drug #3
...	...
XLogP3 Drug #30	XLogP3 Drug #31
XLogP3 Drug #2	XLogP3 Drug #1
XLogP3 Drug #3	XLogP3 Drug #1
...	...
XLogP3 Drug #31	XLogP3 Drug #30

FIGURE 5.3: Feature vector introduced to random forest classifier.

5.3.2 Q_PC- Feature

We observed features which are not correlated with synergicity can be equally a good means to predict interaction types. Q_PC- is one of these features that is neither correlated with synergicity nor lipophilicity, Figure 5.4.

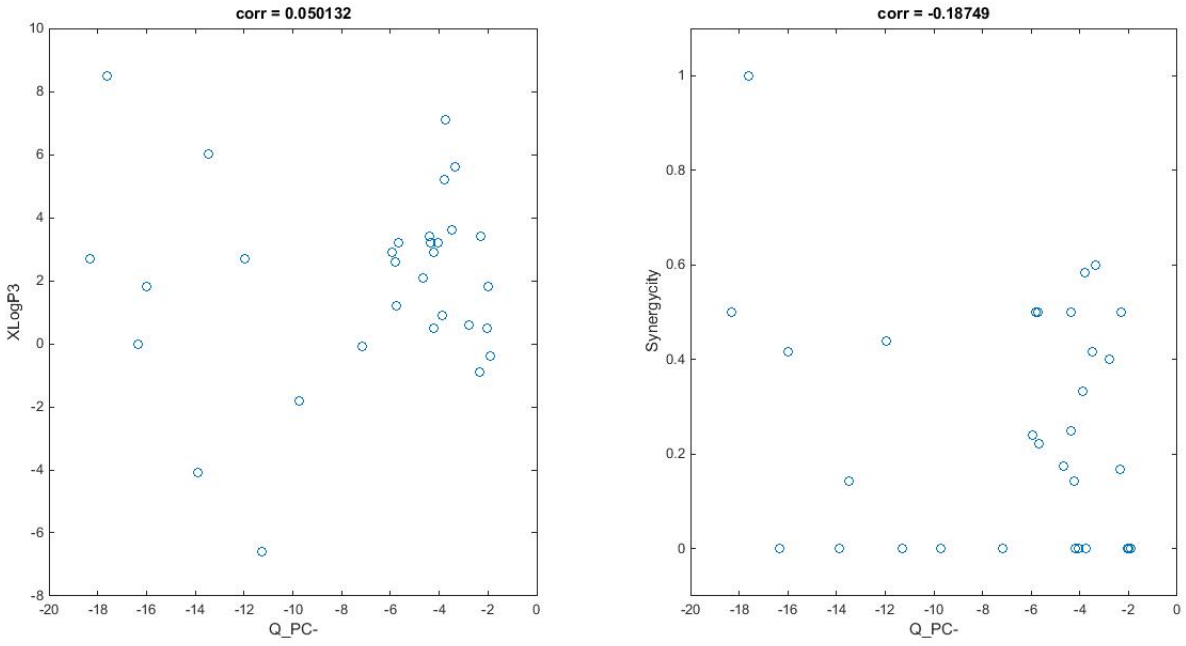

FIGURE 5.4: Q_PC- is not correlated with XLogP3 nor synergicity.

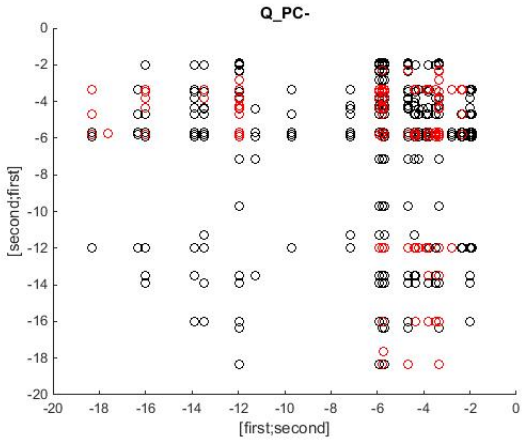

FIGURE 5.5: Q_PC- features were used in the classifier by concatenating the both values for Q_PC- of drugs involved in an interaction

We used the same pattern of concatenating the features of first and second drugs to be introduced to classifier, Figure 5.5. If we apply a simple rule based on Figure 5.5 to identify interactions with Q_PC- value of more than -6 for both drugs as synergistic, an AU-ROC of 0.55 will be achieved. This result promises that if the random forest divides the feature space into more accurate subspaces, more accurate predictions can be obtained.

