2014

Similarity based learning method for drug target interaction prediction

Allapalli Bharadwaja
University of Windsor

Similarity based Learning Method for Drug Target Interaction Prediction

By

Bharadwaja Allapalli

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of **Computer Science**
in Partial Fulfillment of the Requirements for
the Degree of **Master of Science**
at the University of Windsor

Windsor, Ontario, Canada

2014

Similarity based Learning Method for Drug Target Interaction Prediction

by

Bharadwaja Allapalli

APPROVED BY:

_____

Dr. K. Tepe
Department of Electrical and Computer Engineering

_____

Dr. L. Rueda
School of Computer Science

_____

Dr. A. Ngom, Advisor
School of Computer Science

October 14, 2014

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that to the best of my knowledge this work does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

In silico prediction of drug target interactions has gained its popularity with the growth of publicly available information in chemical and biological sciences. The old paradigm of 'one drug-one target' is quickly becoming outdated. It was smart way of understanding the drug-protein interactions but the biological systems we are dealing with are made up of myriad of proteins exhibiting multiple functions. To analyze and understand these systems as a whole, we require efficient computational models. In this work we have improved a machine learning method by integrating more correlated information about the drug compounds and extend this method to weighted profile method in order to infer novel interactions for drugs and targets with no prior interaction information, which was not possible with the current model. We have evaluated our method using area under the ROC curve and the results obtained show that the proposed model can predict drug target interactions accurately.

DEDICATION

I dedicate this to the almighty.

ACKNOWLEDGEMENTS

I am honoured and extremely humbled to work under the supervision of Dr. Alioune Ngom. I thank him for his time and support throughout. His constant motivation, ideas and most importantly patience are the foundation pillars of my research work.

My sincere gratitude goes to internal reader Dr. Luis Rueda for taking time of his busy schedule and also for his valuable suggestions to improve my work. I also extend my regards to external reader Dr. Kemal Tepe for his time and advice.

I thank all my family members and finally, I believe that no work is complete without the blessings of the almighty, thank you god.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

1.1 Preface

The study of molecular biology through analysis of bio-molecular networks has been fundamental in leading scientists and researchers venture deep in understanding the natural and chemical sciences at the molecular level. Many still unknown discoveries are yet to be unearthed which can pose as efficient solutions to the problems faced by our community, for example cure for diseases such as cancer which still haunts the most intelligent species on this planet. The field of study which deals with the study of different types of drugs and their action is known as pharmacology. To understand the relationship between the drug compounds and therapeutic targets in traditional lab settings is not only time taking but also expensive. Hence, in silico computational methods are now being employed in order to study, predict and analyze the drug protein interactions, which is the fundamental step of genomic drug discovery, drug design and pharmacology. Machine learning techniques which involve the design of algorithms that can detect useful patterns from existing data are used to learn from available drug target interaction data and infer unknown interactions from different types of heterogeneous data sets related to these drugs and proteins.

1.2 Biological Terms and Definitions

In Silico drug target interaction prediction using machine learning techniques is fast gaining popularity. This section gives a simple introduction to the biology involved in this study.

The basic terms used throughout this work are defined in the context of the problem being addressed.

Drugs: Drugs can be defined as the chemical compounds which have bio-chemical or physiological effects on humans and other living organisms. In the view point of pharmacology, drugs can be defined as chemical substances which are man-made or endogenous used to prevent and cure diseases thus by enhancing physical and mental health of a living being.

Proteins targets: Proteins are bio molecules (functional modules) of living organisms formed by sequence of amino acid chains. In the context of pharmacology, a biological target (protein) can be defined as anything inside a living organism to which ligands or drugs bind. The commonly known biological target families in humans are enzymes, ion channels and receptors.

Drug target interaction networks: The proteins inside a living body are targeted by drugs in order to either enhance or inhibit the function carried out by that protein thus each drug targets certain specific set of proteins and this property of drugs can be studied and analyzed in terms of network topology where each node in a network represents a drug or a target proteins (drug target interactions network).

