

COMPUTATIONAL APPROACHES TO STUDY DRUG RESISTANCE MECHANISMS

by,
ZOYA KHALID

Submitted to the Graduate School of Engineering and Natural Sciences
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy
in
Molecular Biology, Genetics and Bioengineering

Sabanci University

Spring 2017

COMPUTATIONAL APPROACHES TO STUDY DRUG RESISTANCE
MECHANISMS

APPROVED BY:

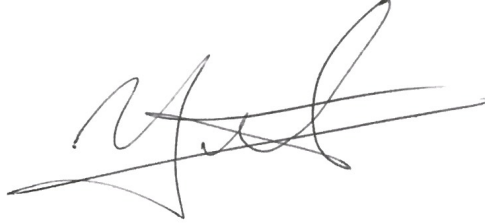
Prof. Dr. Ismail Cakmak.
(Thesis Supervisor)



Prof. Dr. Osman Ugur Sezerman



Prof. Dr. Yucel Saygin.



Assoc. Prof. Dr. Devrim Gozuacik.



Asst. Prof. Dr. Ozgur Asar



DATE OF APPROVAL: 14. 04. 2017

ABSTRACT

Computational Approaches to Study Drug Resistance Mechanisms

Zoya Khalid

Keywords: *Drug Resistance, Text Mining, Relation Extraction, Drug Repurposing, HIV resistance*

Drug resistance is a major obstacle faced by therapists in treating complex diseases like cancer, epilepsy, arthritis and HIV infected patients. The reason behind these phenomena is either protein mutation or the changes in gene expression level that induces resistance to drug treatments. These mutations affect the drug binding activity, hence resulting in failure of treatment. All this information has been stored in PubMed directories as text data. Extracting useful knowledge from an unstructured textual data is a challenging task for biologists, since biomedical literature is growing exponentially on a daily basis. Building an automated method for such tasks is gaining much attention among researchers.

In this thesis we have developed a disease categorized database ZK DrugResist that automatically extracts mutations and expression changes associated with drug resistance from PubMed. This tool also includes semantic relations extracted from biomedical text covering drug resistance and established a server including both of these features. Our system was tested for three relations, Resistance (R), Intermediate (I) and Susceptible (S) by applying hybrid feature set. From the last few decades the focus has changed to hybrid approaches as it provides better results. In our case this approach combines rule-based methods with machine learning techniques. The results showed 97.7% accuracy with 96% precision, recall and F-measure. The results have outperformed the previously existing relation extraction systems thus facilitating computational analysis of drug resistance against complex diseases and further can be implemented on other areas of biomedicine.

Literature is filled with HIV drug resistance providing the worth of training data as compared to other diseases, hence we developed a computational method to predict HIV resistance. For this we combined both sequence and structural features and applied SVM and Random Forests classifiers. The model was tested on the mutants of HIV-1 protease and reverse transcriptase. Taken together the features we have used in our method, total contact energies

among multiple mutations have a strong impact in predicting resistance as they are crucial in understanding the interactions of HIV mutants. The combination of sequence-structure features offers high accuracy with support vector machines as compared to Random Forests classifier. Both single and acquisition of multiple mutations are important in predicting HIV resistance to certain drug treatments. We have discovered the practicality of these features; hence these can be used in the future to predict resistance for other complex diseases.

Another way to deal drug resistance is the application of drug repurposing. Drug often binds to more than one targets defined as polypharmacology which can be applied to drug repositioning also referred as therapeutic switching. The traditional drug discovery and development is a high-priced and tedious process, thus making drug repurposing a popular alternate strategy. We have proposed a method based on similarity scheme that predicts both approved and novel targets for drug and new disease associations. We combined PPI, biological pathways, binding site structural similarities and disease-disease similarity measures. We used sixty drugs for training the algorithm and tested it on eight separate drugs. The results showed 95% accuracy in predicting the approved and novel targets surpassing the existing methods. All these parameters help in elucidating the unknown associations between drug and diseases for finding the new uses for old drugs. Hence repurposing offers novel candidates from existing pool of drugs providing a ray of hope in combating drug resistance.

ÖZET

İlaç Direnç Mekanizmaları için İşlemsel Yaklaşımlar

Zoya Khalid

Anahtar Kelimeler: *İlaç Direnci, Metin Madenciliği, İlişkisel Çıkarım, İlaç Repurposing, HIV direnci*

İlaç direnci, kanser, epilepsy, artrit ve HIV gibi kompleks hastalıkların tedavisi sürecinde terapistlerin karşılaştığı büyük bir engeldir. Protein mutasyonu veya gen ifadesindeki değişiklik düzeyi bu olayın arkasındaki sebeptir. Bu mutasyonlar ilaç bağlanma aktivitesini etkilemekte, bu nedenle de tedavinin başarısızlıkla neticelenmesine sebep olmaktadır. Bu bilgiler PubMed rehberlerinde metin verisi olarak arşivlenmektedir. Biyomedikal literatürü günlük olarak katlanarak büyümekte olduğundan, biyologlar için, yapısal olmayan metin verilerinden kullanılabilir bilgiyi seçip çıkartmak zorlu bir iştir. Bu ve buna benzer görevleri gerçekleştirebilmek için otomatikleştirilmiş bir yöntem geliştirmek araştırmacıların ilgisini çekmektedir.

Bu tez çalışmasında, ilaç direnciyle ilişkili mutasyonları ve ifade değişikliklerini PubMedden otomatik olarak seçip çıkararak, hastalıklara göre kategorilenmiş ZK DrugResist adlı bir veri bankası geliştirdik. Bu araç, ilaç direnciyle ilgili olan biyomedikal metinlerden elde edilen anlamsal ilişkileri de içermektedir ve bu iki özelliği de içeren bir sunucu kurulmuştur. Sistemimiz hibrit özellik seti uygulanarak üç ilişki bakımından test edildi, Direnç (R), Orta derece (I) ve Duyarlılık (S). Son on yıllık süreden beri odak noktası, daha iyi sonuçlar verdiği için hibrit yaklaşımlara değişmiştir. Bizim vakamızda, bu yaklaşım kurula-dayalı yöntem ile özdevimli öğrenme tekniklerini birleştirmektedir. Sonuçlar %96 hassasiyet, geri çekme ve F-ölçümü ile %97.7 doğruluk göstermiştir. Sonuçlar, daha önceden varolan bağıntı çıkarım sistemlerinden daha iyi olduğunu göstermiş, böylece kompleks hastalıklara karşı gelişen ilaç direncinin işlemsel analizlerini kolaylaştırmıştır ve daha ötesi farklı dirimsel tıp alanlarına da uygulanabilmektedir.

Literatür, diğer hastalıklara göre kıyaslandığında, HIV ilaç direnciyle ilgili çalışma verilerinin değerini gösteren araştırmalarla doludur, bu nedenle HIV direncini öngörmek amaçlı işlemsel bir yöntem geliştirdik. Bunun için dizi ve yapı özelliklerini birleştirdik ve Destekçi Vektör Makinesi (SVM) ve Rastgele Orman klasifikatörleri uygulandı. Model, HIV-1 proteaz ve ters transkriptaz mutantları üzerinde test edilmiştir. Yöntemimizde kullandığımız

özellikleri birleştirdik, birden fazla mutasyon arasındaki total temas enerjilerinin, HIV mutantları arasındaki etkileşimi anlayabilmek için önemli olmalarından dolayı, direnç tahmini üzerinde güçlü etkisi vardır. Dizi-yapı özelliklerinin kombinasyonu SVMler ile, rastgele orman klasifikatörüne göre yüksek doğruluk sağlamaktadır. Tek ve birden çok mutasyonun her ikisinin de kazancı belli ilaç tedavilerine karşı oluşan HIV direncinin tahmini için önemlidir. Bu özelliklerin kullanılabilirliğini keşfettik, bu sayede bu özellikler aynı zamanda diğer kompleks hastalıklara karşı gelişen direnç durumunu öngörmek için gelecekte kullanılabilir. İlaç direnci ile ilgilenmenin bir diğer yolu ilaç repurposing uygulamasıdır. Terapötik değişim olarak da adlandırılan, ilacın genellikle birden fazla hedefe bağlanabilmesini tanımlayan polifarmasi ilaç yeniden konumlandırmasında kullanılabilir. Geleneksel ilaç keşif ve geliştirme yöntemi pahalı ve meşakkatli bir süreçtir, bu nedenle ilaç repurposing yöntemi popüler alternatif bir stratejidir. Biz de ilaç ve yeni hastalık ilişkileri için kabul edilen ve yeni hedefleri öngören benzerlik şemasına dayanan bir yöntem önerdik. PPI, biyolojik yolaklar, bağlanma bölgelerinin yapısal benzerlikleri ve hastalık-hastalık benzerlik ölçülerini birleştirdik. Algoritmayı eğitmek için altmış ilaç kullandık, ve algoritmayı sekiz ayrı ilaçta test ettik. Sonuçlar, doğrulanmış ve yeni hedefleri öngörmede %95 doğruluk göstererek varolan yöntemlerden üstünlüğünü göstermiştir. Bütün parametreler ilaç ve hastalıklar arasındaki bilinmeyen ilişkileri açıklayarak eski ilaçların yeni kullanımlarının bulunmasına yardım etmiştir. Böylece repurposing varolan ilaç havuzundan yeni adaylar önererek, ilaç direnciyle mücadelede bir umut olabilir.

©Zoya Khalid 2017
All Rights Reserved

This work is dedicated to

To My Parents

The reason of what I become today

To My Sister

The reason of my happy days

To My Husband

My Inspiration and my soulmate

Acknowledgements

The path towards this dissertation has been circuitous. Its completion is thanks in large part to the people who challenged, supported, motivated and stuck with me along the way. I am tremendously fortunate to have my thesis advisor **Prof.Dr.Uğur Sezerman**. I would like to express my deepest gratitude to him for the patient guidance and encouragement. At many stages, I benefited from his advice, particularly so when exploring new ideas. His positive outlook and confidence in my research inspired me and gave me confidence. His careful editing contributed enormously to the production of this thesis.

I also want to thank the respected jury members Prof Dr İsmail Çakmak, Prof Dr. Yücel Saygin, Doc.Dr Devrim Gözüaık and Yrd.Doc.Dr.Özgür Asar for their valuable feedback. I gratefully acknowledge the financial support received from Higher Education Commission of Pakistan (HEC) to complete PhD degree.

Special thanks to all my friends and to the members of Sezerman lab for all kind of moral and technical support.

Bundle of Thanks to my dear husband Nayyar Mehmood for his continued support and understanding, I can't thank you enough for encouraging me throughout this experience and makes the completion of this thesis possible.

I would also like to express my gratitude to my dearest sister Maria Khalid and my friend Kousar Aslam for patiently tolerating me in my thesis and encouraging me throughout this time.

Lastly, I want to thank my parents for their prayers, unconditional love and support.

Contents

List of Figures	x
List of Tables	xi
1 Introduction	1
1.1 Mechanisms of Drug Resistance	1
1.2 A disease categorized drug response database via Text Mining	3
1.3 Combining Sequence and Structural Features for Predicting Drug Resistance	5
1.4 Drug Repurposing	7
1.5 Motivation	8
1.6 Structure of the Thesis	9
1.7 Published Articles	9
2 Background	10
2.1 Text Mining	10
2.2 Identifying Drug Resistance by combining sequence and structure features	12
2.3 Drug Repurposing	14
3 Materials and Methods	17
3.1 Text Mining	17
3.1.1 Drug Resistance Vs. Others	17
3.1.2 Mutation Vs. Expressions	17
3.1.3 Protein Vs. DNA	18
3.1.4 Cancer Vs. Others	18
3.1.5 Relation Extraction	19
3.1.6 Features and Model Building	19
3.2 Implementation and Usage	21
3.3 Sequence and Structural Features	22

3.3.1	Frequency count	22
3.3.2	Conserved Mutations using PSSM	22
3.3.3	Measuring Flexibility and Rigidity	22
3.3.4	Disordered Regions	23
3.3.5	Hydrophobicity Measure	23
3.3.6	Volume Measure	23
3.3.7	Secondary Structure Features	24
3.3.8	Solvent Accessibility	24
3.3.9	Structure and contact residues	24
3.3.10	Interactions between multiple mutations	24
3.4	Preprocessing Filters and Feature Selection	25
3.4.1	SMOTE and SpreadSubsample Filters	25
3.5	Model Building	26
3.5.1	Support Vector Machines	26
3.5.2	Random Forests	26
3.6	Drug Repurposing	26
3.6.1	Computing Drug Disease Network	27
3.6.2	Building a Scoring System	27
4	Results	30
4.1	Text Mining	30
4.2	Sequence and Structure Features	37
4.3	Drug Repurposing	42
5	Discussion	45
6	Conclusions	49
	Bibliography	51
	Appendices	64
A	List of FDA Approved HIV and Cancer Drugs	65
B	Drug Repurposing Datasets	68

List of Figures

3.1	Modules for extracting Abstracts	18
3.2	Modules for Relation Extraction and Relation Classification	21
3.3	Workflow of Drug Repurposing Methodology	29
4.1	ZK DrugResist snapshot of mutations associated with Cancer	33
4.2	ZK DrugResist snapshot of mutations associated with HIV	34
4.3	Relation Extraction: Resistance(R) , Intermediate (I), Susceptible (S) . . .	35
4.4	ZK DrugResist snapshot of expression changes associated with Cancer	36
4.5	Classification Performance with varying Feature Numbers	38
4.6	Classification Performance with Varying window size	38

List of Tables

2.1	Comparison of Accuracy Measure with the Existing Literature	14
4.1	SVM classification on PubMed Abstracts	31
4.2	Relation Extraction classification results	31
4.3	Regular Expressions	32
4.4	Relations from Sentences	32
4.5	SVM Classification Results on IDV and SQV Drugs of PIs	39
4.6	Random Forests Classification Results on IDV and SQV Drugs of PIs	39
4.7	SVM Classification on PIs Inhibitors	39
4.8	SVM Classification Results on NRTIs and NNRTIs of RTs s	40
4.9	Random Forests Classification Results on PIs, RTs	40
4.10	Random Forests Classification Results on PIs, RTs	41
4.11	Comparison of Accuracy Measure with the Existing Literature	41
4.12	Results on Pathway based Drug Repurposing	43
4.13	Novel Predictions for Drug Repurposing	44
4.14	Total True and False Predictions	44
A.1	Protease Inhibitors PIs	66
A.2	Nucleoside Reverse Transcriptase NRTIs and Non Nucleoside Reverse Tran- scriptase NNRTIs	66
A.3	List of FDA Approved Cancer Drugs	67
B.1	Training set for Drug Repurposing	69
B.2	Test set for Drug Repurposing	70

List of Abbreviations and Symbols

SVM	Support Vector Machines
RF	Random Forests
ADMET	Absorption Distribution Metabolism Excretion and Toxicity
HAART	Highly Active Antiretroviral Therapy
PI	Protease Inhibitors
RT	Reverse Transcriptase
NRTI	Nucleoside Reverse Transcriptase
SQV	Saquinavir
IDV	Saquinavir
NVP	Nevirapine
BOW	Bag of Words
TFIDF	Term Frequency Inverse Document Frequency
NLP	Natural Language Processing
UMLS	Unified Medical Language System
GWAS	Genome Wide Association Studies
CTD	Comparative Toxicogenomics Database
TTD	Therapeutic Target Database

Chapter 1

Introduction

1.1 Mechanisms of Drug Resistance

Before the advent of molecular biology and genetics, phenotypic assays were employed for the drug discovery which involved very little knowledge of molecular mechanisms. Years later this approach has been taken over by target based drug discovery. Target is often defined as a single gene, gene products (proteins) or a biological mechanism. According to one study the target is classified into two classes: Gene based and mechanistic based [1]. Gene based targets are the genes or gene products that tend to carry mutations hence bear the high risk of developing a disease. On the other hand, mechanistic targets are based on how the drug is administered that is its specific mode of action which can be inferred from drug mechanism of actions. The drug functions by binding to other small molecules referred to as substrates or ligands. It binds to the active site of the drug and resulting complex is called as Ligand- Protein complex. When drug binds to its ligand it prevents ligand binding to its natural substrate hence the normal function of the protein is aborted.

Drug treatment is often faced by an obstacle called "drug resistance" generally meaning the decrease in the efficacy of drug in curing a particular disease. Drug resistance is becoming a major health hazard spreading from viral diseases to cancer. Resistance to chemotherapeutic is of two types: intrinsic or acquired. The pre-existence of resistance cells indicates intrinsic resistance, while if the resistance cells appear after treatment it is acquired resistance. Another interesting description of cancer resistance is the appearance of tumor heterogeneity which means different cancer cells can have different morphology in terms of gene expression, metabolism and proliferation. This causes serious impediment in making an effective treatment for cancer, hence providing good reasons to apply pharmacogenomics

for cancer therapy. Furthermore, the use of high throughput techniques is coupled with bioinformatics to identify molecular signatures for predicting drug response.