5.4 Results

5.4.1 E. Coli Results

Resulted AU-ROC and AU-PR curve of performing 10-fold cross validation for 300 iterations are represented in Figures 5.6 and 5.7 respectively.

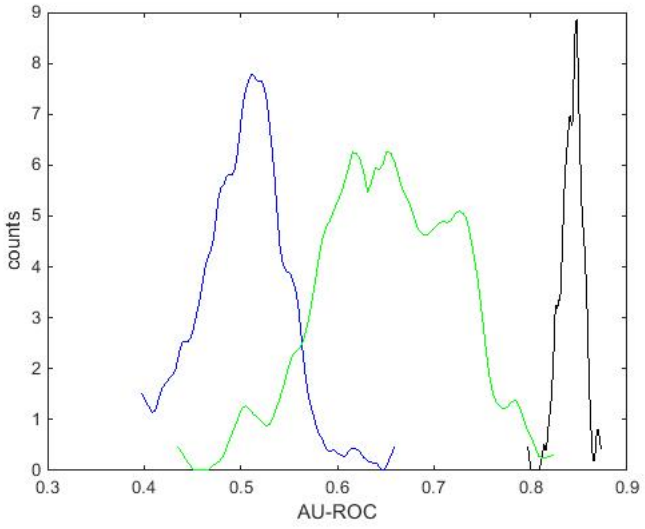

FIGURE 5.6: Distribution of AU-ROC when using random forest as a classifier along with XLogP3 as feature.

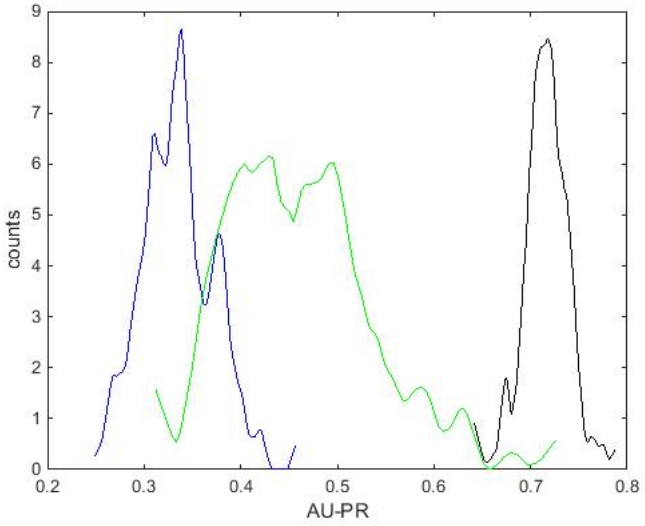

FIGURE 5.7: Distribution of AU-PR when using random forest as a classifier along with XLogP3 as feature.

In Figures 5.6 and Figure 5.7 black color represents the results for the original, green for the node shuffled and blue for the edge shuffled synergy graph. As

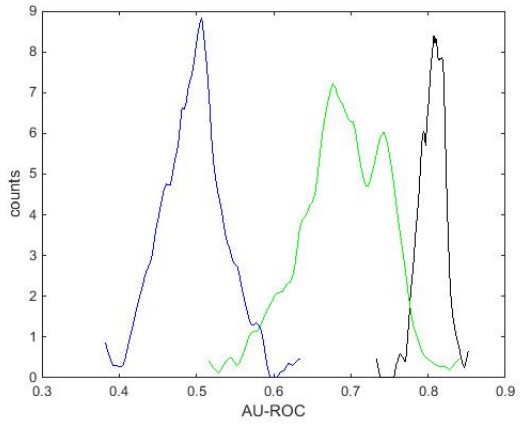

FIGURE 5.8: AU-ROC when using Q_PC- feature in a random forest classifier. Maximum AU-ROC for original network is 0.86

described in Chapter 3, the results are as expected, i.e. original graph's AU-ROC is more than node-shuffled AU-ROC and node-shuffled AU-ROC is more than edge-shuffled AU-ROC. By adopting random forest as the classification method, maximum AU-ROC for original network was improved from 0.80 in [2] to 0.87 and maximum AU-PR to 0.79.