1.3 Thesis Outline

In this work we propose a model which predicts true drug target interactions based on the RLS model predicted in [27] using more sophisticated drug and target kernels and extend the result to weighted profile method to infer novel interactions for new drug (protein targets). The chapters are organised as follows: Chapter 1 gives an introduction to the

problem being solved and explains the basic terminologies used in this work. Chapter 2 starts by explaining the problem being solved, motivation to solve the problem and the contribution made by this work. In chapter 3, we discuss the background study required to understand the proposed method which includes previous methods employed to solve the problem, how these past methods tie up into the current model and related work which explains the concepts used in the proposed algorithm. Chapter 4, we define a problem framework and then explain the steps implemented in the current model. We also explain the techniques employed to evaluate the performance followed by the results obtained using the proposed method in Chapter 5, which clearly indicate that the small yet effective improvement yields a better performance than state of art algorithms. Chapter 6 we conclude by summarizing the method, it superior performance and we discuss some of the ways to explore in future which can improve the current model to make it more effective in inferring true drug target interactions.

CHAPTER 2

2.1 Drug Target Interaction Prediction Problem

All living organisms are made up of cells and different functions exhibited in the living body are because the protein molecules present in these cells. Any desired or an undesired effect in the body is because of the functions carried out by these proteins. The proteins inside a living body are targeted (bound) by drug compounds to enhance or inhibit functions carried out by proteins. Each drug target specific set of proteins inside the body of a living organism. This property of drugs can be studied and analysed by representing drug-target interactions using network graphs.

Huge amounts of publicly available chemical data and genomic data motivates multidisciplinary researchers to bridge the gap between biology and chemistry by integrating and analyzing the molecular information. For example, the Molecular Libraries Roadmap initiative taken up by National Institutes of Health is promoting the development of public databases such as PubChem [28] which is fostering new research areas related to pharmaceuticals. The available information on drugs and targets can be integrated to form high throughput data sets for understanding their relationship which help to discover new targets for existing drugs and novel targeting drugs for existing proteins. This not only helps in analyzing the therapeutic effects of unknown drugs but also adverse effects of known drugs. Therefore, many unknown so called off targets of these drug compounds can be significant data for further clinical trials in wet labs, crucial step in drug discovery and drug design.

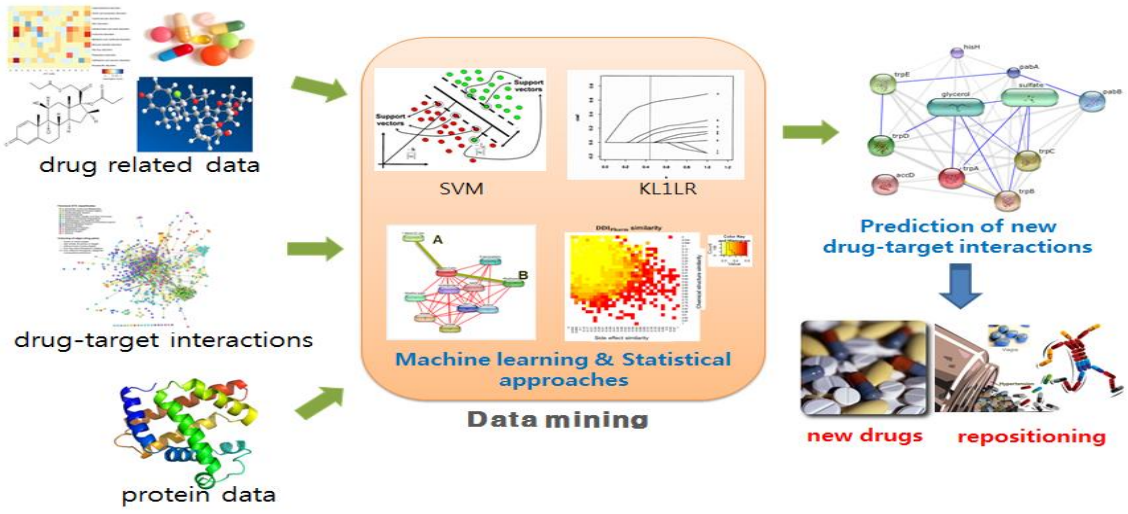


Figure 2.1 Similarity based drug target interaction inference model [33]