As it is a primitive cause of treatment failure, many studies have been performed in order to combat drug resistance. One such approach uses change in the treatment that is first line, second line and third line drugs in response to resistance being observed [2]. Sadly, this could not become a very popular choice as choosing first line drugs has less side effects and is also effective to large population compared to second and third line drugs. The alternative is to use multiple drugs also known as drug cocktails. On one hand if this approach increases the efficacy by lowering down the chances of resistance, its side effects making it less effective in patients [3]. Many clinical trials have been conducted to overcome drug resistance, nevertheless drug target develops multiple drug resistance referred as MDR. These trials usually are performed once the resistance is being observed, there should be theoretical knowledge available beforehand that would be helpful in selecting drug target for overcoming resistance.

Redox regulation is also contributing in developing drug resistance. Reactive oxygen species (ROS) causes oxidation of amino acid residues and protein backbone which results in protein fragmentation. These oxidized proteins disrupt the protein functions hence, making cells to adopt altered molecular pathways. ROS modulation is considered as one of the prerequisite for tumor development and also important to measure drug resistance [4].

Effective drug treatments against complex diseases like cancer, epilepsy, arthritis, HIV is greatly affected by drug resistance. The phenomena underlying drug resistance is surrounded by multiple factors which are not well understood as yet. Few notions about it are, generally either the point mutations in drug target or modifications of expression levels makes drug insensitive to treatment [5].

However, the literature shows that the most prevailing method of drug resistance is the acquisition of point mutation in drug target that causes alteration in amino acids at certain residues. These mutations develop at the binding site of proteins hence affecting the drug binding activity making it insensitive to treatment. Few mutations are reported in literature, for instance Dasatinib resistance caused by V299L, T315A, and F317I/L mutations. And Nilotinib resistance is caused by mutations at Y253F/H, E255K/V, and F359C/V [6] [7]. Moreover, those mutations that are not exactly at the binding sites but are located away do not directly participates in drug resistance but are really ambiguous as these might changes the structure of the protein [8]. In addition to that the changes in expression levels which are over-expression or down expression of certain genes also contributes in drug resistance.

The epidermal growth factor (EGFR) is over-expressed in almost 30% of breast cancer patients [9]. In other studies, it is reported that the over-expression of efflux pumps is the contributing cause of Imatinib resistance [10]. Despite of having limited knowledge of drug resistance mechanisms, point mutation of drug targets could be considered as the starting point for predicting and overcoming drug resistance.

1.2 A disease categorized drug response database via Text Mining

Biomedical publications are exponentially increasing with the passage of time making it hard for the researchers to keep themselves updated. Information seems to be overburdening with the addition of new articles on already existing corpus. Accessing relevant information from these online knowledge sources has become a big challenge as it takes huge amount of manual labor leaving it as a time consuming and laborious task. This brings researchers to build a way which should be more advanced than searching keywords. The automated methods aids in providing the prior knowledge before conducting clinical experiments.

Text mining is the technique nearly related to Information Retrieval (IR) that extracts structured data from unstructured text. In other way, it is the application of data mining to the text data. Nevertheless, data mining usually works on structured data and textual data is unstructured most of the times. Text mining has benefits over other methods as it uses algorithms instead of applying manual filters. Secondly, it helps in deriving new relations from the available text. Preprocessing of the text is required before applying data mining which involve natural language processing techniques. Text mining is generally referred as an interdisciplinary field that combines NLP, machine learning and statistics altogether. Textual data is the only way of storing all sorts of information, hence making text mining a popular choice for digging out meaningful knowledge out of huge pool of data.

First step in text mining is to categorize the documents on the basis of their content also called as Text Categorization which in our case is separating drug resistance articles with others. PubMed is queried with terms which include and are not limited to "drug resistance, "mutations", "expression changes" and "complex diseases". The problem is, Biology has a rich source of vocabulary and one term can be referred by different names, this goes especially for proteins and gene names. To address this, using Bag of Words and representing it

as vector space model usually acts as a savior. A classifier using supervised machine learning is then applied to categorize documents. Textual data is way much complex as one term has synonyms, so creating valid set of training and test data is also tedious task because text varies from one domain to another. For instance, the training set created for classifying drug resistance cannot used as test set for protein-protein interactions classification.

Text mining has various techniques. One such technique is applying clustering to text data also referred as document clustering. It is a fast filtering technique which allows to extract information from text that has been grouped according to the similarity between documents. The similarity could be either text-based that includes term frequency and latent semantics or citation-based that helps in providing other related documents [11]. Another approach is Text Summarization which intends to provide a recap containing the highlights of a document. This is beneficial in coping with the information overload. A good summary grabs the important points that helps in deciding whether to read entire document or not. This data reduction process has various techniques reported in literature. The most effective technique for biomedical data is concept chaining that uses the concepts rather than terms. This works by taking in semantically related concepts using UMLS, identifying strong chains by assigning scores and then extract sentences to create the summary. Another way of doing this is making use of frequency distribution technique in which each sentence is being assigned a score based on the frequency count of term or concept [12]. An unsupervised technique of text mining is topic modeling which is a probabilistic model containing a combination of topics. This approach has also been applied on biomedical data for instance a well known example is the classification of drugs based on safety measures and the therapeutic use of drugs. One reported methods for topic modeling is Latent Dirichlet Allocation that calculates the posterior probabilities of words on an input document [13]. Biomedical text has been broadly facilitated by text mining algorithms specially called as biomedical natural language processing BIONLP. Much real life applications have been performed using NLP making it an applied science rather than just theoretical. However, the named entity recognition has always been a dilemma, since the biological terms varies from document to document as one gene or protein has many synonyms. Nowadays in biomedical domain more work is being done in finding entities relationships for example protein-protein or disease-gene interactions [14] [15] [16]. Commonly, there are two sides of relation extraction, either using interaction based methods or using text mining methods. For interaction methods the databases available are BIND [17], IntAct [18], BIOGRID [19], Strings [20] and MINT [21]. Although these databases contain a large collection of manually extracted

relations that exists between the entities, it still requires a considerable amount of time and effort because of the rapid expansion in literature. On the other hand, text mining shows diversity in relation extraction starting from simple approaches like co-occurrence to complex systems like hybrid approaches, hence making it a popular method among researchers [22]. In this study we propose an automated method " ZK DrugResist" for extracting drug response relations via text mining. This tool categorizes abstracts based on mutations, gene expression and disease types associated with drug resistance. Also this tool extracts semantic relations from the sentences for three drug response relations (Resistance, Intermediate, Susceptible). ZK DrugResist is freely available for non commercial use, it does not require any registration and can be accessed using Safari, Google Chrome and Firefox web browsers. ZK DrugResist requires user to pick any disease category for input. The server will display the results page for the selected disease type summarizing drug resistance.

1.3 Combining Sequence and Structural Features for Predicting Drug Resistance

Aids is an epidemic disease spreading worldwide since early 1980s which occurs due to HIV infection. Even today there is no effective treatment that provides a complete remedy for AIDS [23]. Commonly it is treated by providing a combination of drugs termed highly active antiretroviral therapy (HAART), which controls the virus transmission, hence increasing the survival chances. As reported in earlier studies, the two important drug targets for HIV infected patients are protease inhibitors and reverse transcriptase. Among these, four falls in the nucleoside RT inhibitors (NRTIs) category, three in the non-nucleoside RT inhibitors (NNRTIs) and seven in the PR inhibitors (PIs) [24]. Despite this fact, the HIV therapy is restricted to the emerging drug resistance phenomena. The reason behind these phenomena is either the protein mutation or the changes in gene expression level that induces drug resistance [7] [6]. These mutations either alter the residues at binding site or at the distal regions hence affecting the drug binding activity which results in failure of treatment [25], [26]. Therefore, it is necessary to conduct resistance testing in order to carry out HIV effective therapy. Previously, both phenotypic and genotypic analysis was used to measure HIV drug resistance. Phenotyping is more difficult to be performed as it is time consuming and laborious, while genotyping on the other hand is faster, but there remain

some challenges in predicting drug resistance from genotypic data [27]. However, due to the acquisition of mutations and multiple mutation patterns for drug resistance, it is difficult to associate genotypic with phenotypic methods.

Many computational methods have already been developed for analyzing drug resistance using genotypic data. These methods are either sequence based or structure based. For sequence based, various studies have been reported that use the statistical and supervised learning approaches to evaluate resistant mutation sequences. For instance, for HIV these methods usually work by taking the protein sequences of Protease PI and reverse transcriptase RT. Further on the basis of fold value they are assigned as resistant, intermediate or susceptible. Sequence methods require training set of sufficient size for predicting resistance with higher accuracy. Despite of the efficiency of these methods, they are unable to predict the new inhibitors as the same training data cannot be applicable for training other predictors. For HIV these methods usually work by taking the protein sequences of Protease PI and reverse transcriptase RT. Further on the basis of fold value they are assigned as resistant, intermediate or susceptible. The structure based methods normally use the 3D representation to calculate binding energies between protease and inhibitors. Although no large training set is required, still the performance measure is compromised because of the noise in calculating this free energy [28].

To overcome the limitations of both methods, one can combine the two approaches that can help in predicting resistance before testing the drug clinically. We examined HIV resistance against protease and reverse transcriptase drug treatments. The data available on Stanford HIV database has been used for seven PIs, four NRTIs, and three NNRTIs. Additionally for the two drugs of PIs: Indinavir (IDV) and Saquinavir (SQV) we also used the datasets reported by Dragchi and group [29]. Using SVM and Random forests classifiers we examined both single and multiple mutations of HIV resistance and also inferred the interactions among them.

Our goal is to look for the best features for determining HIV resistance. Unlike the linear sequence representation, we combined both sequence and structure features implying a qualitative depiction of a feature set. This feature set includes hydrophobicity measure, evolutionary conservation, flexibility measure, frequency occurrence count, solvent accessibility, disordered proteins and amino acid volume information as sequence features. For structural features, the 2D and 3D structure representation along with the contact energies of the interacting residues followed by average hydrophobicity and average volume were used. Along with single point mutations it is equally important to look for the combina-

tion of multiple mutations for HIV resistance. Therefore, we calculated the total contact energies between these multiple mutations for every instance that in turns helps to infer the impact of multiple interactions of HIV mutants. This gives a better understanding of how these multiple mutations bring resistance in response to certain drug treatments. Our study shows the possibility of using combinatorial approaches with optimized features to combat HIV resistance.

1.4 Drug Repurposing

Growth in drug research and development has been dropping down for few years as it is getting real expensive with chances of failure attempts. Pharmaceutical companies are unsuccessful in keeping pace to bringing new drugs to the market, reasons behind this are manifold, broadly the safety and efficacy factors are coupled with the overpricing. To cope with this dilemma, the theory of reusing the existing drugs gained a lot of thoughtfulness among bio-pharmaceutics. The phenomena of using the old drugs with new indications paired with the original indications is termed as repurposing or repositioning. The concept of drug repurposing causing a huge cut down in terms of price in drug discovery domain as it costs almost half the price of a new drug causing an overhaul in drug development. Before carrying out drug repositioning, few aspects should be given a look and these include drug side effects and its relevancy with the disease. Since the approved drugs have already been passed through various validation steps which include target identification and ADMET (absorption, distribution, metabolism, excretion and toxicity) characteristics, thus can facilitate in identifying new uses for the same drug [30].

Among various approaches for finding new uses of old drugs, using GWAS data, gene expression data and structural features are considered as the most distinguished methods. A drug can bind to several different targets making it promiscuous in nature. This is a well-known phenomenon which is also considered as a significant factor for the efficacy of drugs. Drug promiscuity reveals the drug off-targets, thus making them prime candidates for drug repurposing. Among the in-silico approaches for finding drug targets, binding site structure similarity from structural bioinformatics domain is highly contributing. The proteins sharing similar binding sites tend to bind same ligands hence paving ways for drug repositioning. Previous studies reported that flexibility is also one of the physiochemical properties of ligand contributing in drug promiscuity. With sharing similar binding sites,

the ligand flexibility is similarly influencing drug binding to multiple targets, hence these features are staple for drug repositioning [31] [32].

In addition to that, recent studies have also worked on analyzing drug associated biological pathways. These are beneficial in exploring mechanism of action of drugs and also the upstream or downstream genes in a pathway [33]. It is important to look for the pathway associated genes, as there are possibilities if the drug is not directly binding to its target but the target gene is interacting with other genes in a pathway hence binding to the ligand. Among drugs retrieved biological pathways, some of the drug targets share common pathways which in turn helps in revelation of clinical functions along with the drugs mechanism of action. Also these pathways might have associated with some other diseases than those they were initially used for, hence providing a substitute for drug repurposing.

1.5 Motivation

Drug Resistance is the major bottleneck in treating complex diseases making it the limelight of current research. Over the decades much work has been conducted both experimentally and computationally to understand drug resistance mechanisms. Machine learning has widely been applied for studying drug resistance. Our motivation is to develop methods for understanding drug resistance by applying machine learning techniques. The aim of the thesis is as follows

- To provide a disease categorized database that contains a collection of all the literature summarizing the association of drug resistance with point mutations and expression changes also including the drug response relations (resistance, intermediate , susceptible)
- Second is to integrate sequence and structural features for predicting HIV resistance.
- Third to propose a drug repurposing strategy by combining PPI, biological pathways, binding site structure similarity and disease-disease similarity unlike one drug one target models.

1.6 Structure of the Thesis

This thesis is organized into five chapters. Each chapter has three separate sections, Text Mining, Drug Resistance Prediction and Drug Repurposing.

Chapter 2: Provides a comprehensive background of computational analysis of drug resistance. The approaches developed to predict drug resistance including Text Mining, Sequence and Structure Features and Drug repositioning are discussed.

Chapter 3 : A design of model developed to study drug resistance mechanisms is described in detail. This chapter covers the features taken for building a model in order to establish a database , second the features for building a classification model for predicting HIV resistance. And last section provides in detail the algorithm proposed for drug repurposing.

Chapter 4: Provides in detail the results obtained.

Chapter 5: Discussion and conclusions are highlighted. Also possible ideas for extending this work are discussed.

1.7 Published Articles

The part of the thesis on disease categorized database and prediction of HIV resistance has been published [34] [35]

- Khalid, Zoya, and Osman Ugur Sezerman. "Prediction of HIV Drug Resistance by combining Sequence and Structural Properties." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2016).
- Khalid, Zoya, and Osman Ugur Sezerman. "ZK DrugResist 2.0: A TextMiner to Extract Semantic Relations of Drug Resistance from PubMed." *Journal of Biomedical Informatics* (2017).

The part of thesis on designing an algorithm for drug repurposing is in preparation phase.

Chapter 2

Background

This chapter begins with the overview of previous work on drug resistance. The computational strategies that aimed to predict resistance and ways of combating drug resistance is summarized in this chapter.

2.1 Text Mining

Biomedical literature provides a rich source of information which provide a good start in establishing the state of the art in particular domain. PubMed database has the vast collection of biomedical literature that provides approximately 24 million articles from different journals including Medline and life science [36]. Biological terms such as genes and proteins has no fixed terminology, therefore interpreting this heterogeneous nature of the data needs to be automated to facilitate the researchers. This introduces Biological Natural Language Processing (BIONLP) as a light in disguise. The components of BIONLP are: Text classification, Named Entity Recognition, Relation Extraction and Relation Classification.

Much work has been performed in analyzing protein-protein interactions, but most of the databases developed for that purpose are manually curated ones. There is a high need of developing an automated method for the ease of extracting required information. The automation can be performed by using machine learning or rule based methods however, these methods require an applicable training set which varies from one domain to another. Many data mining techniques have been applied to biomedical data but the proteins, genes and drugs nomenclature heterogeneity complicates the process. Many authors suggested using context free grammars or specialized biomedical parsers [37] to tackle this problem.

Considering drug resistance, literature shows some already developed databases. One of

them is BacMat which is based on genetic alterations that are associated with antibiotics resistance [38]. One more database is Biozyne P-gp Predictor which is based on SVM classifier that differentiates the substrates from efflux pumps [?]. [39] reported a database GEAR which associates genomic elements like SNPs, microRNA with drug resistance by measuring the probabilities of co-occurrence.

Many techniques for relation extraction have been proposed lately that broadly include co-occurrence, rule based, machine learning and pattern-based. Co-occurrence as the simplest approach, provides high sensitivity but very low specificity as biomedical texts have complex sentences that make it difficult to find related words [40]. Secondly for the Rule-based approaches usually rules are defined manually by using features derived by applying natural language processing (NLP) chunking and parsing techniques. However, variations in biological terms make rule-based methods less accurate providing more precision but at the cost of lower recall [41]. Third approach works on building classifiers that use features derived from either shallow parsing or full dependency parsing. This method is the most preferable method as it provides good measures but it also requires fully annotated training sets of enough size [15]. Lastly pattern based methods need patterns to extract relations which provide low recall if they are manually generated ones. Bootstrapping can be performed for automatic patterns to improve recall measure [42].