5.4.2 Q_PC- Results

A similar random forest was applied to the feature space and the results in Figures 5.8 and 5.9 were achieved for AU-ROC and AU-PR respectively. Analyses were repeated for 300 iterations of 10 fold cross validation. Reviewing the AUROC distribution suggests that Q_PC- can be a as discriminative in predicting synergies as XLogP3. In order to further investigate this feature we conducted the analyses using decision trees in MATLAB with default settings and obtained similar results.

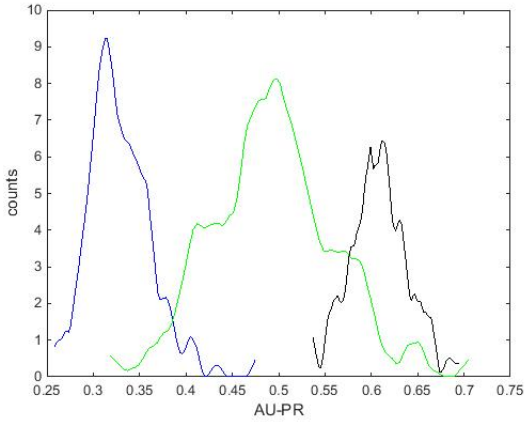

FIGURE 5.9: AU-PR when using Q_PC- feature in a random forest classifier.

Chapter 6

Discussion and Conclusion

In this thesis, we studied the drug synergy prediction problem in two important datasets, E. Coli and yeast. Chemogenomic profiles and chemical drug features were used in the analysis. We used three different approaches.

In our first survey, we introduced a new method in Chapter 3 to predict synergies based on the topological information gained from chemogenomic profiles in both E. Coli and yeast datasets. This algorithm resulted in Area Under ROC curve and the accuracy of *0.71* and *0.91*, respectively, in E. Coli dataset.

Next, in Chapter 4 we used meta-heuristic algorithms for feature subset selection and synergy prediction in E. Coli dataset. We employed genetic algorithm with two point crossover procedure to select a subset of features and maximize the defined objective function. In the second part of the chapter, we defined a GRASP algorithm to the first perform feature subset selection and predict synergies in E. Coli dataset in the second step. Applying this method yielded in *0.94* accuracy and *0.82* Area Under ROC curve for E. Coli dataset.

Finally, Chapter 5 was dedicated to utilize chemical features. We used decision trees and random forest in order to predict synergies in yeast dataset. Two chemical features were utilized, *XLogP3* and *Q_PC-*. Our analysis suggests that Q_PC-chemical feature can be as discriminative as XLogP3, despite the fact that *Q_PC-* does not have a significant correlation with synergicity scores of the drugs. XLogP3 has been used in literature previously [2]. It has been seen that utilizing XLogP3 will lead to more accurate predictions in yeast dataset.

We observed that utilizing chemical features in yeast dataset results in more accurate synergy predictions in comparison to other methods. A future step can perform the same analysis on E. Coli dataset. More chemical features can also be explored to find other discriminateive features. In topological based methods introduced in Chapter 3, more scores can be extracted from distance matrices to observe their effect on the final accuracy of predictions.

Bibliography

[1] Richard D Forrest. Early history of wound treatment. *Journal of the Royal Society of Medicine*, 75(3):198, 1982.

[2] Kaan Yilancioglu, Zohar B Weinstein, Cem Meydan, Azat Akhmetov, Isil Toprak, Arda Durmaz, Ivan Iossifov, Hilal Kazan, Frederick P Roth, and Murat Cokol. Target-independent prediction of drug synergies using only drug lipophilicity. *Journal of chemical information and modeling*, 54(8):2286–2293, 2014.