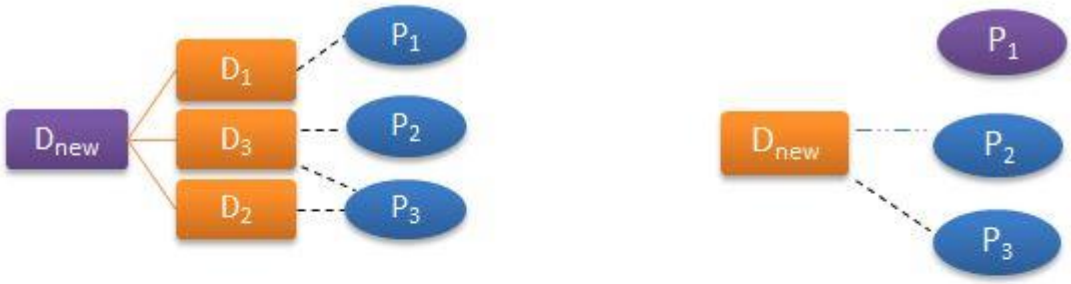
The figure-2.1 represents how recent machine learning methods are applied to infer novel drug target interactions by integrating different information of drug compounds and target proteins and using this data a model is trained to predict unknown interactions. With the recent advancement in technology and huge amount of biological and chemical data available through online public databases such as KEGG [12], BRENDA [22], SuperTarget [8] etc., there is a need for predicting drug target interactions accurately using computational methods. Different types of computational approaches have been proposed for predicting drug target interactions. Two well-known approaches are text mining [36] and docking simulations. Though these methods produced useful insights into drug target relationships, they have some serious limitations with respect to the type of data required to implement these methods for example, docking techniques require 3D structures of target proteins and scarcity of data available on 3 dimensional structures, this approach cannot be implemented on a large scale. Literature text mining on the other

hand requires keyword searching which is plagued by redundant names of drugs and proteins and also cannot be employed on novel findings in chemical and bio sciences.

Statistical learning algorithms are well known for the tasks such as classification and regression, data repositories such as KEGG [12], STICH [14], SuperTarget [8], Matador [8], BRENDA [22] and DrugBank [29] which store information about drug target interactions and other useful data such as information about genomic sequences, drug responses of target proteins, chemical structures, side effect information and molecular descriptors of drugs which can detect hidden drug protein interactions often prove to unearth adverse effects of drugs during drug design.

So, the data available is in the form of matrices. A drug target network can be represented using a bipartite graph where one set of nodes represent the drugs, other set of nodes represent the proteins being targeted and the edges connecting pair of nodes denotes the interactions. From this graph we can obtain an adjacency matrix in which each row represents a drug interaction vector and each column represents interaction vector of a target protein in which a known interaction between a drug target pair is set to 1 and unknown interactions are marked as 0. Applying regression, we can define a function over the relationship between the drug- protein interactions and similarity measures of individual drugs and targets respectively. Molecular and functional information available such as chemical structure similarity of the drugs, then similarity of drugs based on side effect keywords and genomic sequence similarity of the target proteins can be integrated for inferring the relationship between the nodes of the bipartite graph i.e., drugs and target proteins. The figure-2.2 gives the intuition of current machine learning models based on drug similarity. Let $D = \{D_1\ D_2\ D_3\ D_{new}\}$ represent set of drug compounds and $P = \{P_1\ P_2$