All of the reported relation extraction methods generally worked on the protein-protein interaction or protein-gene interactions corpus [43] [44]. Most of the times, the other type of relations is overlooked by the researchers. For predicting HIV resistance there are two methods reported, one is rule-based while the other one centered on computational based approach [45] [46]. Furthermore, there are already developed databases like Stanford HIVDB, RegaDB and CancerDR which are updated regularly by experts. One study published in 2012 has analyzed the casual relations extraction focuses on HIV resistance caused by certain mutations. This study was based on Rule-based methods. The results showed the precision of 97% on HIVDB dataset while on PubMed abstracts it is 87% [47]. Another method named EDGAR based on natural language processing tools focuses on extracting drug, gene and cell names from biomedical text along with extracting semantic relations with respect to gene expression and drug resistance. The authors have not mentioned any accuracy measure to be compared with. This method has an application called semantic Medline which is accessible at skr3.nlm.nih.gov [48].

2.2 Identifying Drug Resistance by combining sequence and structure features

As described earlier, predicting drug resistance has been divided into two categories: sequence based and structure based. For applying these methods there is a need of prior data to train a model for analyzing the effect of mutations on resistance. As the existing data provides information that helps in associating the mutations to the phenotype (resistant, susceptible). Literature is filled with HIV drug resistance providing the worth of training data as compared to other diseases, hence limits the sequence based approaches to only analyzing HIV resistance. The largest warehouse is the Stanford HIV drug database (HIVDB) containing genotypic sequence data with the phenotypes [49]. Generally, the gene sequence is the genotypic data while phenotypic data is measured by calculating the mutation effect on resistance. This resistance factor is defined as follows:

$$RF = \frac{MutantIC50}{WildtypeIC50}$$

For predicting phenotype from genotype data building regression models is one way of doing that. These are probabilistic models that can be obtained by applying least square regression, neural networks, decision trees and support vector machines as cited in [50]. In a recent study authors have used the combination of both linear and cross-validated regression for predicting HIV resistance. They have tested their model on reverse transcriptase and non-nucleoside reverse transcriptase (NNRTI) [51]. Using artificial neural networks for predicting HIV resistance is one preferred model. [52] used 7500 HIV sequences for training and testing their model. Furthermore, Random Forests and ANN were also employed to analyze bevirimat HIV resistance [53]. All these methods have reported higher degree of accuracy in predicting HIV resistance, but quality of a model cannot be determined only by considering accuracy measure. More True positives and less False positives should be examined to avoid misleading results.

Even if these learning methods are good enough to predict resistance, still it is difficult to determine which mutations confer resistance or in other words how many mutations are enough for causing resistance. For this multiple mutations patterns should be evaluated as reported in [54]. The authors used Bayesian models to generate probabilities for determining the impact of multiple mutations on drug resistance in HIV.

The above discussed methods used the sequence of protein or gene to build learning models. There is another category of sequence based methods that uses chemical properties of the

proteins for predicting resistance. The sequence data is represented as physiochemical descriptors which are either hydrophilic, hydrophobic or soluble. The same could be applied for determining the protein ligand interactions. The striking outcome of using this, is the generalization of a model for instance, based on these properties the same training set could be applied to two inhibitors having similar chemical features. However, the model accuracy will compromise if the drugs aren't similar much [55] [56].

Moreover, with the computational learning methods, rule-based methods also contributed in predicting HIV resistance. These rules are defined by experts by analyzing clinical data for finding the impact of single or multiple mutations on resistance. Publicly available databases are Stanford HIV, HIV-grade, REGA, ANRS and Visible Genetics [57] [50] [58] [59].

On the contrary, the structure based methods use the drug ligand complex, docking, molecular dynamics and molecular modeling. These approaches start with the wild type structure followed by introducing mutations and checking for the structural changes. The scoring function is built to compute how the mutant proteins are effecting resistance. One study used molecular modeling software SYBYL for determining that mutations are actually disrupting the binding site of the drug [60]. Similarly, protease inhibitor ampenavir was studied using molecular dynamics which shows the impact of double point mutations on resistance. Similarly, for Tuberculosis another study reported for ACP reductase indicating that S94A mutation confer resistance to isoniazid [61]. Furthermore, multidrug HIV resistance was also conducted using molecular dynamics simulations as cited by [62]. One more study [63] studied mutation at residue 50 with binding to two drugs atazanavir and amprenavir. Mutation of I50V for atazanavir and I50L to amprenavir causes decrease in binding affinity hence results in HIV protease resistance.

Various other studies for analyzing these resistant mutations were examined by applying statistical and machine learning methods including artificial neural networks, SVM, Random Forests, decision trees and Regression Analysis to examine the relationship of genotype and phenotype [64] [65] [66]. All these published methods usually relied on the genotypic data hence used non-parametric methods. Previously for Saquinavir (SQV) and Indinavir (INV) of PIs linear discriminant analysis and cluster analysis was performed to determine resistance mutations [67]. Similarly, for structure based analysis SVM and Random Forests learning models were tested for Nevirapine (NVP) of PIs drug treatment [68]. Moreover, Graph Theory techniques such as Delaunay Triangulation and Sparse Dictionary were also used for structure based analysis of resistance mutation patterns [69].

One advantage of structure based methods over sequence based methods is that, they don't

really depend on the availability of training data. Still it faced one limitation, the molecular docking and molecular simulations require predefined set of mutants. Hence these methods cannot be applied extensively for predicting structure based HIV resistance.

Table 2.1 summarizes the existing literature defining methods used and accuracy measured.

Table 2.1: Comparison of Accuracy Measure with the Existing Literature

Types of Features	Technique	Reported Accuracy	Reference
Sequence Feature: Hydrophobicity	Molecular Dynamic Simulation	NA	[70]
Sequence Feature: Conserved Residues	Molecular Dynamic Simulation	NA	[71]
Structure Features : Molecular Dynamic simulations	Molecular Dynamic Simulation	NA	[72]
Sequence Features : Frequencies of Occurrence	Neural Networks	85%	[29]
Structure Features :Ligand Protein binding complex	Neural Networks	85%	[29]
Combined Sequence and Structure representation	Sparse Dictionary	85-97%	[69]
Structure Features : Contact Energies	Support Vector Machines	83-91 %	[73]
Structure Features: Interaction among Multiple Mutations	Bayesian Variable partition	NA	[74]
Sequence and Structure	Random Forests	80-94%	[75]
	Support Vector Machines	65-87%	
Sequence Features	Decision Trees	77-89%	[50]
	Neural Networks		
	Support Vector Machines		
	Regression		
Combined sequence-structure: Delaunay tessellation	Support Vector Machines	90-93%	[76]
	Random Forests		
Combined sequence-structure: Delaunay tessellation	Support vector Machines	88%	[77]
Sequence Features: Frequencies of Occurrence	Regression Model	76-80%	[78]

2.3 Drug Repurposing

Drug repositioning is the idea of exploring new therapeutic uses of old drugs, hence contributing in saving money from bringing out a new drug to the market. Various studies have been reported on drug repositioning making it a hot topic for the researchers. The classical method of drug discovery is one drug one target model which has expected to provide less efficacy and with more side effects. This model doesn't consider the biological mechanisms that makes drug to bind with more than one target hence limiting the efficiency of this model [79] [80].

To overcome these limitations much work has been done on computational and network based approaches that provides a new direction towards drug repositioning. In one study the authors used chemical structures of drug and its target starting from this the network

will expand linking new indications [81]. [82] used pairwise similarity to conduct drug repositioning. The similarity measures include drug similarity, drug target similarity and target-target interaction. A drug-drug similarity network was proposed by [83] where the network has gene expression profiles as features. Many studies have been proposed lately that used the drug and disease expression profiles to provide plausible candidate for drug repositioning [84] [85] [86]. Further literature mining and pathway analysis was also proposed to build drug disease network [87] [88].

One study used the microarray gene expression data for finding drug-disease interactions. Their network contains disease-disease, disease-drug and drug-drug associations that provides insights about drug repositioning. The methodology was based on scoring system that calculates the similarity scores among the drug and disease pairs. Using this method, the authors have discovered many new indications for the approved drugs [89].

Another method used gene expression profiles to check the effect of drug on various treatments. The network contains those nodes that either have similar mechanism of action or targeting same biological pathway. This network was developed on consensus transcriptional response which shows the transcriptional activity of a drug towards drug treatments. This approach helps in capturing the similarities and differences in drug responses, hence is useful for drug repositioning [90].

For creating a disease-drug, disease-disease and drug-drug network Guanghai and Agarwal used gene expression profiles, they have used two approaches namely correlation and enrichment. Correlation takes in profile-profile similarity while enrichment measures signature-profile similarity. This helps in identifying novel relations among drug and disease hence can be used for repositioning [91].

[92] used both efficacy and side effects measure for drug repurposing. The gene regulation has been observed before and after drug treatment to measure drug efficacy and the number of essential genes and correlated genes were measured for side effects. Based on this a scoring scheme was developed to align drug-disease association for repurposing.

Another study performed drug repositioning without using the gene signatures. A scoring function was formulated to compute drug-gene-disease network, which takes in both the contribution of a gene and effect of a drug on a gene. Drug-disease association can be computed by measuring the similarity and dissimilarity of their gene expression profiles [93].

Various studies have already been conducted for pathways based analysis of drug repurposing. Creating a network from drugs-targets-pathways-gene-disease helps in interpreting mechanism of action and can contribute in drug repurposing [?] [94].

Machine learning is contributing much for building drug repositioning strategies. It combines multiple information including how similar their chemical structures are, closeness in a PPI network of drug targets and correlation among the gene expression patterns [95].

In addition to the computational approaches few studies are also devoted to manually analyze the drug associated pathways for drug repositioning [96] [97]. For instance, bexarotene which was used for cancer treatment can also be used for Alzheimers disease [98]. The manual curation has performed which is based on drug target, target associated pathways, transcriptional responses of pathway and the gene based analysis for understanding mechanism of disease.

Chapter 3

Materials and Methods

3.1 Text Mining

For the text mining first the abstracts were collected by querying PubMed with terms which include but not limited to "drug resistance, mutations", "expression changes" and "complex diseases". These abstracts were categorized based on four modules as described below.

3.1.1 Drug Resistance Vs. Others

The first stage is to separate the drug resistance abstracts from the rest. The total abstracts are 15,580. For each of the downloaded abstract the feature vector is constructed. In order to distinguish them the frequency of each word is counted as a feature value. These words were then further processed using tokenization and porter stemming algorithms. After breaking the abstract into words the term frequency of words was counted by using term "drug resistance" and calculates its total occurrence in an abstract. Next calculate the Term document inverse frequency TFIDF, it is the weight calculated by taking number of documents in the corpus by the number of documents containing term. This module will separate drug resistance abstracts from the rest.

3.1.2 Mutation Vs. Expressions

In the second category we picked these drug resistance documents and scanned them to associate either with mutation or with expression level changes. The documents cited like over-expression, down regulation kind of terms are marked as expression abstracts while the others that are cited by protein mutations are marked as Mutation abstracts.

3.1.3 Protein Vs. DNA

To extract mutations associated with drug resistance we used Perl Regular Expressions. Protein mutations are ambiguous there are chances they might have mixed up with the DNA mutations like the one letter code amino acid mutation A456G can easily be misinterpreted with the nucleotide letters. We worked on this ambiguity with regular expressions which classified the protein mutations and DNA mutations separately.

3.1.4 Cancer Vs. Others

The last step categorizes the association of drug resistance with the complex diseases like Cancer and others which include but not limited to neuro-degenerative, autoimmune and metabolic disorders. This gives us 5,965 abstracts reported on cancer resistance, while the others 9,615 are on HIV resistance. For cancer the abstracts containing mutations are 1,224 while from HIV it is 5,615. Additionally, there are 520 abstracts which contained expression changes with drug resistance.

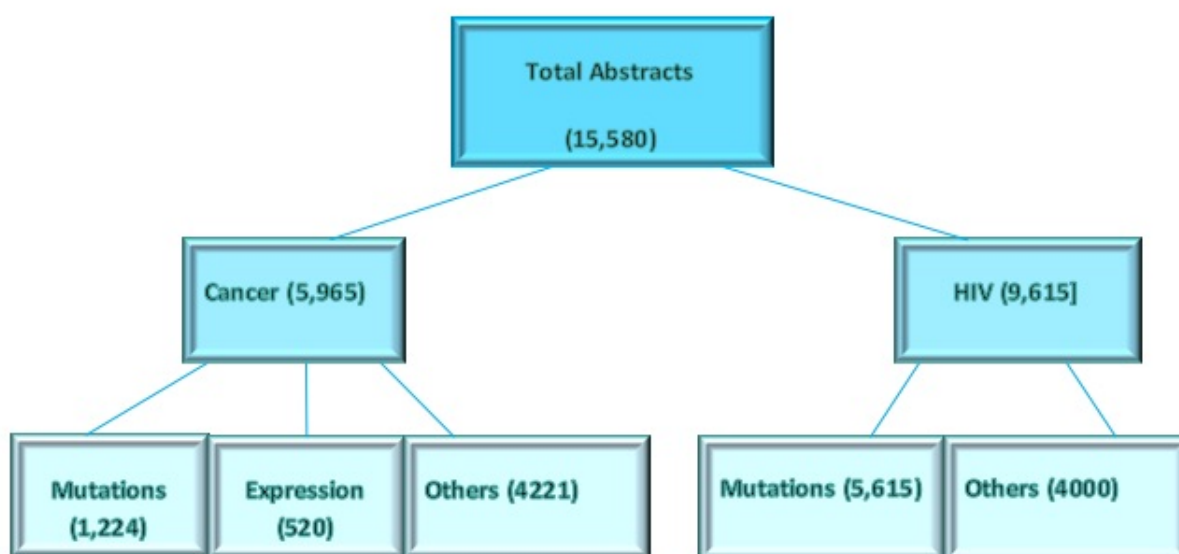


Figure 3.1: Modules for extracting Abstracts

3.1.5 Relation Extraction

For extracting relations first, we collected sentences containing drug name, mutation type or gene expression changes together in a sentence. This gives us 4000 sentences in total. Along with the identification of mutation, drugs and gene names our system also searched for relation words and predicates. We created a list of these words as shown in Table 4. These words were manually created for the ease of parsing the sentences later. Each sentence is looked for the pattern <mutation, relation/predicate, drug/disease/gene >. For parsing we used Stanford parser that generates the output in the form of pen Treebank.

3.1.6 Features and Model Building

This section describes the features set for training and testing our model, which in our case is the hybrid feature set.

3.1.6.1 Vector representation

We have used Bag of words (BOW) model for document representation in vector form. This model uses term frequency count as one of the features which counts the occurrence of each term in a sentence and assigns score. Second task is to assign weights to the features, for this Term frequency inverse document frequency (TFIDF) was used. It is the product of two statistics Term Frequency and Inverse Document Frequency, term frequency deals with the raw calculation of a term in a document while inverse document frequency deals with the significance of a word count. We used a Perl module TEXT: TFIDF for this.

3.1.6.2 NLP and Concept Ranking

For this part variety of text processing algorithms were used namely, part of speech tagging (POS) and Natural Language Processing (NLP) for noun and verb phrases identification. Part of speech tag could be a noun, verb or adjective, this tagger provide annotation to the text in the dataset. Specifically, for NLP the Perl module GeniaTagger was used to extract syntactical information from the text. This online system takes in biomedical data as input and performs chunking. The verb and noun chunks obtained were considered as syntactical features for further classification.

Next step is to rank noun and verb concepts obtained in the previous step by applying a mapping function reported in [99]. This ranking is important to filter out the irrelevant

features from the relevant ones as it provides more meaningful information about the relation words hence helps in finding related concepts for relation extraction. UMLS has three knowledge sources, Metathesaurus, Semantic Network, and the Specialist Lexicon. MetaMap uses Metathesaurus for mapping named entities like drug, gene name with more accuracy to the medical concepts as each concept is associated with a semantic type like Pharmacological substance, Gene or Genome? etc. Also, it assigns scores to each medical concept for each input sentence which were further used for ranking by applying the concept ranking algorithm [99]. This algorithm worked by extracting related concepts from UMLS by setting two different thresholds for noun and verb. For noun the scores obtained from MetaMap should exceed 600 while for verb this number is 700. The concepts include Therapeutic or Preventive procedures, Functional concept and so on details are referenced [44]. The concepts that passed this threshold criterion were filtered out as features for our feature set.

3.1.6.3 Name Entity Recognition

The gene names following the protein mutation are also extracted from the abstracts. For this purpose the complete list of official gene names was downloaded from HUGO database <http://www.genenames.org/>. Any gene name mentioned in the abstract is programmed to match with the list of the genes stored and the results are displayed on web. We followed MugeX approach for this module.