[3] Gregor Jansen, Anna Y Lee, Elias Epp, Amélie Fredette, Jamie Surprenant, Doreen Harcus, Michelle Scott, Elaine Tan, Tamiko Nishimura, Malcolm Whiteway, et al. Chemogenomic profiling predicts antifungal synergies. *Molecular systems biology*, 5(1):338, 2009.

[4] Scott Treadway. Exploring the universe of ayurvedic botanicals to manage bacterial infections. *Clinical Nutrition Insights*, 6(17):1–3, 1998.

[5] William Foster and Alain Raoult. Early descriptions of antibiosis. *JR Coll Gen Pract*, 24(149):889–894, 1974.

[6] Barry R. Bloom. *Tuberculosis: Pathogenesis, Protection, and Control*. American Society Mic Series. ASM Press, 1994. ISBN 9781555810726.

[7] Thomas P Van Boeckel, Sumanth Gandra, Ashvin Ashok, Quentin Caudron, Bryan T Grenfell, Simon A Levin, and Ramanan Laxminarayan. Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. *The Lancet Infectious Diseases*, 14(8):742–750, 2014.

[8] Maique Weber Biavatti. Synergy: an old wisdom, a new paradigm for pharmacotherapy. *Brazilian Journal of Pharmaceutical Sciences*, 45(3):371–378, 2009.

[9] Elisabeth M Williamson. Synergy and other interactions in phytomedicines. *Phytomedicine*, 8(5):401–409, 2001.

[10] CK Atal, Usha Zutshi, and PG Rao. Scientific evidence on the role of ayurvedic herbals on bioavailability of drugs. *Journal of ethnopharmacology*, 4(2):229–232, 1981.

[11] Xing Chen, Biao Ren, Ming Chen, Ming-Xi Liu, Wei Ren, Quan-Xin Wang, Li-Xin Zhang, and Gui-Ying Yan. Asdcd: antifungal synergistic drug combination database. *PLoS One*, 9(1):e86499, 2014.

[12] Grant R Zimmermann, Joseph Lehar, and Curtis T Keith. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug discovery today*, 12(1):34–42, 2007.

[13] S Loewe. The problem of synergism and antagonism of combined drugs. *Arzneimittel-Forschung*, 3(6):285, 1953.

[14] Markus Bredel and Edgar Jacoby. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nature Reviews Genetics*, 5(4):262–275, 2004.

[15] M.T. Madigan, J.M. Martinko, K.S. Bender, D.H. Buckley, and D.A. Stahl. *Brock Biology of Microorganisms, Global Edition: UEL*. Pearson Education Limited, 2015. ISBN 9781292068312.

[16] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507): 1304–1351, 2001.

[17] Robert J Nichols, Saunak Sen, Yoe Jin Choo, Pedro Beltrao, Matylda Zietek, Rachna Chaba, Sueyoung Lee, Krystyna M Kazmierczak, Karis J Lee, Angela Wong, et al. Phenotypic landscape of a bacterial cell. *Cell*, 144(1):143–156, 2011.

[18] Maureen E Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P St Onge, Mike Tyers, Daphne Koller, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320(5874):362–365, 2008.

[19] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Molecular systems biology*, 2(1), 2006.

[20] Sriram Chandrasekaran, Melike Cokol-Cakmak, Nil Sahin, Kaan Yilancioglu, Hilal Kazan, James J Collins, and Murat Cokol. Chemogenomics and orthology-based design of antibiotic combination therapies. *Molecular systems biology*, 12(5):872, 2016.

[21] Coen Bron and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

[22] Thomas A Feo and Mauricio GC Resende. Greedy randomized adaptive search procedures. *Journal of global optimization*, 6(2):109–133, 1995.

[23] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[24] Ferdinand Van Der Heijden, Robert Duin, Dick De Ridder, and David MJ Tax. *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*. John Wiley & Sons, 2005.

[25] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[26] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.