P_3} represent protein targets where the known interacting pairs are {(D_1P_1) (D_2P_3) (D_3P_2) (D_3P_3)} and we have to infer interactions of D_{new}. In the below figure we can see the corresponding adjacency matrix and drug similarity matrix of D_{new} with respect to all the drugs. Inferring interactions of D_{new} using nearest neighbor algorithm that D_{new} has highest probability of interaction with P3 as it nearest neighbor D3 interacts with P3 and as the similarity between drugs decreases the probability of sharing common target decreases. This is a simple method where the interaction with a given protein is weighted using the similarity of neighboring drugs and their interaction with the corresponding protein. All the machine learning methods proposed are based on a common assumption that similar drug compounds share similar target proteins and vice versa.



Given	P_1	P_2	P_3
D_1	1	0	0
D_2	0	0	1
D_3	0	1	1

Known Interactions

D_{new}	
D_1	0.1
D_2	0.5
D_3	0.7

Drug similarity

Then

	P_1	P_2	P_3
D_{new} =	0	0.23	0.40

Predicted Interactions

Figure 2.2 A graphical representation of nearest neighbor algorithm to infer novel drug target interactions

2.2 Research Motivation

The traditional method adopted in the pharmaceutical industry by testing the effects of known drug compounds systematically is considered as obsolete, superfluous amount of human genome data available publicly altered the industry's internal working model. As scientists can now employ a bottom-up approach, working through genomic data to find relationships between certain genotypes and diseases and then screening drug data to identify therapeutic candidates. The old paradigm of 'one drug-one target' is quickly becoming outdated. It was smart way of understanding the drug-protein interactions but the biological systems we are dealing with are made up of myriad of proteins exhibiting multiple functions. To analyze and understand these systems as a whole, we require efficient computational models, and then we will be able to predict the adverse effects of a drug or the therapeutic effects of a drug efficiently. The currently available drug target interactions across various data sources are not experimentally validated. So, there is a strong need to predict these interactions accurately.

This problem is being tackled from various angles. Methods such as docking simulations based on the *3D* structure of proteins, approaches based on machine learning algorithms using text mining and similarity based methods. As 3D simulation is time consuming and text mining medical documents for drug target relationships is affected with redundant names of genes and compounds the researchers started working with drug-drug and target-target relationships to understand drug-target interactions based on similarity measures.

So, current trend has shifted towards more simple yet sophisticated machine learning techniques based on statistical data analysis such as classification and regression using various similarity measures of drugs and proteins and known drug target interactions. The underlying assumption of these methods is that similar proteins are targeted by similar drugs and similar drug compounds tend to bind with similar proteins. This is an important property and very useful to infer interactions of drug target networks. The similarity between two given drugs can be described from various perspectives like physiochemical properties of molecules or number of 2D/3D structures, number of side effect keywords in drug package inserts and molecular descriptors like number of chemical double bonds etc. on the other hand similarity between two given proteins can be calculated based the sequence alignment of amino acid chains, gene ontology similarity or PPI closeness. Learning algorithms can be designed to integrate these different types of data available at disposal that can be used to predict drug target interactions accurately.

Hence, there is a strong need for design and development of new algorithms which can infer true drug target relationships effectively and our work is a small effort in this direction.

2.3 Thesis Contribution

The contribution of our work is two-fold: First we define more sophisticated kernels on different similarity measures of drugs and target proteins available using the radial basis function. Then integrate two different types of similarity measures for drugs namely chemical sub structure similarity and pharmacological similarity based on side effect key words and obtain a new kernel for drug compounds. We incorporate newly obtained kernels into Kernel Regularized Least Squares method to infer new interactions for

protein targets (drugs) with at least one known interaction. Second, we extend the result of RLS method, using the newly predicted interaction scores to infer the interactions of new drugs and proteins for which there is no known interaction data hence, solving the unknown candidate problem of method proposed by the authors in [27].