3.1.6.4 Relation Classification

This component deals with classifying extracted relations from the candidate sentences. Three relations were focused namely resistance, intermediate and susceptible which were mentioned in text as high, low or reduced resistance. Hence class labels are Resistance (R), Intermediate (I) and Susceptible (S). We used Support Vector Machines (SVM) for performing classification for the three classes.

3.1.6.5 Support Vector Machines (SVM)

Support vector machines are supervised learning algorithms that optimize feature vector to separate classes by using a margin it constructs hyperplane in high dimensional space. As in our case the relation classification is multiclass problem rather than binary, hence LibSVM Perl implementation was used [100]. Moreover, as the number of features set is larger so

linear kernel was preferred with $c=0.5$. For training our model we have used features derived from the dataset of 15,580 abstracts, while for testing we used complete independent set of 500 abstracts. We used separate training and test sets for evaluating the performance of our SVM model. From 4000 sentences, 3000 candidate sentences are for training the model for three class Relation classification while for testing 1000 sentences were selected as test dataset.

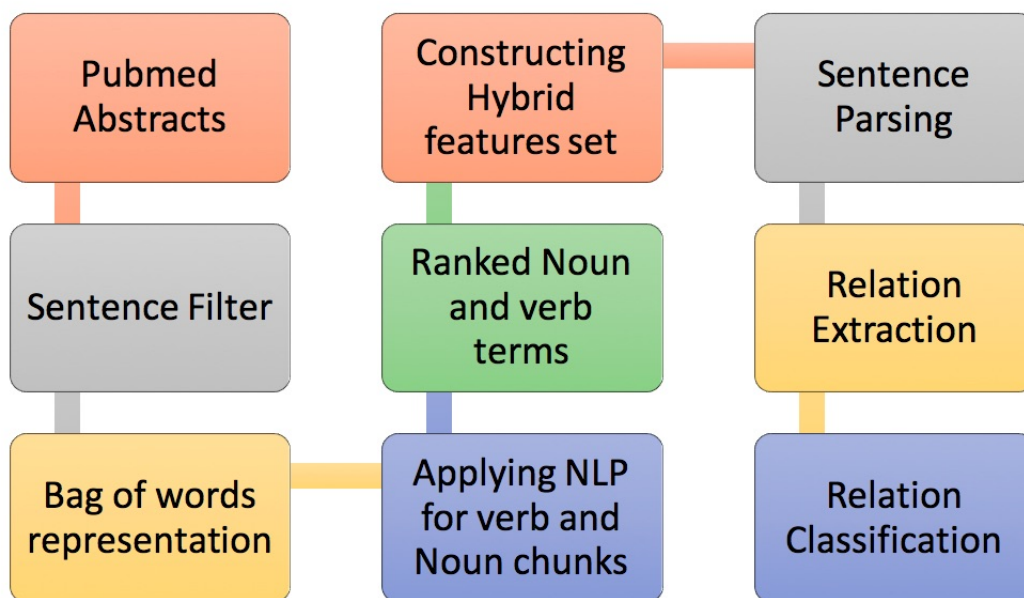


Figure 3.2: Modules for Relation Extraction and Relation Classification

3.2 Implementation and Usage

We named our tool as ZK DrugResist which is web based implemented in Strawberry Perl, Python and PHP. The MySQL database was built on XAMPP Server. The Web designing was performed by using WordPress, further CGI, DBI and DBD Perl modules was used to connect MySQL with the web interface. The core model of this tool was SVM model. ZK DrugResist is freely available for non commercial use, it does not require any registration and can be accessed using safari, google chrome and Firefox web browsers. ZK DrugResist requires user to pick any disease category for input. The server will display the results page for the selected disease type summarizing drug resistance.

3.3 Sequence and Structural Features

The amino acid sequences of HIV-1 Protease and Reverse transcriptase was downloaded from Stanford HIVdb. The thirty most common mutations were identified and labeled in both of the sequences. These labeled mutations were represented as a sequence window with the mutation residue at the center followed by n residues at c terminal and n residues at n terminal. For the optimal sequence window length, we tried different lengths from 5 to 21 by measuring classification accuracy.

3.3.1 Frequency count

Each position in a sequence window is represented by a 20 dimensional vector representing 20 amino acid residues. Only one dimension may contain a value of 1 the rest are zeroes. Furthermore, the amino acids were also grouped through Sezerman grouping method [101] which makes an 11 dimensional vector for each position. In addition to that we have also added frequency count of each residue in a sequence as a separate feature. The frequency measure of each amino acid is calculated as total frequency of amino acids in a sequence window over the average occurrence in overall protein sequence. Again in this case each position is represented by 20 dimensional and 11 dimensional vectors for grouped and non-grouped amino acids.

3.3.2 Conserved Mutations using PSSM

Evolutionary conserved sequences are important as of the fact that these are very critical to the function of the protein. Mutations in these conserved regions might affect the functioning of the protein. In order to evaluate this, we used PSI-BLAST which helps in identifying the conserved residues by assigning them scores. We then compared the effect of mutation on the conserved regions. PSI-BLAST program from BLAST+ toolkit (version 2.2.26+) with three iterations (*num_iterations*3), and inclusion e-value threshold of $1e - 5$ (*-inclusion_ethresh* $1e - 5$) were used. The higher the score, more conserved is the residue.

3.3.3 Measuring Flexibility and Rigidity

Measuring the flexibility and rigidity of the amino acid residues before and after mutation is an important feature for drug resistance mutations. We used an online server FlexPred to

measure this feature. This server is freely available at <http://flexpred.rit.albany.edu/> [102] The binary feature value has been generated for flexibility and rigidity measures. Value of 1 is assigned if the overall window sequence shows flexibility measure, 0 otherwise. This server predicts the residue positions which are involved in conformational switches of a protein. Each residue has been assigned label flexible or rigid based on probability values generated. Higher the probability, more confident is the prediction.

3.3.4 Disordered Regions

Unstructured regions also called as disordered are those proteins which dont have any proper well defined structure. IUPred server was used to predict these regions in both HIV-1 protease and RT sequences [103]. This server takes in the amino acid sequence and predicts the disordered regions based on the pairwise contact energy as these proteins do not have many interactions to form hence no stable structure. The information obtained from IUPred server was converted into two kinds of feature vectors. One is based on the overall disordered tendency in a sequence window while second one relies on a threshold value of 0.5 hence creating binary feature vector. If the average sequence window disordered value is greater than 0.5 the feature vector is assigned a value of 1 and in the other case its 0.

3.3.5 Hydrophobicity Measure

Hydrophobicity scales are the measures of hydrophobicity for amino acid residues. More positive the value is more hydrophobic the residue is. The mutation affects the hydrophobicity scale hence disrupting the overall structure. We used Hopp & Woods hydrophobicity scales to estimate average hydrophobicity of sequence window [104].

3.3.6 Volume Measure

The smaller volume residue mutating into larger sized amino acid residue might change the overall structure of the protein. Hence it is important to find out volume of the sequence window before and after mutation. We used Kharakoz's estimated amino acid volumes [105]. The volume of each sub window added two different features to the feature vector.

3.3.7 Secondary Structure Features

The secondary structure of HIV1 protease and reverse transcriptase was predicted by PSIPRED web server freely available at <http://bioinf.cs.ucl.ac.uk/psipred/> [106]. It creates a 9 dimensional vector with three values for each position of a sequence window. There are three features in secondary structure coils, helices and beta sheets (C H E). For each sequence window either 0 or 1 has been assigned based on which secondary structure feature is present in particular sequence window.

3.3.8 Solvent Accessibility

Those residues which are accessible to the solvent if mutated might cause stability changes in protein. To estimate this feature we used WESA tool which is freely available online <http://pipe.scs.fsu.edu/wesa.html>. Again a binary feature vector was created for solvent accessibility feature.

3.3.9 Structure and contact residues

The 3D structure of HIV-1 protease and RT was downloaded from PDB RSCB. The RSCB Ligand explorer was used in order to visualize the contacts between the inhibitor and the drug. Two type of contacts Hydrogen bonds (HBs) and non-bonding interactions (NBIs) were analyzed. Each of the contacts was considered separately hence giving two different features to a feature vector. The better way to represent a protein 3D structure is to analyze the protein contact map which shows the pairwise distance between protein residues. This measure is important to infer relationship between two residues which are key factors for protein structure prediction. We used RaptorX Contact Predict online server freely accessible at <http://raptorx.uchicago.edu/ContactMap/> [107]. The threshold between the contact atom residues was 7\AA . The value is assigned as 1 if the two residues are in contact, 0 otherwise.

3.3.10 Interactions between multiple mutations

The dataset we are using for this study shows that there are also multiple mutations making drug insensitive to treatment. In order to capture the combination of multiple mutations and the interaction between them we calculated the total contact energies which determines if these mutations are interacting or not. This feature was also used in one of the reported

studies to determine impact of interaction on drug resistance [74] [54]. We used the Amino Acid empirical contact energies published in an earlier study [108]. Moreover, we also calculated the average volume and average hydrophobicity for these combinatorial mutation patterns.

3.4 Preprocessing Filters and Feature Selection

For feature selection we used minimum redundancy and maximum relevance algorithm (mRMR) [109]. Its a two-step process, first it orders the features based on minimum redundancy and maximum relevance, later these ordered feature list is used for further analysis. For our data we used mRMR algorithm with a discretization threshold of 1 with rest of the parameters left with default settings. We did the incremental feature selection to reduce the dimensionality of features. We added the highest score features to the lowest one and tested the performance of the classifier. Average area under the curve (AUC) was used as performance evaluation measure to find the optimal feature set [110]

Our dataset is highly unbalanced; to avoid the biased output from classifier we have applied two filters. One is for over-sampling the minority class and the second is for under-sampling of majority class.

3.4.1 SMOTE and SpreadSubsample Filters

For over-sampling we have applied Synthetic Minority Over-Sampling Technique (SMOTE) filter. Applying only over-sampling techniques to imbalanced data often results in overfitting. To avoid this, we combined over-sampling of minority class with the under-sampling of majority class which results in achieving better classifier performance. For under-sampling we used SpreadSubsample filter [111]. For the two drugs IDV and SQV of PIs we trained our model on the datasets reported in [112] [113] and performed the model testing on the dataset present at Stanford HIVdb for these two drugs. This was performed to analyze our model on completely independent test set containing new mutation positions with few of them overlapping with the training set. While for the rest of the five drugs of PIs and seven drugs of RTs the training and testing were performed on the datasets available on Stanford HIVdb by dividing it as 20% for test set and rest as training set. We used two classes for our case: Resistance and Non-Resistance as multiclass classification is difficult with SVM and Random Forests, both resistance and intermediate were considered as resistance class

while susceptible is non-resistance [69].

3.5 Model Building

Two classifiers Support Vector Machines and Random Forests are used to predict HIV drug resistance.

3.5.1 Support Vector Machines

Support vector machines are supervised learning algorithms that optimize feature vector to separate classes by using a margin, it constructs hyperplane in high dimensional space. LibSVM implementation via wrappers method in scikit-learn python module was used [100] [114]. The parameters were tweaked in order to get the best possible classification accuracy. One of the limitations of SVM is its sensitivity to class imbalance, if the data is biased the chances of misclassification increase. In order to overcome this, we used sampling techniques on our dataset before applying SVM as described above.

3.5.2 Random Forests

Random Forests generally are ensemble learning methods that are used for classification and regression tasks. It works by bootstrapping from the training set. The python scikit-learn module was used to apply random forests on our dataset. The performances of the models were assessed by three basic measures namely sensitivity (sn), specificity (sp) and accuracy (Acc) respectively. For training set accuracy, we used 10-fold cross validation. In addition to that the classification performance was further analyzed by ROC analysis. One of the benefits of using ROC is that it is insensitive to class imbalance, as described earlier our dataset has skewed class distribution that might result in biased output, ROC is a good evaluation measure in this case as it is indifferent to this problem.

3.6 Drug Repurposing

We have proposed a drug prioritization algorithm to reposition drugs by using benchmarked dataset reported in [115]. We developed a ranking algorithm to find diseases that a drug can be repurposed for. We trained our method on 60 FDA approved drugs with their associated

disease indications. Further for testing the proposed model, we used independent test set of 8 drugs with the same attributes [116].

3.6.1 Computing Drug Disease Network

Our aim is to start with the drug, its target and old disease indication as a seed value and ends up finding the new/repurposed disease labels. We extracted drug targets from DrugBank, comparative Toxicogenomics database (CTD) and Therapeutic Target Database (TTD), target involved pathways from Pathway commons and KEGG. To look for common pathways among targets we used GenesLikeMe from GeneCards. It works by finding the shared pathways with the query gene by assigning weighted score. Our threshold was 1.0 in this case. For finding binding site structural similarity we used online tool PROBIS. This server takes in PDB structures as query proteins and compare it with 42270 structures available in the database which shares similar binding sites. The threshold of 1 was selected which filtered the significant similarity scores from the non- significant ones. The tool is freely available at <http://probis.cmm.ki.si/>. Binding sites can be similar in two proteins, a ligand binding to one protein can bind to another protein sharing the similar binding site that was not binding with this ligand at first place.

For checking disease- disease similarity we used a web-server DisGeNET available at <http://www.disgenet.org/web/DisGeNET/menu/search?0>. This tool calculates similarities between two diseases based on the number of shared genes among them. Furthermore, we also used CTD web-server for this feature. This server has a a module called Analyze that also compares the two diseases based on the common genes the two diseases have. Score for this feature was calculated by the following formula :

$$Score = \frac{X - Min}{Max - Min} \times 5$$

3.6.2 Building a Scoring System

Our scoring system takes in gene-based similarity measures and disease-disease similarity measures together. The algorithm works by computing weights for three gene-based similarity measures and one disease-disease similarity measures. For drugs associated with more than one gene, we looked for the highly cited association between the two in PMID. Details are as follows.

Algorithm: Drug-Disease Ranking based on similarity measures

Input: Drug Dr, Disease O, Drug Target Dt,

Output: Similarity Scores S, disease list.

1. Initialize a string array of size 5 for storing proteins Prot [5]
2. Initialize integer array of size 5 for storing scores of proteins protScore[5]
3. Initialize counter1 and counter2 =0
4. Get proteins which are 5 nodes away in a PPI network from the Target protein Dt, loop counter1 and store in Prot [5]
5. Loop counter2 and fill array by assigning protScore[]= 5,4,3,2,1, closest protein being the highest
6. **while** *protein counter is less than or equal to 5* **do**
 - | Check common pathways between Dt and Prot[]
 - | **if** *there is a common pathway* **then**
 - | Add pathway score to protScore[] for the proteins in Prot []
 - | Update Prot[]
 - | **end**
- end**
7. **for** Prot[] **do**
 - | Check for the binding site structural similarity **if** *similarity exists* **then**
 - | Add +5 to protScore[] and update Prot[]
 - | **end**
- end**
8. **for** Prot[] **do**
 - | Extract associated diseases add in diseaseN[]
 - | Look for disease similarity between diseaseO[] and diseaseN[]
 - | Update protScore[]
- end**
9. **if** *protScore[]* > 5 **then**
 - | Repeat checking binding site structural similarity with downstream and upstream proteins of a pathway.
- end**
10. **Print** protScore[], diseaseD[] and Prot[]

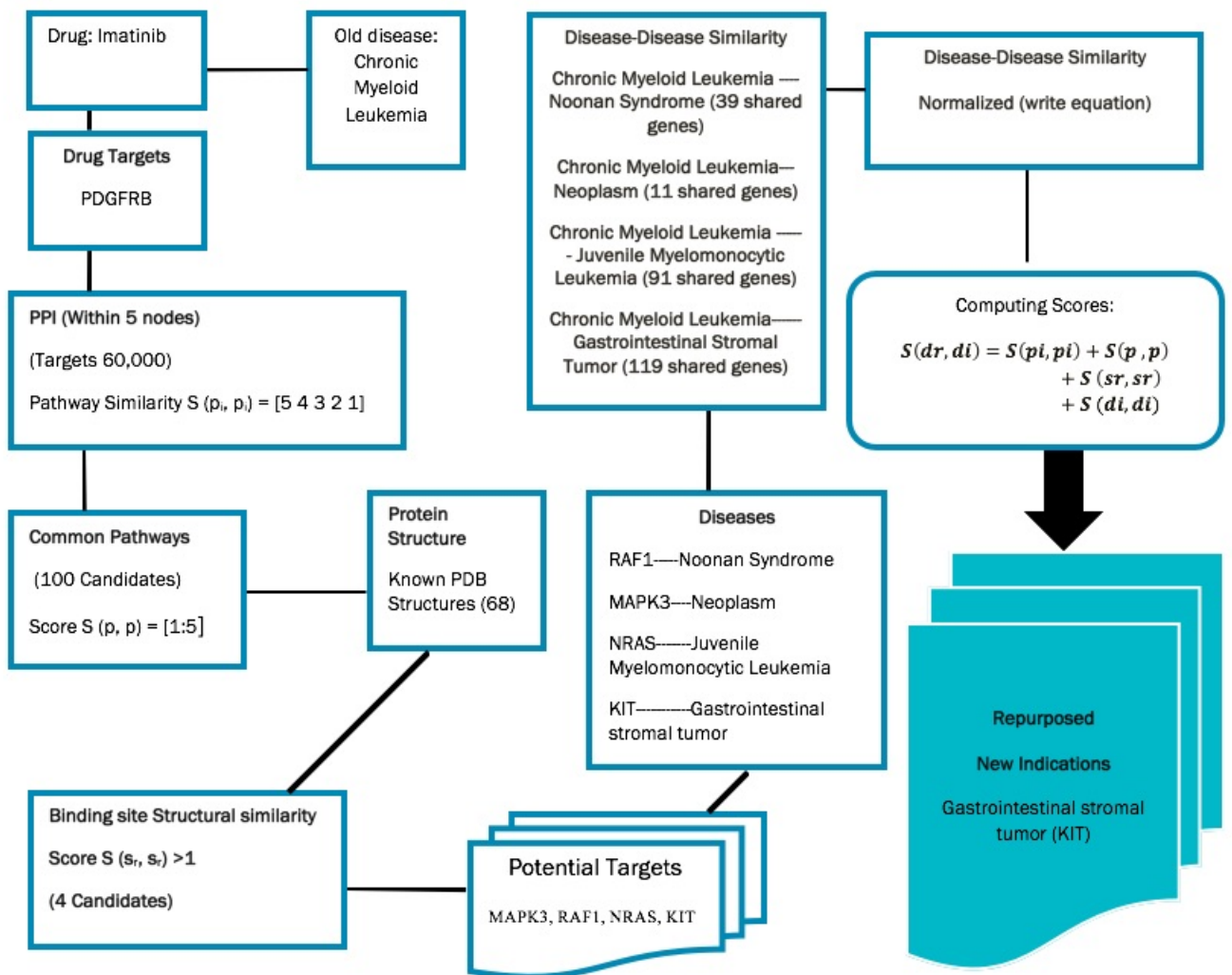


Figure 3.3: Workflow of Drug Repurposing Methodology

Chapter 4

Results

In this chapter we will present our database ZK DrugResist followed by the analysis carried out to predict HIV resistance by combining sequence and structure features and also how our proposed algorithm works on drug repurposing data.

4.1 Text Mining

We applied our SVM classification model on the corpus obtained from PubMed. There were 5,965 abstracts reported on cancer resistance, while the others 9,651 are on HIV resistance. For cancer the abstracts containing mutations are 1,224 while from HIV it is 5,615. Additionally, there are 520 abstracts on cancer which contained expression changes with drug resistance. The model generalization abilities were tested with 10-fold cross validation and we obtained 97% accuracy for the training set. For testing we have supplied completely independent set of 500 abstracts and have obtained 97.9% accuracy measure. Furthermore, for the mutation extraction from texts we used Perl regular expression library and it is working with 100% accuracy on the corpus. The results are tabulated in Table 4.3 shows some regular expressions of point mutations observed in PubMed abstracts.

Relation extraction was tested for three relations namely Resistance (R), Intermediate (I) and Susceptible (S). These three relations were used because these were found as most momentous relations obtained from the textual data. The part of our dataset was proposed by [47] which has already been used for relation extraction tasks but with a different approach. Specifically, for this module we have proposed hybrid features set that combines bag of words representation, natural language processing techniques and semantic features from

Table 4.1: SVM classification on PubMed Abstracts

Evaluation Measures SVM	Accuracy	Precision	Recall	F-Measure
10-fold Cross validation	97.2%	0.96	0.97	0.98
Test Set	97.9%	0.97	0.97	0.98

UMLS MetaMap. The results showed 98.99% accuracy on the training set while the test set accuracy was 97.98%. Also we evaluated our model with precision and recall measure along with the accuracy which are tabulated in Table 4.2. Out of 3000 candidate sentences, 2600 sentences belong to Resistance class, 167 to Intermediate and 190 to Susceptible class, while rest 43 sentences were the ambiguous sentences. For the test set out of 1000, 845 labeled as resistance, 25 as Intermediate, 124 as susceptible and 6 were the ambiguous sentences. The results are tabulated in Table 4.4.

These results can be accessed from ZK DrugResist which is available at <http://zkdrugresist.sabanciuniv.edu/>. The server categorizes drug response relations based on disease type. Figures 4.1, 4.2 and 4.3 provides a snapshot of how ZK DrugResist works.

Table 4.2: Relation Extraction classification results

Evaluation Measures SVM	Accuracy	Precision	Recall	F-Measure
10-fold Cross validation	98.98%	0.97	0.97	0.98
Test Set	97.7%	0.96	0.96	0.96

Table 4.3: Regular Expressions

Mutations Regular Expressions
ALA128 -- >GLU
ARG56 to TRP
Valine to Glutamine at position 168
Cystine (122) to Methionine
L148K
ARG-145-- >MET
GLU45 by LEU
GLU45 with LEU
ALA-22 was replaced by GLY
HIS-to-VAL substitution at position 54
THY to CYS at residue 124
150(ALA)- - -MET
VAL-156 and MET-128 to CYS

Table 4.4: Relations from Sentences

Total Sentences	Resistance(R)	Intermediate (I)	Susceptible (S)	Ambiguous
3000	2600	167	190	43
1000	845	25	124	6



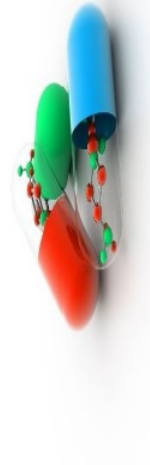
ZK DrugResist

Selected Disease: Cancer.

Results are below !.

Mutation/ Drug Receptor	PMID	Drug Name	Title	Disease Name
T315I/ Bcr	23666688		Detection of BCR-ABL1 kinase domain mutations causing imatinib resistance in chronic myelogenous leukemia	Cancer
A216V/RARA	25229938		Resistance to Therapy in Acute Promyelocytic Leukemia	Cancer
T790M/EGFR	24953979		Novel therapeutic strategies for patients with NSCLC that do not respond to treatment with EGFR inhibitors	Cancer
C65S/Apurinic	21865600		Knock-in reconstitution studies reveal an unexpected role of Cys-65 in regulating APE1/Ref-1 subcellular trafficking and function	Cancer
F584C F584L V654A L656P T670I R804W D816F D816V D816Y N822K Y823D E839K/KIT	20140688		Exploring the cause of drug resistance by the detrimental missense mutations in KIT receptor: computational approach	Cancer

Figure 4.1: ZK DrugResist snapshot of mutations associated with Cancer



ZK DrugResist

Selected Disease: HIV.

Results are below !.

Number	Mutation/Drug Receptor	PMID	Drug Name	Title	Disease Name
1	I50V	17971713	ATV	Efficacy and safety of atazanavir, with or without ritonavir, as part of once-daily highly active antiretroviral therapy regimens in antiretroviral-naïve patients	HIV
2	I50V	16127058	ATV	Atazanavir signature I50L resistance substitution accounts for unique phenotype of increased susceptibility to other protease inhibitors in a variety of human immunodeficiency virus type 1 genetic backbones	HIV
3	01V/F, 20R/M/I, 24I, 33I/F/V, 36I/L/V, 46I/L, 48V, 54V/L, 63P, 71V/T/I, 73C/S/T/A, 82A/F/S/T	16035256	ATV	Evaluation of atazanavir C trough, atazanavir genotypic inhibitory quotient, and baseline HIV genotype as predictors of a 24-week virological response in highly drug-experienced, HIV-infected patients treated with unboosted atazanavir	HIV

Figure 4.2: ZK DrugResist snapshot of mutations associated with HIV



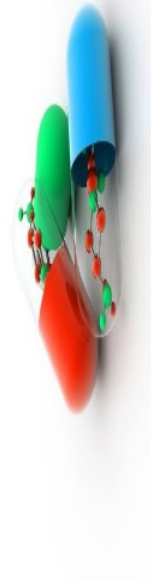
ZK DrugResist

Semantic Relations

Results are below !.

Sentence	Extracted Relation	Mutation	Relation Class
Fifteen different types of mutations (T315I, E255K, G250E, M351T, F359C, G251E, Y253H, V289F, E355G, N368S, L387M, H369R, A397P, E355A, D276G), including 2 novel mutations were identified, with T315I as the predominant type of mutation	T315T a predominant type of mutation	T315I, E255K, G250E, M351T, F359C, G251E, Y253H, V289F, E355G, N368S, L387M, H369R, A397P, E355A, D276G, T315T	NA
Acquired resistance to vemurafenib associated with reactivation of MAPK signaling as observed by elevated ERK1/2 phosphorylation levels in progressive lesions and the appearance of secondary NRAS(Q61) mutations or MEK1(Q56P) or MEK1(E203K) mutations	These mutations NRAS(Q61) MEK1(Q56P) or MEK1(E203K) causes acquired resistance to Vemurafenib	vemurafenib/ NRAS(Q61) MEK1(Q56P) MEK1(E203K)	Resistance (R)
All the subjects carrying I164T mutation showed some feature of metabolic syndrome, including hypertension, hyperlipidemia, diabetes, and atherosclerosis	I164T causes hypertension, hyperlipidemia, diabetes, and atherosclerosis	I164T	NA

Figure 4.3: Relation Extraction: Resistance(R) , Intermediate (I), Susceptible (S)



ZK DrugResist

Expression based Drug Resistance

Number	Gene/Inhibitor	PMID	Title	Disease Name
		27073322	YAP induces cisplatin resistance through activation of autophagy in human ovarian carcinoma cells	Cancer
01	EGFR	25871436	Cloning and Expression of a Novel Target Fusion Protein and its Application in Anti-Tumor Therapy	Cancer
02	EGFR, HER3, Met ERK1/2	25482142	Biomarkers for predicting response to tyrosine kinase inhibitors in drug-sensitive and drug-resistant human bladder cancer cells	Cancer
03	LMPTP	25834896	Allosteric Small Molecule Inhibitors of LMPTP	Obesity
04	ABCB1	25837780	Laurus nobilis L. Seed Extract Reveals Collateral Sensitivity in Multidrug-Resistant P-Glycoprotein-Expressing Tumor Cells	Cancer
05	Y-Box-binding protein 1	25750333	Cyclin A Correlates with YB1, Progression and Resistance to Chemotherapy in Human Epithelial Ovarian Cancer	Cancer
06	ALPP CALCOO1 CAV1 CYP1A2 IGFBP3	25710561	mRNA profiling reveals determinants of trastuzumab efficiency in HER2-positive breast cancer	Cancer

Figure 4.4: ZK DrugResist snapshot of expression changes associated with Cancer

4.2 Sequence and Structure Features

This section describes the prediction of HIV drug resistance by combining both sequence and structural properties through supervised machine learning techniques. The model was tested on PIs and RTs drug treatments for HIV resistance. For sequence features mining, first we have identified the optimal sequence window length by using the maximum features count. The classification results showed that the highest AUC obtained was at size 5, so the rest of the analysis was continued with this window length. Our feature set comprised of 13050 features, we reduced the features dimensionality by applying the mRMR algorithm. The incremental feature selection was employed to construct a feature set, initially started with the highest scoring ten features followed by adding more features until the classification accuracy falls or becomes steady. This gives us 400 features for PIs and 500 features for RTs, which include the frequency occurrences of both grouped and non-grouped amino acids at some positions, evolutionary conservations of specific positions, contact energies, average hydrophobicity, solvent accessibility and secondary structure features. We have noticed that this feature selection algorithm does not rank our flexibility measure count and disordered count as optimal features; hence, we removed it from our feature vector. This means that single point mutation is neither affecting the flexibility of amino acids nor the disordered predicted regions. Few of the selected features are mentioned in Table VI and pictorial representation is shown in Figures 4.1 and 4.2 that depicts classification accuracy on varying feature numbers and window sizes.

Few of the drugs of HIV resistance has class imbalance issue; dataset has more resistance classes as compared to the non-resistance ones. To tackle this, we applied two filters SMOTE and SpreadSubsample before applying classifiers as mentioned in Methods section.

Our classification models were trained for each drug separately. Two classifiers were tested; SVM and Random Forests, and results showed that SVM predicts HIV resistance with accuracy 98-99.2% compared to the Random Forests which provides 87-92% accuracy measures on all drugs. The results are tabulated in Tables 4.5,4.6, 4.7, 4.8, and 4.9.

Furthermore, we applied the 10-fold crossvalidation measure in order to check the generalization abilities of our model. The cross-validation accuracy we obtained by applying SVM was 95-96%, while with the Random Forests it was 82-83%. As accuracy alone is not a sufficient measure for classification, we evaluated our model with other parameters, including F-Measure, precision and Recall. We also plotted ROC in order to further validate the

performance measure of our models. ROC value of 0.985 generated with SVM classifier which showed that our model performs well on independent test sets.

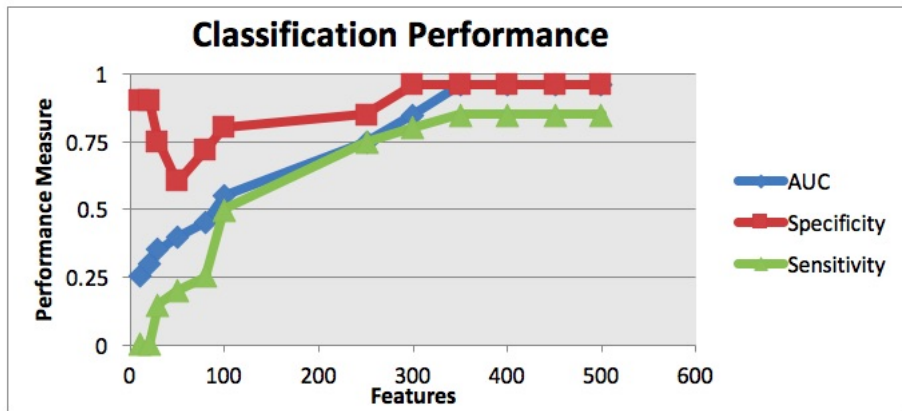


Figure 4.5: Classification Performance with varying Feature Numbers

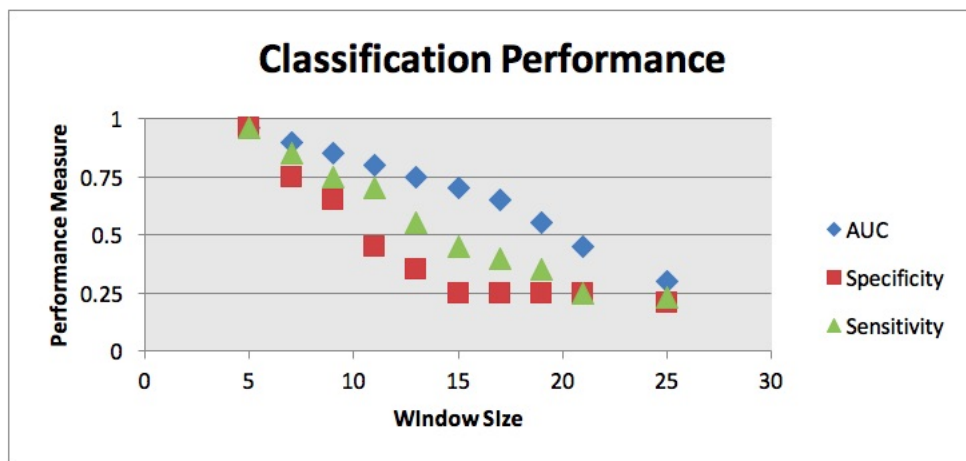


Figure 4.6: Classification Performance with Varying window size

Table 4.5: SVM Classification Results on IDV and SQV Drugs of PIs

Evaluation Measures SVM	Acc	Sn	Sp	Recall	FMeasure	ROC
IDV (Test Set)	98%	0.9	0.9	0.978	0.977	0.98
10-fold cross validation	96%	0.96	0.92	0.94	0.95	0.95
SQV (Test set)	97.65%	0.96	0.96	0.96	0.96	0.98
10-fold cross validation	96%	0.95	0.94	0.94	0.95	0.95

Table 4.6: Random Forests Classification Results on IDV and SQV Drugs of PIs

Evaluation Measures RF	Acc	Sn	Sp	Recall	FMeasure	ROC
IDV (Test Set)	92%	0.9	0.9	0.92	0.89	0.92
10-fold cross validation	88%	0.89	0.88	0.87	0.88	0.83
SQV (Test set)	92%	0.91	0.92	0.93	0.89	0.85
10-fold cross validation	88%	0.89	0.88	0.87	0.88	0.83