The whole process for predicting drug target interactions is divided into two steps: In the first step initially, we apply kernel regression using Regularized Least Squares algorithm to infer new interactions for proteins with at least one known interaction based on already known drug target interactions and kernels defined over chemical structure similarity and pharmacological similarity of drugs respectively then, use the predicted interaction profiles of neighboring proteins to infer interaction profile of target proteins for which all interactions are unknown. The neighbors for target proteins are defined by the score in genomic sequence similarity kernel matrix. Similar approach is followed for drugs initially we infer new interactions for drugs with at least one known interaction and use this new interaction profiles of neighbors to predict the interaction profile of unknown drugs (unknown drugs are the drug compounds with no prior known interaction). We have two independent predictions of same drug target pair and we use a weighted combination of those two scores to obtain a final interaction score. The interaction profiles of neighboring proteins of new target proteins obtained after KRLS algorithm are real valued scores unlike binary and thus be more useful in training phase of weighted profile method to predict true interactions. Thus, we improve the drug target prediction problem by using more informative Gaussian kernels defined over the similarity measures of drugs and target proteins and also by supplying the real valued labelled data (interaction scores) for the training phase of instance based leaning algorithm. The results

clearly indicate that these small yet significant improvements prove to be effective in predicting the true interactions accurately. We have analyzed the predictive power of proposed method using area under ROC curve [6] which is the plot of true positives and false positives in the prediction for varying thresholds.

CHAPTER 3

3.1 Previous Works

This section describes the various state-of-art methods proposed by eminent researchers in order to predict drug target interactions efficiently. The following paragraphs review these methods and how these methods tie up into the model proposed to tackle the drug target interaction prediction problem.

3.1.1 Yamanishi .Y et al, *Prediction of drug-target interaction networks from the integration of chemical and genomic spaces*. 2008.

The method proposed is a supervised bipartite graph learning method [32]. The authors build this model to understand the relationship between drug target network topology, drug chemical structures similarity and protein genomic sequences similarity. In this work they refer chemical structure similarities among all the drugs in the data as chemical space, the protein sequence similarities as genomic space and the existing drug target interactions to pharmacological space respectively. The goal is to infer unknown drug target interactions by integrating the chemical space and genomic space onto pharmacological space and they proceed by mapping these three spaces onto one single unified space such that the interacting drugs and targets are close to each other while non-interacting drug target pairs are placed further apart in the unified space. Once a mapping function is learned the unknown drug target pair is mapped onto this pharmacological space to infer the interaction score by determining how close the queried drug target pair is in the mapped pharmacological space. Known interactions are referred as ‘gold

standard' data and it is used as training set to infer unknown interactions and for evaluation in cross validation as well.

The steps followed in the current model are:

1. Map drugs and targets known to be interacting into 'pharmacological space'.
2. A model between chemical/genomic space and pharmacological space is learned and queried drug/proteins are mapped onto the unified space.
3. Infer new interactions among drug-target pairs by connecting the pairs which are closer than a given threshold in the pharmacological space.