Table 4.7: SVM Classification on PIs Inhibitors

Evaluation Measures SVM	ATV	NFV	RTV	LPV	TPV
Accuracy	98%	98.2%	99.2%	99.2%	98.1%
Standard Deviation (10-fold)	0.21	0.13	0.14	0.25	0.1
Specificity	0.96	0.97	0.95	0.93	0.96
Standard Deviation (10-fold)	0.25	0.18	0.14	0.22	0.21
Sensitivity	0.92	0.95	0.95	0.93	0.96
Standard Deviation (10-fold)	0.12	0.13	0.28	0.23	0.1

Table 4.8: SVM Classification Results on NRTIS and NNRTIs of RTs s

Evaluation Measures SVM	3TC	ABC	AZT	TDF	ETR	EVP	NVP
Accuracy	99.6%	98.2%	99.2%	99.4%	98.5%	99.4%	99.1%
Standard Deviation (10-fold)	0.21	0.13	0.14	0.25	0.31	0.18	0.22
Specificity	0.96	0.97	0.95	0.93	0.97	0.96	0.96
Standard Deviation (10-fold)	0.22	0.21	0.29	0.29	0.15	0.18	0.24
Sensitivity	0.92	0.95	0.95	0.93	0.97	0.96	0.96
Standard Deviation (10-fold)	0.51	0.19	0.18	0.22	0.21	0.28	0.27

Table 4.9: Random Forests Classification Results on PIs, RTs

Evaluation Measures RF	3TC	ABC	AZT	TDF	ETR	EVP	NVP
Accuracy	90.2	87.5	87.5	92.5	92.4	89.5	88.5
Standard Deviation (10-fold)	0.48	0.43	0.54	0.52	0.38	0.48	0.59
Specificity	0.92	0.87	0.89	0.89	0.91	0.91	0.88
Standard Deviation (10-fold)	0.58	0.63	0.44	0.48	0.35	0.39	0.49
Sensitivity	0.89	0.92	0.92	0.92	0.88	0.91	0.91
Standard Deviation (10-fold)	0.52	0.53	0.48	0.55	0.53	0.51	0.48

Table 4.10: Random Forests Classification Results on PIs, RTs

Order	Features	Position
1	mut10_avg_vol	538
2	mut10_E+1	422
63	mut36_CE-1_N	719
143	mut62_Difference	4482
209	mut71_Difference	9863
285	mut74_CE-1_I	6106
292	mut74_avg_vol	6198
317	mut74_CE-2_T	6133
287	mut74_CE+2_K	6188
419	mut41_CE+2_D	1876

Table 4.11: Comparison of Accuracy Measure with the Existing Literature

Methods	PIs Inhibitors						RTs Inhibitors			
	ATV	NFV	IDV	LPV	SQV	TPV	3TC	ABC	AZT	TDF
HIV grade	84.7	81.2	85.1	80.5	80.2	72.8	91.5	89.7	94.6	80.7
ANRS	N/A	78.1	85.1	87.0	N/A	59.7	92.0	83.9	94.4	72.7
HIVdb	N/A	83.4	N/A	83.9	N/A	76.8	94.3	95.0	94.5	79.7
REGA	84.4	82.2	85.6	84.0	69.3	N/A	95.9	86.0	94.0	73.8
SVM	95.5	96.0	94.6	96.2	94.6	96.1	98.7	98.1	98.4	97.5
SVM Combined	98 %	98.2 %	99.2 %	99.2 %	97.65%	98.1 %	99.6 %	98.2 %	99.2 %	99.4 %
ANN	84.7	81.2	85.1	80.5	80.2	72.8	98.2	98.4	98.7	97.0
Sparse Dictionary	N/A	78.1	85.1	87.0	N/A	59.7	91.2	91.5	93.2	85.2

4.3 Drug Repurposing

We developed an algorithm to integrate PPI interactions data, common pathway analysis, binding site structure similarity and disease-disease similarity measure to score the relevance of each component in predicting new diseases for which a drug can be repurposed for. The dataset contains drugs with their old and new disease indications. The half part of the dataset which contains drug, its target and old indication has been used to train our algorithm. And further the method was tested to check if it can predict the hidden part of the dataset (new disease indications).

From human PPI network we filtered proteins that are at least 5 nodes away from the drug target. This gives us approximately 60,000 proteins varying from target to target. We picked these proteins and looked for common pathways among them. This filters the proteins from 60,000 to almost 100 for most of the cases. In third step we checked for binding site structural similarity between the drug target and these filtered protein list. This further shortens the candidate list to at most 4 proteins. The final step was to look for disease-disease similarity associated with the drug target and the new targets. The output of our method shows that the high scoring candidate is always the target of the new disease indication which was already reported as repurposed disease of that particular drug.

From training set, out of forty-seven drugs, the computed scores of the two drug targets were really low. Further scanning of these two cases showed that they do not have any binding site structure similarity between the two targets. Similarly, from the test set out of six, we have found three similar cases. We then picked these less scoring genes and find the pathways they are associated with and further compare the downstream and upstream genes to check binding site structure similarity. The associated pathways were scanned for different levels for downstream and upstream genes which are either activating, phosphorylating or interacting with the target genes. The idea was to uncover if the pathways proteins are actually binding to the drug and blocking its activity. These genes were looked again for structure similarity, all of the drug targets shared similar binding site with the pathway genes, the results are tabulated in Table 4.1.2.

As described, for testing our algorithm, we hid the half part of the dataset and applied it only on the first part which is drug, its target and its old indication. We matched the resulted candidate targets and diseases with the hidden part of the dataset and results showed that our method is successful in predicting the new indications of the drug with

Table 4.12: Results on Pathway based Drug Repurposing

Drug	Target	old Indication	New Indication	Target	Pathway	Target
(Train Set) Mifepristone	Pregnancy termination	FAS	Cushing Syndrome	NR3C1	P53 SIGNALING PATHWAY	P53
(Train Set) Heparin	anticoagulant	SERPINC1	Cystic Fibrosis	TGFB1	COMPLEMENT AND COAGULATION CASCADES	F9
(Test set) Itraconazole	Antifungal	CYP3A4	Cancer	ABCB1	Pathways in Cancer	FIGF

good accuracy. In addition to that it also predicts some novel targets and diseases which in future can be verified experimentally. The novel targets and their associated diseases are tabulated in Table 4.1.3.

We found some interesting findings in our results. For few of the cases, our scoring scheme revealed that the targets from the dataset and the novel target have achieved same scores. For these we extracted the drugs associated with the two and looked for the drug-drug similarity. We calculated Tanimoto coefficient for computing maximum common substructures (MCS) similarity between the two chemicals. This score tells the common substructure shared by two query chemicals. The value of score ranges between 0 to 1; 1 being the highest, our score was between 0.4 to 0.6. Hence we can say that the drugs for these two targets have common substructures so can be repurposed for each other. Results are tabulated in Table 4.1.4

Table 4.13: Novel Predictions for Drug Repurposing

Drug	Target	old Indication	Novel Targets	New Indications	Reported New Indications
Zidovudine	Cancer	TP53	BCL2 MDM2	Breast Neoplasms Hypertension	AIDS
Methotrexate	Cancer	DHFR	MYC RB1	Liver Neoplasms Breast Neoplasms	Rheumatoid arthritis
Memantine	Parkinson Disease	GRIN1	RAF1 RAC1	NOONAN SYNDROME Heart Failure	Alzheimers disease
Thalidomide	Sedative	TNF	IL6 TNFRSF1A	Diabetes Mellitus Liver Syndrome	MULTIPLE MYELOMA
Raloxifene	Osteoposrosis	ESR1	FOS TGFR	Hypertensive Disorders Breast Neoplasms	Breast Neoplasms
Colesevelam	Hyperlipidemia	LDLR	MAPK1 NFKB1 EGFR	Neoplasm Adenocarcinoma Lung Carcinoma	Diabetes Mellitus, Type 2
Imatinib	Chronic Myeloid Leukemia	PDGFRB	MAPK3 RAF1 NRAS	Neoplasm NOONAN Syndrome Juvenile Myelomonocytic	Gastrointestinal stromal tumour

Table 4.14: Total True and False Predictions

Total Drugs	True Predictions	After Pathway Analysis	Correct Predictions
44 (Train Set)	40	42	2
6 (Test Set)	3	5	1

Chapter 5

Discussion

Drug resistance is the major obstacle faced by therapists in treating HIV infected patients. Efficient methods of predicting drug resistance may help to overcome the treatment failure regimens. Besides the proteomic level studies, in silico predictions are also one of the robust solutions to this task. The computational strategies utilize sequence data to dig out HIV mutants to certain drug treatments. Previously, many computational studies have been reported regarding drug resistance prediction.

We built a text miner to associate mutations, expressions and resistance relations with drug resistance. We compared our results with two already published methods on drug resistance, one of them is EDGAR and other method was proposed by [47] which is also used in 5 hospitals from virology lab for prior selection of the resistant data. The HIV relation extraction system [47] was tested on 500 sentences from PubMed and 300 sentences from HIVdb comments with a precision, recall and F- measure of 87%, 82% and 84.5%. Our results showed better performance with 97%,97% and 98% of these evaluation measures hence providing state of the art performance. On the other hand, EDGAR study did not publish any evaluation measure providing an idea of how this approach will work, hence we cannot really make a comparison. There is no gold standard corpus available for drug resistance, which leaves us comparing our method only with the one reported in [47]. Hybrid approaches have a benefit over other methods, such as using only rule based methods yield low recall while machine learning offers low precision if the training set is not of sufficient size; hence combining the two approaches results in better recall and precision measure.

Our study proposed text mining system for drug resistance which has two components. The first component works on categorizing abstracts based on extracting mutations, genes, disease and identifies their associations with drug resistance. The second component works

on classifying all possible relations that can exist between mutations and drugs which are either resistance, intermediate or susceptible relations. The dataset proposed can further be used as a gold- standard corpus for analysis on drug resistance .

Next, we developed a method to predict HIV resistance computationally. Most studies working on HIV resistance have utilized datasets present on the HIV Stanford database. However, there is one study reported in year 2002, in which the authors used the datasets published by Winter and Schinazi research group. In order to check the classifier performance of our selected features in predicting HIV resistance we trained our model on both datasets.

The resistance datasets of few drugs contain imbalanced class information; that is, majority patterns belong to resistance class as opposed to the non-resistance class. It is important to tackle the imbalance dataset issue before applying classifier, to avoid biased output that predicts only the majority class and leaves the minority class out. The probability of bias is higher, if the data is high dimensional and the number of samples is fewer. We checked classification before and after applying filters. The results revealed that both SMOTE and SpreadSubsample improve the overall performance of our classifier. The cross-validation accuracy of the train set was between 95-96% from SVM while with Random Forests it was 82-83%. The accuracy measure on the independent test sets was 98-99.2% by SVM, while with Random Forests it was 87-92%.

The standard ways of predicting HIV resistance are generally the genotype interpretation algorithms like HIV-GRADE, Rega, Stanford HIVdb and ANRS-Rules [37]. For predicting HIV resistance, taking sequence and structure features separately show some limitations. Sequence based features are only limited to the mutations present in the training set as they are less effective in finding completely unseen mutations, while structure based methods can predict resistance for unseen data, but it remains difficult to infer the mechanistic impact of these mutations. Making the right choice of features helps in obtaining a more biologically meaningful representation of protein sequence and structure in order to deduce drug resistance mechanisms.

The results of our strategy showed that our method outperforms the state of the art methods for drug resistance prediction against PIs and RTs for the two classifiers SVM and Random Forests. The sequence and structural features of independent mutations were combined with multiple mutations. The results showed that the accuracy measure for seven drugs ranges between 98-99.2%; while the accuracy reported from standard methods were 59.7-87.0% [18]. One of the studies on drug resistance prediction introduced sparse dictionary

and Delaunay triangulation method as an extension to the standard methods. For PIs and RTs their reported accuracy ranges between 92-97% by SVM [18] while with the sparse dictionary method it ranges between 95-99% for PIs and for RTs it was between 85-91% which is comparable with our results. The accuracy obtained from SVM previously was 95-96%. Our features improved this accuracy to greater than 97%.

Our method was benchmarked against already published methods and showed better results in all of the evaluation measures. The results clearly showed that combining both sequence and structural properties with the added features of contact energies helps in enhancing the accuracy measure. These features are crucial for understanding drug resistance mechanisms. It is important to look for evolutionary conserved residues because mutation in these will disrupt protein structure which in turn affects the binding pocket hence drug will not bind. The residues which are in close proximity in a protein structure when mutated to a large sized amino acid, disrupts the protein structure hence drug binding activity is affected. If a small volume residue mutated into large sized residue it will affect the nearby residues activity hence resulting in the drug not binding with the protein. Similarly, regarding structural features, contact energies between the multiple mutation patterns crucial for ligand molecule binding, hence it is one of the important features. The contact energy changes of a protein structure have a strong impact on unfolding free energy changes which ultimately affects the stability of protein [38]. For drug resistance this feature has always been overlooked. We tested our classifier performance with and without contact energies feature. The results showed the great degree of decrease in accuracy measure without contact energy feature. Hence it is an important feature to be considered for predicting HIV resistance.

One way of combating drug resistance is the concept of drug repositioning. This emerges as an alternate strategy for therapy if the first line and second line treatment fails. We proposed a prioritization algorithm for computational drug repurposing. Drug often binds to more than one target, which is defined as polypharmacology one application of which is drug repurposing also referred as drug repositioning or therapeutic switching. Drug polypharmacology can act both ways; on the one hand it is beneficial for drug repurposing while at the same time it is highly unwanted in drug discovery domain because of probable side effects. Two reasons behind this promiscuous phenomena of drugs are the binding site similarities among the drug targets and the flexibility of drugs that makes them bind to multiple targets. There have been a number of approaches proposed lately for drug repurposing including side effects, gene expression profiles and structure similarity. Drug repositioning is in the limelight of current research as bringing a new drug to the market is becoming more

and more expensive.

Our method integrates four different types of similarity measures between drug and disease. We developed a ground basis for evaluating drug disease relationship by prioritizing the candidates for drug repurposing. All these parameters help in elucidating the unknown associations between drug and diseases for finding the novel targets to reiterate old drugs. We compared our method with previous studies reported on drug repositioning. [33] used pathway information to unravel new uses of old drugs. The authors used all possible associations from drug down to diseases. Their method predicts some novel associations too. We noticed that they did not use binding site structure information in building their network. Unlike them our method uses this parameter which turns out as an important feature for drug repurposing. Another study reported [94] pathway based analysis for drug repositioning. Biological pathways are important for analyzing biological mechanisms that are associated with diseases. Its important to elucidate the pharmacological effect of the drugs which can be inferred by analyzing pathways associated with the drug targets.

[89] reported a method named PREDICT which proposed a strategy of drug repositioning that can be applied to personalized medicine as well. They used five different levels for computing drug-drug and disease-disease similarities. Their measure includes chemical based, sequence based, genetic based, closeness in a PPI, side effect and phenotypic based for disease-disease similarity. They have achieved high rates of accuracy measures $AUC=0.9$. [117] studied the correlation of drug promiscuity with binding site structural similarity which states that one drug can bind to multiple targets because of sharing similar binding site. We have used this concept to build our algorithm. All these published methods did not use these parameters altogether for drug repositioning. Our algorithm helps in predicting novel indications, hence can be applied to a large scale for conducting drug repurposing.

Chapter 6

Conclusions

The aim of this study are manifold. First is to develop a platform that contains all literature about drug resistance combined with mutations and expression changes associations. For this purpose we have built a database "ZK DrugResist" that queries PubMed abstracts and categorized them on the basis of disease association. This tool provides a quick way of getting informed about all the information regarding drug resistance that has been published till now. Thus saving time and energy from searching online one after another through Medline repositories. Up to the best of our knowledge no such Text Miner has been built before on drug resistance, hence our tool can be used on large scale for the analysis of drug resistance against complex diseases.

The second goal was to build such methods that can predict resistance with more precision. We have revised a strategy for determining HIV resistance by merging both sequence and structural properties. We have introduced a novel combination of features that give promising results in terms of performance measures. In the previous studies sequence and structure features were either applied separately or if combined the features like volume measure, contact energies and multiple interactions were usually missed out. Up to the best of our knowledge no study has ever used this combination of features for predicting HIV resistance, hence our feature set could be beneficial for predicting HIV resistance with greater accuracy. All of these features are crucial in inferring valuable information for understanding mechanistic insights of drug resistance. Even with all the odds that may emerge, we are optimistic that the methods and results shown here will break down the ways that will improve drug activity to overcome resistance. For future studies, these features can be used to predict resistance to the drugs of other complex diseases.

The third goal was to propose drug repurposing strategy that serves as a ray of hope in bat-

ting drug resistance. Our method is based on a similarity scheme that can handle both approved and novel targets for drug-disease association. Our model integrates protein-protein interaction data, biological pathways, binding site similarity and disease-disease similarity unlike one drug one target models. The algorithm tests the relevance of each parameter and scores accordingly. Results showed that our method is successful in predicting already reported new indications of a drug and along with that some novel indications were also found. The novel targets can serve as leads that requires further experimental validation. Repurposed drugs provide a best alternative for treating drug resistance.