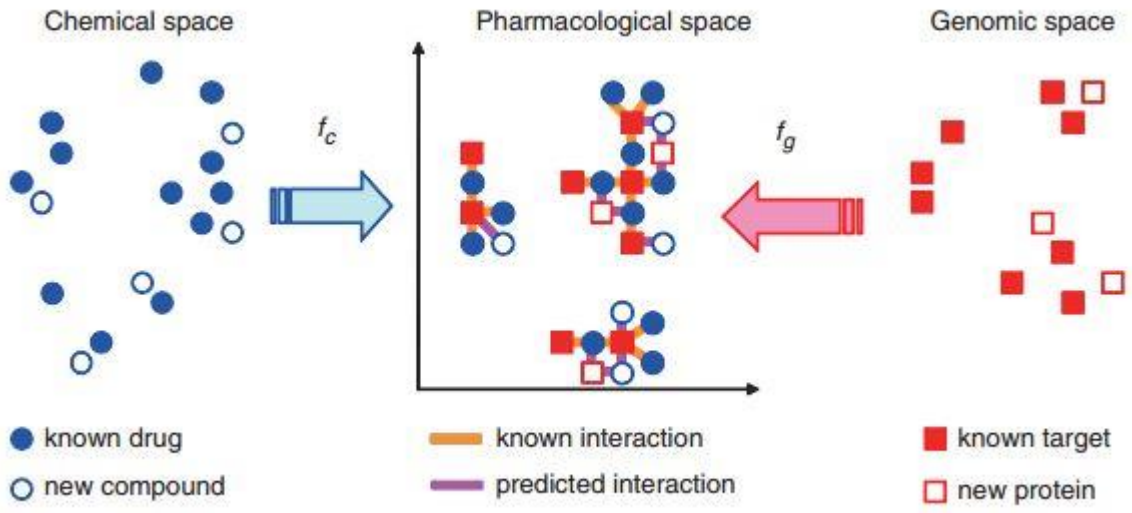


Figure 3.1 Representation of supervised bipartite graph inference method [32]

The figure-3.1 illustrates the idea of method proposed by the authors. The circles represent drugs, squares represent target proteins, colored ones are known and uncolored ones are new drugs and targets respectively. We can see that using corresponding mapping functions f_c and f_g the chemical space and genomic space are mapped onto pharmacological space.

The drug target interaction network is represented by a bipartite graph $G = (V_1+V_2, E)$ where V_1 is set of drugs, V_2 is set of proteins and E is set of interactions between interacting elements of V_1 and V_2. A kernel similarity matrix is calculated in the following way:

$$K = \begin{pmatrix} K_{cc} & K_{cg} \\ K_{cg}^T & K_{gg} \end{pmatrix} \quad (3.1)$$

Where $(K_{cc})_{ij} = exp(-d^2{}_{c_i c_j} / h^2)$, for i,j = 1, …, n_c, $(K_{gg})_{ij} = exp(-d^2{}_{g_i g_j} / h^2)$, for i,j = 1,…,$n_g$ and $(K_{cg})_{ij} = exp(-d^2{}_{c_i g_j} / h^2)$, for i =1, …, n_c , j = 1, …, n_g

Here d is the shortest distance between all objects on the graph and h is the width parameter optimize via cross validation experiments. An appropriate identity matrix is added to K to make it positive definite.

The dimensions of the matrix K is $(n_c + n_g)$ x $(n_c + n_g)$ it is not easy to operate on so, eigenvalue decomposition of K is computed as:

$$K = \Gamma \wedge^{1/2} \wedge^{1/2} \Gamma^T = UU^T \quad (3.2)$$

Where the diagonal elements of $\wedge$ are eigenvalues, columns of matrix Γ are eigenvectors, $U = \Gamma \wedge^{1/2}$ and row vectors of the matrix, $U = (\boldsymbol{u}_{c_1}, \ldots, \boldsymbol{u}_{c_{n_c}}, \boldsymbol{u}_{g_1}, \ldots, \boldsymbol{u}_{g_{n_g}})^T$. The space spanned by $\mathbf{u}_c$ and $\mathbf{u}_g$ is '*pharmacological space*'.

A kernel regression model is proposed for correlating chemical/genomic space and pharmacological space as:

$$\boldsymbol{u} = f(x, x_1) = \sum_{i=1}^{n} s(x, x_1)\boldsymbol{w}_i + \epsilon \quad (3.3)$$

Where s(· , ·) is a similarity score function, $\boldsymbol{w}_i$ is a weight vector and ϵ is a noise vector. Thus, we obtain the following loss function.

$$L = ||UU^T - SWW^TS^T||_F^2 \quad (3.4)$$

S is *n* x *n* similarity matrix, W = $(\mathbf{w}_1,\ldots,\mathbf{w}_n)^T$ and $||.||_F$ is Frobenius norm.