As of future work we will further improve our drug repurposing methodology by adding transcriptional responses which states that, if two drugs have similar transcriptional responses they will have similar MOA hence can be repurposed. Also, we will perform parameter optimization of our scores for obtaining better results.

Bibliography

- [1] F. Sams-Dodd, “Target-based drug discovery: is something wrong?,” *Drug discovery today*, vol. 10, no. 2, pp. 139–147, 2005.
- [2] L. H. Miller, H. C. Ackerman, X.-z. Su, and T. E. Wellems, “Malaria biology and disease pathogenesis: insights for new treatments,” *Nature medicine*, vol. 19, no. 2, pp. 156–167, 2013.
- [3] N. Daniel, V. Schneider, G. Pialoux, A. Krivine, S. Grabar, T. H. Nguyen, P.-M. Girard, W. Rozenbaum, and D. Salmon, “Emergence of hiv-1 mutated strains after interruption of highly active antiretroviral therapy in chronically infected patients,” *Aids*, vol. 17, no. 14, pp. 2126–2129, 2003.
- [4] A. Maiti, “Reactive oxygen species reduction is a key underlying mechanism of drug resistance in cancer chemotherapy,” *Chemotherapy*, vol. 1, no. 2, pp. 1–5, 2012.
- [5] S. Remy, S. Gabriel, B. W. Urban, D. Dietrich, T. N. Lehmann, C. E. Elger, U. Heinemann, and H. Beck, “A novel mechanism underlying drug resistance in chronic epilepsy,” *Annals of neurology*, vol. 53, no. 4, pp. 469–479, 2003.
- [6] A. Hochhaus, P. L. Rosée, M. C. Müller, T. Ernst, and N. C. Cross, “Impact of bcr-abl mutations on patients with chronic myeloid leukemia,” *Cell cycle*, vol. 10, no. 2, pp. 250–260, 2011.
- [7] R. Chrisanthar, S. Knappskog, E. Løkkevik, G. Anker, B. Østenstad, S. Lundgren, E. O. Berge, T. Risberg, I. Mjaaland, L. Mæhle, *et al.*, “Chek2 mutations affecting kinase activity together with mutations in tp53 indicate a functional pathway associated with resistance to epirubicin in primary breast cancer,” *PLoS One*, vol. 3, no. 8, p. e3062, 2008.

- [8] H. Ode, S. Neya, M. Hata, W. Sugiura, and T. Hoshino, “Computational simulations of hiv-1 proteases multi-drug resistance due to nonactive site mutation 190m,” *Journal of the American Chemical Society*, vol. 128, no. 24, pp. 7887–7895, 2006.
- [9] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, “Cancer drug resistance: an evolving paradigm,” *Nature Reviews Cancer*, vol. 13, no. 10, pp. 714–726, 2013.
- [10] T. Nambu, N. Araki, A. Nakagawa, A. Kuniyasu, T. Kawaguchi, A. Hamada, and H. Saito, “Contribution of bcr–abl-independent activation of erk1/2 to acquired imatinib resistance in k562 chronic myeloid leukemia cells,” *Cancer science*, vol. 101, no. 1, pp. 137–142, 2010.
- [11] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner, “Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches,” *PloS one*, vol. 6, no. 3, p. e18029, 2011.
- [12] L. H. Reeve, H. Han, and A. D. Brooks, “The use of domain-specific concepts in biomedical text summarization,” *Information Processing & Management*, vol. 43, no. 6, pp. 1765–1776, 2007.
- [13] S. ElShal, M. Mathad, J. Simm, J. Davis, and Y. Moreau, “Topic modeling of biomedical text,”
- [14] L. J. Jensen, J. Saric, and P. Bork, “Literature mining for the biologist: from information retrieval to biological discovery,” *Nature reviews genetics*, vol. 7, no. 2, pp. 119–129, 2006.
- [15] S. Ananiadou, D. B. Kell, and J.-i. Tsujii, “Text mining and its potential applications in systems biology,” *Trends in biotechnology*, vol. 24, no. 12, pp. 571–579, 2006.
- [16] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, *et al.*, “Literature-curated protein interaction datasets,” *Nature methods*, vol. 6, no. 1, pp. 39–46, 2009.
- [17] G. D. Bader, D. Betel, and C. W. Hogue, “Bind: the biomolecular interaction network database,” *Nucleic acids research*, vol. 31, no. 1, pp. 248–250, 2003.

- [18] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, *et al.*, “The intact molecular interaction database in 2012,” *Nucleic acids research*, p. gkr1088, 2011.
- [19] A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. ODonnell, *et al.*, “The biogrid interaction database: 2013 update,” *Nucleic acids research*, vol. 41, no. D1, pp. D816–D823, 2013.
- [20] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguéz, T. Dörcks, M. Stark, J. Müller, P. Bork, *et al.*, “The string database in 2011: functional interaction networks of proteins, globally integrated and scored,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D561–D568, 2011.
- [21] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, “Mint: a molecular interaction database,” *FEBS letters*, vol. 513, no. 1, pp. 135–140, 2002.
- [22] C. Quan, M. Wang, and F. Ren, “An unsupervised text mining method for relation extraction from biomedical literature,” *PloS one*, vol. 9, no. 7, p. e102039, 2014.
- [23] Y. Mehellou and E. De Clercq, “Twenty-six years of anti-hiv drug discovery: where do we stand and where do we go?,” *J. Med. Chem*, vol. 53, no. 2, pp. 521–538, 2010.
- [24] L. Menéndez-Arias, “Molecular basis of human immunodeficiency virus type 1 drug resistance: overview and recent developments,” *Antiviral research*, vol. 98, no. 1, pp. 93–120, 2013.
- [25] I. T. Weber and J. Agniswamy, “Hiv-1 protease: structural perspectives on drug resistance,” *Viruses*, vol. 1, no. 3, pp. 1110–1136, 2009.
- [26] L. Menéndez-Arias, “Molecular basis of human immunodeficiency virus drug resistance: an update,” *Antiviral research*, vol. 85, no. 1, pp. 210–231, 2010.
- [27] D. Wang and B. Larder, “Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks,” *Journal of Infectious Diseases*, vol. 188, no. 5, pp. 653–660, 2003.
- [28] X. Chen, I. T. Weber, and R. W. Harrison, “Molecular dynamics simulations of 14 hiv protease mutants in complexes with indinavir,” *Journal of molecular modeling*, vol. 10, no. 5-6, pp. 373–381, 2004.

- [29] S. Drăghici and R. B. Potter, “Predicting hiv drug resistance with neural networks,” *Bioinformatics*, vol. 19, no. 1, pp. 98–107, 2003.
- [30] T. T. Ashburn and K. B. Thor, “Drug repositioning: identifying and developing new uses for existing drugs,” *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 673–683, 2004.
- [31] K. Shabana, K. A. Nazeer, M. Pradhan, and M. J. Palakal, “A computational method for drug repositioning using publicly available gene expression data,” in *Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference on*, pp. 1–2, IEEE, 2014.
- [32] H.-M. Lee and Y. Kim, “Drug repurposing is a new opportunity for developing drugs against neuropsychiatric disorders,” *Schizophrenia research and treatment*, vol. 2016, 2016.
- [33] J. Li and Z. Lu, “Pathway-based drug repositioning using causal inference,” *BMC bioinformatics*, vol. 14, no. 16, p. S3, 2013.
- [34] Z. Khalid and O. U. Sezerman, “Prediction of hiv drug resistance by combining sequence and structural properties,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.
- [35] Z. Khalid and O. U. Sezerman, “Zk drugresist 2.0: A textminer to extract semantic relations of drug resistance from pubmed,” *Journal of Biomedical Informatics*, vol. 69, pp. 93–98, 2017.
- [36] W. W. Fleuren and W. Alkema, “Application of text mining in the biomedical domain,” *Methods*, vol. 74, pp. 97–106, 2015.
- [37] P. V. Marsden and N. E. Friedkin, “Network studies of social influence,” *Sociological Methods & Research*, vol. 22, no. 1, pp. 127–151, 1993.
- [38] C. Pal, J. Bengtsson-Palme, C. Rensing, E. Kristiansson, and D. J. Larsson, “Bacmet: antibacterial biocide and metal resistance genes database,” *Nucleic acids research*, vol. 42, no. D1, pp. D737–D743, 2014.
- [39] Y.-Y. Wang, W.-H. Chen, P.-P. Xiao, W.-B. Xie, Q. Luo, P. Bork, and X.-M. Zhao, “Gear: A database of genomic elements associated with drug resistance,” *Scientific Reports*, vol. 7, p. 44085, 2017.

- [40] R. A. Erhardt, R. Schneider, and C. Blaschke, “Status of text-mining techniques applied to biomedical text,” *Drug discovery today*, vol. 11, no. 7, pp. 315–325, 2006.
- [41] J. Šarić, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork, “Extraction of regulatory gene/protein networks from medline,” 2006.
- [42] J. Vercauteren and A.-M. Vandamme, “Algorithms for the interpretation of hiv-1 genotypic drug resistance information,” *Antiviral research*, vol. 71, no. 2, pp. 335–342, 2006.
- [43] R. Chowdhary, J. Zhang, and J. S. Liu, “Bayesian inference of protein–protein interactions from biological literature,” *Bioinformatics*, vol. 25, no. 12, pp. 1536–1542, 2009.
- [44] S. Kim, J. Yoon, and J. Yang, “Kernel approaches for genic interaction extraction,” *Bioinformatics*, vol. 24, no. 1, pp. 118–126, 2008.
- [45] T. Lengauer and T. Sing, “Bioinformatics-assisted anti-hiv therapy,” *Nature Reviews Microbiology*, vol. 4, no. 10, pp. 790–797, 2006.
- [46] H. Saigo, T. Uno, and K. Tsuda, “Mining complex genotypic features for predicting hiv-1 drug resistance,” *Bioinformatics*, vol. 23, no. 18, pp. 2455–2462, 2007.
- [47] Q.-C. Bui, B. Ó. Nualláin, C. A. Boucher, and P. M. Sloot, “Extracting causal relations on hiv drug resistance from literature,” *BMC bioinformatics*, vol. 11, no. 1, p. 101, 2010.
- [48] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter, “Edgar: extraction of drugs, genes and relations from the biomedical literature,” in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 517, NIH Public Access, 2000.
- [49] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, “Human immunodeficiency virus reverse transcriptase and protease sequence database,” *Nucleic acids research*, vol. 31, no. 1, pp. 298–303, 2003.
- [50] S.-Y. Rhee, J. Taylor, G. Wadhwa, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer, “Genotypic predictors of human immunodeficiency virus type 1 drug resistance,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17355–17360, 2006.

- [51] K. Van der Borgh, E. Van Craenenbroeck, P. Lecocq, M. Van Houtte, B. Van Kerckhove, L. Bacheler, G. Verbeke, and H. van Vlijmen, “Cross-validated stepwise regression for identification of novel non-nucleoside reverse transcriptase inhibitor resistance associated mutations,” *BMC bioinformatics*, vol. 12, no. 1, p. 386, 2011.
- [52] E. Pasomsub, C. Sukasem, S. Sungkanuparph, B. Kijirikul, W. Chantratita, *et al.*, “The application of artificial neural networks for phenotypic drug resistance prediction: evaluation and comparison with other interpretation systems,” *Jpn. J. Infect. Dis*, vol. 63, no. 2, pp. 87–94, 2010.
- [53] D. Heider, J. Verheyen, and D. Hoffmann, “Predicting bevirimat resistance of hiv-1 from genotype,” *BMC bioinformatics*, vol. 11, no. 1, p. 37, 2010.
- [54] J. Zhang, T. Hou, W. Wang, and J. S. Liu, “Detecting and understanding combinatorial mutation patterns responsible for hiv drug resistance,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 4, pp. 1321–1326, 2010.
- [55] M. Junaid, M. Lapins, M. Eklund, O. Spjuth, and J. E. Wikberg, “Proteochemometric modeling of the susceptibility of mutated variants of the hiv-1 virus to reverse transcriptase inhibitors,” *PloS one*, vol. 5, no. 12, p. e14353, 2010.
- [56] M. Lapins, M. Eklund, O. Spjuth, P. Prusis, and J. E. Wikberg, “Proteochemometric modeling of hiv protease susceptibility,” *BMC bioinformatics*, vol. 9, no. 1, p. 181, 2008.
- [57] M. Obermeier, A. Pironti, T. Berg, P. Braun, M. Däumer, J. Eberle, R. Ehret, R. Kaiser, N. Kleinkauf, K. Korn, *et al.*, “Hiv-grade: a publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge,” *Intervirology*, vol. 55, no. 2, pp. 102–107, 2012.
- [58] J. Kjaer, L. Høj, Z. Fox, and J. Lundgren, “Prediction of phenotypic susceptibility to antiretroviral drugs using physiochemical properties of the primary enzymatic structure combined with artificial neural networks,” *HIV medicine*, vol. 9, no. 8, pp. 642–652, 2008.
- [59] C. Reid, R. Bassett, S. Day, B. Larder, V. DeGruttola, and D. Winslow, “A dynamic rules-based interpretation system derived by an expert panel is predictive of virological failure,” in *Antiviral Therapy*, vol. 7, pp. S121–S121, INT MEDICAL PRESS LTD 2-4 IDOL LANE, LONDON EC3R 5DD, ENGLAND, 2002.

- [60] C.-Y. Chen, I. Georgiev, A. C. Anderson, and B. R. Donald, "Computational structure-based redesign of enzyme activity," *Proceedings of the National Academy of Sciences*, vol. 106, no. 10, pp. 3764–3769, 2009.
- [61] H. A. Wahab, Y.-S. Choong, P. Ibrahim, A. Sadikun, and T. Scior, "Elucidating isoniazid resistance using molecular modeling," *Journal of chemical information and modeling*, vol. 49, no. 1, pp. 97–107, 2008.
- [62] X.-L. Zhu, H. Ge-Fei, C.-G. Zhan, and G.-F. Yang, "Computational simulations of the interactions between acetyl-coenzyme-a carboxylase and clodinafop: resistance mechanism due to active and nonactive site mutations," *Journal of chemical information and modeling*, vol. 49, no. 8, pp. 1936–1943, 2009.
- [63] S. Mittal, R. M. Bandaranayake, N. M. King, M. Prabu-Jeyabalan, M. N. Nalam, E. A. Nalivaika, N. K. Yilmaz, and C. A. Schiffer, "Structural and thermodynamic basis of amprenavir/darunavir and atazanavir resistance in hiv-1 protease with mutations at residue 50," *Journal of virology*, vol. 87, no. 8, pp. 4176–4184, 2013.
- [64] D. Wang, B. Larder, A. Revell, J. Montaner, R. Harrigan, F. De Wolf, J. Lange, S. Wegner, L. Ruiz, M. J. Pérez-Elías, *et al.*, "A comparison of three computational modelling methods for the prediction of virological response to combination hiv therapy," *Artificial intelligence in medicine*, vol. 47, no. 1, pp. 63–74, 2009.
- [65] N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig, "Diversity and complexity of hiv-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 8271–8276, 2002.
- [66] N. Beerenwinkel, M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter, "Geno2pheno: Estimating phenotypic drug resistance from hiv-1 genotypes," *Nucleic acids research*, vol. 31, no. 13, pp. 3850–3855, 2003.
- [67] A. D. Sevin, V. DeGruttola, M. Nijhuis, J. M. Schapiro, A. S. Foulkes, M. F. Para, and C. A. Boucher, "Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type 1 genotype with applications to aids clinical trials group 333," *Journal of Infectious Diseases*, vol. 182, no. 1, pp. 59–67, 2000.