Two models f_c and f_g are learned one for c_{new} and the other for g_{new} respectively. The scores for three types of drug target pair are calculated by inner product as follows

(1). $corr(c_{new}, g_j) = \mathbf{u}_{cnew} \cdot \mathbf{u}_{gj}$ (3.5)

(2). $corr(c_i, g_{new}) = \mathbf{u}_{ci} \cdot \mathbf{u}_{gnew}$ (3.6)

(3). $corr(c_{new}, g_{new}) = \mathbf{u}_{cnew} \cdot \mathbf{u}_{gnew}$ (3.7)

Finally high scoring drug target pairs are predicted to interact with each other by measuring the distance in the unified space. High scoring pairs are close to each other in the pharmacological space. A more detailed derivation of the model can be obtained from the paper.

SIMCOMP is used to calculate the chemical similarity score between two compounds and normalized Smith-Waterman score is computed to measure the sequence similarity for targets. The authors conclude by claiming that their method is first of its kind in formalizing drug target interaction prediction problem as supervised learning problem for a bipartite graph and also in the integration of chemical and genomic spaces. They also suggest one or two important improvements that can be made to the current model like usage of more sophisticated kernel similarity functions for drug and target similarities and incorporating additional information related to drugs and proteins in the data to improve the efficiency of a model and also accuracy in inferring true drug target interactions.

3.1.2 Campillos .M et al, *Drug Target Identification Using Side-Effect Similarity*. 2008.

A propitious method based on the use of 'phenotypic side effect similarities' to predict drug target interactions has been proposed by the authors. Drug perturbations occur when the given drug not only interacts with its primary target but also with the additional off-targets in an organism. The surprising behaviour of these drugs by targeting off-targets is generally considered harmful but in some cases it has been proved to be a good one, leading us to new therapeutic results of known drugs. "Similar side effects of unrelated drugs can be caused by their common off-targets" [2]. Identification of additional targets through chemical similarity and docking simulations has been effective in a smaller perspective of human system. This has motivated the authors to explore the use of side effect information to predict drug target interactions and in this process they computed a side effect similarity measure for marketed drugs and analyzed the likelihood of two given drugs sharing protein targets. Their study has shown that two given drugs even though dissimilar based on their chemical structure, their high side effect similarity is one of the reasons to share common targets and thus identified additional targets for known drugs. Drug package inserts of marketed drugs were used to extract side effect information and they have classified side effects based on Unified Medical Language System ontology. The method proposed here does not actually predict drug target interactions instead it calculates the probability of a given drug pair to share a common target. This probability is calculated based on two drug-drug relationships namely chemical similarity and side-effect similarity, calculated based on relevant terms in the ontology by fitting a different functions to the corresponding data.

The best fitting function to infer the probability of sharing a common target based on the chemical similarity is given by:

$$\boldsymbol{P}_{2D}(y) = (1 + e^{(B-y)/A})^{-1} \quad (3.8)$$

Where P_{2D} is the probability of sharing a common target as a function of chemical similarity of given drug pair. A (6.91) and B (0.68) are the parameters of the sigmoid function. Chemical similarities were ranked and the function is fitted based on the percentiles [2].

In the second case, the probability of sharing a common target for a given drug pair is calculated based on side effect similarity measure computed using the logarithmic percentiles of their ranked similarity values as follows:

$$\boldsymbol{P}_{SE}(x) = A \cdot x + B \quad (3.9)$$

Where P_{SE} is the probability of sharing a common target for a given drug pair, A (-0.084) and B (0.047) are parameters fitted to the function.