- [68] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
- [69] X. Yu, I. T. Weber, and R. W. Harrison, “Prediction of hiv drug resistance from genotype with encoded three-dimensional protein structure,” *BMC genomics*, vol. 15, no. 5, p. S1, 2014.
- [70] J. E. Foulkes-Murzycki, W. R. P. Scott, and C. A. Schiffer, “Hydrophobic sliding: a possible mechanism for drug resistance in human immunodeficiency virus type 1 protease,” *Structure*, vol. 15, no. 2, pp. 225–233, 2007.
- [71] W. Wang and P. A. Kollman, “Computational study of protein specificity: the molecular basis of hiv-1 protease drug resistance,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 26, pp. 14937–14942, 2001.
- [72] A. Gupta, S. Jamal, S. Goyal, R. Jain, D. Wahi, and A. Grover, “Structural studies on molecular mechanisms of nelfinavir resistance caused by non-active site mutation v77i in hiv-1 protease,” *BMC bioinformatics*, vol. 16, no. 19, p. S10, 2015.
- [73] I. Cruz, M. G. Lorenzo, R. Abalo, and R. Rodriguez, “Prediction of human immunodeficiency virus drug resistance using contact energies,” in *Neural Networks and Brain, 2005. ICNN&B’05. International Conference on*, vol. 1, pp. 490–493, IEEE, 2005.
- [74] J. Zhang, T. Hou, Y. Liu, G. Chen, X. Yang, J. S. Liu, and W. Wang, “Systematic investigation on interactions for hiv drug resistance and cross-resistance among protease inhibitors,” *journal of Proteome Science and Computational Biology*, vol. 1, no. 1, p. 2, 2012.
- [75] M. Masso and I. I. Vaisman, “Sequence and structure based models of hiv-1 protease and reverse transcriptase drug resistance,” *BMC genomics*, vol. 14, no. 4, p. S3, 2013.
- [76] V. L. Ravich, M. Masso, and I. I. Vaisman, “A combined sequence–structure approach for predicting resistance to the non-nucleoside hiv-1 reverse transcriptase inhibitor nevirapine,” *Biophysical chemistry*, vol. 153, no. 2, pp. 168–172, 2011.
- [77] M. Masso and I. I. Vaisman, “A novel sequence-structure approach for accurate prediction of resistance to hiv-1 protease inhibitors,” in *Bioinformatics and Bioengineering*,

2007. *BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pp. 952–958, IEEE, 2007.
- [78] M. Masso, “Sequence-based predictive models of resistance to hiv-1 integrase inhibitors: An n-grams approach to phenotype assessment,” *Current HIV research*, vol. 13, no. 6, pp. 497–502, 2015.
- [79] J. T. Dudley, E. Schadt, M. Sirota, A. J. Butte, and E. Ashley, “Drug discovery in a multidimensional world: systems, patterns, and networks,” *Journal of cardiovascular translational research*, vol. 3, no. 5, pp. 438–447, 2010.
- [80] E. E. Schadt, S. H. Friend, and D. A. Shaywitz, “A network view of disease and compound screening,” *Nature reviews Drug discovery*, vol. 8, no. 4, pp. 286–295, 2009.
- [81] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, *et al.*, “Predicting new molecular targets for known drugs,” *Nature*, vol. 462, no. 7270, pp. 175–181, 2009.
- [82] J. Li and Z. Lu, “A new method for computational drug repositioning using drug pairwise similarity,” in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference On*, pp. 1–4, IEEE, 2012.
- [83] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, *et al.*, “Discovery of drug mode of action and drug repositioning from transcriptional responses,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 33, pp. 14621–14626, 2010.
- [84] G. Hu and P. Agarwal, “Human disease-drug network based on genomic expression profiles,” *PLoS one*, vol. 4, no. 8, p. e6536, 2009.
- [85] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte, “Discovery and preclinical validation of drug indications using compendia of public gene expression data,” *Science translational medicine*, vol. 3, no. 96, pp. 96ra77–96ra77, 2011.
- [86] D. Shigemizu, Z. Hu, J.-H. Hung, C.-L. Huang, Y. Wang, and C. DeLisi, “Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer,” *PLoS Comput Biol*, vol. 8, no. 2, p. e1002347, 2012.

- [87] J. Li, X. Zhu, and J. Y. Chen, “Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts,” *PLoS Comput Biol*, vol. 5, no. 7, p. e1000450, 2009.
- [88] Y. Li and P. Agarwal, “A pathway-based view of human diseases and disease relationships,” *PloS one*, vol. 4, no. 2, p. e4346, 2009.
- [89] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, “Predict: a method for inferring novel drug indications with application to personalized medicine,” *Molecular systems biology*, vol. 7, no. 1, p. 496, 2011.
- [90] F. Iorio, T. Rittman, H. Ge, M. Menden, and J. Saez-Rodriguez, “Transcriptional data: a new gateway to drug repositioning?,” *Drug discovery today*, vol. 18, no. 7, pp. 350–357, 2013.
- [91] D. Emig, A. Ivliev, O. Pustovalova, L. Lancashire, S. Bureeva, Y. Nikolsky, and M. Bessarabova, “Drug target prediction and repositioning using an integrated network-based approach,” *PLoS One*, vol. 8, no. 4, p. e60618, 2013.
- [92] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, “Drug target identification using side-effect similarity,” *Science*, vol. 321, no. 5886, pp. 263–266, 2008.
- [93] F. Vitali, L. D. Cohen, A. Demartini, A. Amato, V. Eterno, A. Zambelli, and R. Bellazzi, “A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer,” *PloS one*, vol. 11, no. 9, p. e0162407, 2016.
- [94] Y. Pan, T. Cheng, Y. Wang, and S. H. Bryant, “Pathway analysis for drug repositioning based on public database mining,” *Journal of chemical information and modeling*, vol. 54, no. 2, pp. 407–418, 2014.
- [95] F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. DAmato, and D. Greco, “Drug repositioning: a machine-learning approach through data integration,” *Journal of cheminformatics*, vol. 5, no. 1, p. 30, 2013.
- [96] W. J. Strittmatter, “Old drug, new hope for alzheimer’s disease,” *Science*, vol. 335, no. 6075, pp. 1447–1448, 2012.

- [97] A. Sivachenko, A. Kalinin, and A. Yuryev, “Pathway analysis for design of promiscuous drugs and selective drug mixtures,” *Current drug discovery technologies*, vol. 3, no. 4, pp. 269–277, 2006.
- [98] P. E. Cramer, J. R. Cirrito, D. W. Wesson, C. D. Lee, J. C. Karlo, A. E. Zinn, B. T. Casali, J. L. Restivo, W. D. Goebel, M. J. James, *et al.*, “ApoE-directed therapeutics rapidly clear β -amyloid and reverse deficits in ad mouse models,” *science*, vol. 335, no. 6075, pp. 1503–1506, 2012.
- [99] A. W. Muzaffar, F. Azam, and U. Qamar, “A relation extraction framework for biomedical text using hybrid feature set,” *Computational and mathematical methods in medicine*, vol. 2015, 2015.
- [100] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [101] A. S. Yavuz, B. Ozer, and O. U. Sezerman, “Pattern recognition for subfamily level classification of gpcrs using motif distillation and distinguishing power evaluation,” in *IAPR International Conference on Pattern Recognition in Bioinformatics*, pp. 267–276, Springer, 2012.
- [102] I. B. Kuznetsov and M. McDuffie, “Flexpred: a web-server for predicting residue positions involved in conformational switches in proteins,” *Bioinformatics*, vol. 3, no. 3, pp. 134–136, 2008.
- [103] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon, “Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content,” *Bioinformatics*, vol. 21, no. 16, pp. 3433–3434, 2005.
- [104] T. P. Hopp and K. R. Woods, “Prediction of protein antigenic determinants from amino acid sequences,” *Proceedings of the National Academy of Sciences*, vol. 78, no. 6, pp. 3824–3828, 1981.
- [105] D. Kharakoz, “Partial volumes and compressibilities of extended polypeptide chains in aqueous solution: additivity scheme and implication of protein unfolding at normal and high pressure,” *Biochemistry*, vol. 36, no. 33, pp. 10276–10285, 1997.
- [106] L. J. McGuffin, K. Bryson, and D. T. Jones, “The psipred protein structure prediction server,” *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

- [107] M. Källberg, G. Margaryan, S. Wang, J. Ma, and J. Xu, “Raptorx server: a resource for template-based protein structure modeling,” *Protein Structure Prediction*, pp. 17–27, 2014.
- [108] M. Berrera, H. Molinari, and F. Fogolari, “Amino acid empirical contact energy definitions for fold recognition in the space of contact maps,” *BMC bioinformatics*, vol. 4, no. 1, p. 8, 2003.
- [109] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [110] A. S. Yavuz, N. B. Sözer, and O. U. Sezerman, “Prediction of neddylation sites from protein sequences and sequence-derived properties,” *BMC bioinformatics*, vol. 16, no. 18, p. S9, 2015.
- [111] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [112] M. A. Winters, J. M. Schapiro, J. Lawrence, and T. C. Merigan, “Human immunodeficiency virus type 1 protease genotypes and in vitro protease inhibitor susceptibilities of isolates from individuals who were switched to other protease inhibitors after long-term saquinavir treatment,” *Journal of virology*, vol. 72, no. 6, pp. 5303–5306, 1998.
- [113] J. Hammond, C. Calef, B. Larder, R. Schinazi, J. W. Mellors, *et al.*, “Mutations in retroviral genes associated with drug resistance,” *Human retroviruses and AIDS*, pp. 11136–11179, 1998.
- [114] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [115] M. Kissa and G. Tsatsaronis, “A benchmark dataset for computational drug repositioning,” *Biomed Data J*, vol. 1, no. 2, pp. 10–12, 2015.
- [116] J. S. Shim and J. O. Liu, “Recent advances in drug repositioning for the discovery of new anticancer drugs,” *Int J Biol Sci*, vol. 10, no. 7, pp. 654–63, 2014.

- [117] V. J. Haupt, S. Daminelli, and M. Schroeder, “Drug promiscuity in pdb: protein binding site similarity is key,” *PLoS one*, vol. 8, no. 6, p. e65894, 2013.

Appendices

Appendix A

List of FDA Approved HIV and Cancer Drugs

Table A.1: Protease Inhibitors PIs

Name	Abbreviation
Amprenavir	APV
Atazanavir	ATV
Darunavir	DRV
Indinavir	IDV
Nelfinavir	NFV
Ritonavir	RTV
Saquinavir	SQV
Tipranavir	TPV

Table A.2: Nucleoside Reverse Transcriptase NRTIs and Non Nucleoside Reverse Transcriptase NNRTIs

Name	Abbreviation
Lamivudine	3TC
Abacavir	ABC
Zidovudin	AZT
Tenofovir	TDF
Etravirine	ETR
Encenicline	EVP
Nevirapine	NVP

Table A.3: List of FDA Approved Cancer Drugs

Drugs for Cancer	Drugs for cancer
Cabometyx	Proleukin
lenvatinib	Valstar
nivolumab	Xeloda
atezolizumab	Zofran
venetoclax	Anzemet
alectinib	Bromfenac
cobimetinib	letrozole
daratumumab	Neumega
elotuzumab	Taxol
panobinostat	flutamide
palbociclib	iodixanol
pembrolizumab	amifostine
trifluridine and tipiracil	sargramostim
ixazomib	ibrutinib
sonidegib	afatinib
necitumumab	obinutuzumab
osimertinib	ceritinib
dinutuximab	idelalisib
rolapitant	nivolumab
belinostat	olaparib
blinatumomab	pembrolizumab
ramucirumab	ibrutinib

Appendix B

Drug Repurposing Datasets

Table B.1: Training set for Drug Repurposing

DrugbankId	Drug Name	Old Indication	New Indication	Year of Repositioning
DB01611	Hydroxychloroquine Sulphate	Antiparasitic and antimalarian agent	Lupus Erythematosus, Systemic	1955
DB00437	Allopurinol	Tumor Lysis Syndrome	Gout	1967
DB00915	Amantadine	Influenza, Human	Parkinson Disease	1969
DB00495	Zidovudine	Cancer	AIDS	1987
DB00563	Methotrexate	Cancer	Rheumatoid arthritis	1988
DB00350	Minoxidil	Hypertension	Alopecia	1988
DB00704	Naltrexone	Opioid Dependence	Alcohol Withdrawal	1994
DB00755	Retinoic acid (Tretinoin)	Acne Vulgaris, Keratosis Pilaris	Leukemia, Promyelocytic, Acute	1995
DB00182	Amphetamine (Aderall)	Stimulant, Obesity	Attention Deficit Disorder with Hyperactivity	1996
DB00441	Gemcitabine	Antiviral	Pancreatic Neoplasms, Bronchogenic Carcinoma	1996
DB00681	Amphotericin B	Antifungal	Leishmaniasis	1997
DB01156	Bupropion	Depression	Smoking Cessation	1997
DB01216	Finasteride	Benign Prostatic Hyperplasia	Alopecia	1997
DB01105	Sibutramine	Depression	Obesity	1997
DB01005	Hydroxycarbamide (Hydroxyurea)	Myeloproliferative Disorders	Anemia, Sickle Cell	1998
DB00203	Sildenafil	Angina Pectoris	Erectile Dysfunction	1998
DB01041	Thalidomide	Sedative	Erythema Nodosum Leprosum	1998
DB00482	Celecoxib	Analgesia, Osteoarthritis and adult rheumatoid arthritis	Familial Adenomatous Polyposis	1999
DB00065	Infliximab	Crohn Disease	Rheumatoid arthritis	1999
DB01222	Budesonide	Asthma, Rhinitis, Nasal Polyps	Crohn Disease	2001
DB00254	Doxycycline	Bacterial Infections	Periodontitis	2001
DB00472	Fluoxetine	Depression	Premenstrual Syndrome	2001
DB00674	Galantamine	Myopathic Conditions	Alzheimer Disease	2001
DB01169	Arsenic	Syphilis, African Trypanosomiasis	Leukemia, Promyelocytic, Acute	2002
DB00289	Atomoxetine	Depression	Attention Deficit Disorder with Hyperactivity	2002
DB00996	Gabapentin	Epilepsy	Neuralgia	2002
DB01043	Memantine	Parkinson Disease	Alzheimers disease	2003
DB00441	Gemcitabine	Antiviral	Breast Neoplasms	2004
DB01229	Paclitaxel	Cancer	Restenosis	2004
DB00273	Topiramate	Epilepsy	Migraine	2004
DB00268	Ropinirole	Parkinson Disease	Restless Legs Syndrome	2005
DB00203	Sildenafil	Angina Pectoris	Hypertension, Pulmonary	2005
DB00393	Nimodipine	Hypertension	Vasospasm, Intracranial	2006
DB00441	Gemcitabine	Antiviral	Ovarian Neoplasms	2006
DB00073	Rituximab	Non-Hodgkin Lymphoma, Chronic Lymphocytic Leukemia	Rheumatoid arthritis	2006
DB01041	Thalidomide	Sedative	Multiple Myeloma	2006
DB00476	Duloxetine	Diabetic Neuropathies	Depression	2007
DB00230	Pregabalin	Epilepsy, Diabetic Neuropathies	Anxiety Disorders	2007
DB00230	Pregabalin	Epilepsy, Diabetic Neuropathies	Fibromyalgia	2007
DB00481	Raloxifene	Osteoporosis	Breast Neoplasms	2007
DB00905	Bimatoprost	Glaucoma	Hypotrichosis	2008
DB00930	Colesevelam	Hyperlipidemia	Diabetes Mellitus, Type 2	2008
DB00476	Duloxetine	Diabetic Neuropathies	Fibromyalgia	2008
DB00692	Phentolamine	Hypertension	Anesthesia	2008
DB04896	Milnacipran	Depression	Fibromyalgia	2009
DB00820	Tadalafil	Erectile Dysfunction	Hypertension, Pulmonary	2009
DB01142	Doxepin	Depression	Insomnia	2010
DB00476	Duloxetine	Diabetic Neuropathies	Musculoskeletal Pain (Osteoarthritis, Low Back Pain)	2010
DB00820	Tadalafil	Erectile Dysfunction	Prostatic Hyperplasia, Benign	2011
DB00834	Mifepristone	Pregnancy termination	Cushing Syndrome	2012
DB00273	Topiramate	Epilepsy	Obesity	2012
DB01222	Budesonide	Asthma, Rhinitis, Nasal Polyps	Colitis, Ulcerative	2013
DB00480	Lenalidomide	Multiple Myeloma	Lymphoma, Mantle-Cell	2013

Table B.2: Test set for Drug Repurposing

Drugs	Original Use	Anticancer Mechanisms	Developmental Status
Itraconazole	Treatment of fungal infections	Inhibiting 20S proteasome and AKT signaling Inhibiting Hedgehog pathway	phase I and II
Nelfinavir	Treatment for HIV infections	Inhibiting endothelial cell cholesterol trafficking and angiogenesis Inhibiting HSP90 and HER2 signaling Inducing ER stress and autophagy, and inhibiting angiogenesis	phase I and I
Digoxin	Treatment for cardiac diseases	Inhibiting Na ⁺ /K ⁺ -ATPase Acting as a phytoestrogen and inhibiting androgen receptor signaling Inhibiting HIF-1 synthesis	phase I and I
Nitroxoline	Treatment for urinary tract infections	Inhibiting human MetAP2 and sirtuins in endothelial cells Inducing premature senescence and inhibiting angiogenesis Inhibiting cathepsin B	Preclinical trials
Riluzole	Treatment for Amyotrophic lateral sclerosis	Inhibiting the release of glutamate Inhibiting cell proliferation of metabotropic glutamate receptor 1 (GRM1)-expressing human melanoma cells	Phase I and II
Disulfiram	Treatment for chronic alcoholism	Inhibiting proteasome when complexed with metals Inhibiting DNA methyltransferase 1 (DNMT1)	Phase II and III