In the third case the combination of two similarity measure is used to calculate the probability and the sigmoid function fitted is as follows:

$$P_{SE,2D}(x,y) = H \cdot (1 + e^{\left(A + D.(\arctan \frac{y}{1-x})^{E}\right).\left(B + F.(\arctan \frac{y}{x})^{G} - \sqrt{(C-y)^2+(1-x)^2}\right)})^{-1} \quad (3.10)$$

$P_{SE,\,2D}$ represents the probability of a given drug pair to share a common target with side effect similarity *P* value and chemical similarity. The parameters A (0.0167) B (55.507) C (-80.16) D (-129.6) E (455.6) F (617.3) G (0.415) H (0.83) are fitted accordingly.

The authors noted that the integrated method improved the sensitivity and specificity measures and conclude that side effect similarity of drugs indeed has a strong correlation towards the probability of sharing common targets and should be further explored.

3.1.3 Jacob .L et al, *Protein-Ligand Interaction Prediction: an improved chemogenomics approach.* 2008.

The authors propose a method to combine the chemical and biological information which infer interactions of a small molecule with any given targets. Support vector machine algorithm is used for prediction in a combined space by training the SVM with known drug target interactions to predict interactions for targets with no known ligand information from ligand similarity and target similarity kernels. The SVM classification algorithm makes use of a pair-wise product kernel, a combination of ligand similarity kernel and kernel(s) of target proteins by supplying known labeled information of drug-target pairs. They have used different types of protein similarities.

The drug-target pair-wise similarity is computed as follows:

$$s((d,p),(d_i,p_j)) = s_d(d,d_i) \cdot s_p(t,t_j) \quad (3.11)$$

And the SVM classifier is trained using this information to infer unknown drug-target interaction scores. It is one of the efficient machines learning approaches, as only one single classifier has to be trained in order to infer any number of unknown drug target interactions. The basis of this method is to represent a pair of drug target by vector and estimate a linear function fit to the data whose sign will predict the ligand protein interaction [11]. The limitation of this method surfaces when we are dealing with large

datasets for example, if a dataset has m drugs and n proteins, the size of kernel matrix is (m x n) and considering a case where n = 700 and m = 900 due to limitations of memory all the instances cannot be used in training and in worst case we have stick to random sampling of negative instances and not much efficient practically [5].

3.1.4 Xia .Z et al, *Semi-supervised Drug-Protein Interaction Prediction Using Heterogeneous Spaces.* 2009.

All network based supervised prediction algorithms works only with the help of labeled data supplied to it in the training phase. Sometimes unlabeled data can also be helpful in revealing hidden drug target interactions in the network. Authors propose a semi-supervised prediction framework to infer unknown drug target interactions on a large scale. As they tackle the major issue of biological databases, having very little validated data regarding the target information of drug compounds. The semi-supervised algorithm integrates known drug target interaction information with chemical structure similarity and genomic sequence similarity of drugs and target proteins respectively as they successfully move past the traditional classification methods which infer interactions of a single given protein based on the chemical structure similarity of drugs in the data. The current model is based on the Laplacian regularized least square algorithm which predicts interaction for drugs and targets separately and the individual scores are combined to yield an average score for the interactions. General LapRLS is improved using a kernel based on the known drug target information and is referred as NetLapRLS in this work.

"*In LapRLS and NetLapRLS, the data-dependent regularization terms are normalized Laplacian operation on graphs*".[31] A data dependent model using geometry of

probability distribution is implemented during which two individual classifiers using LapRLS are trained using chemical similarity and genomic similarity matrices and then are combined using average function. Two types of proposed method are employed first one is the standard LapRLS which is trained using only chemical and genomic spaces and the second is the extended version of the first NetLapRLS where training the model incorporates drug target interaction information along with chemical and genomic spaces.

The prediction function with respect to drug interaction domain is as follows:

$$F_d^* = W_d \, (W_d + \beta_d L_d W_d)^{-1} \, Y \tag{3.12}$$

The prediction function with respect to target interaction domain is as follows:

$$F_p^* = W_p \, (W_p + \beta_p L_p W_p)^{-1} \, Y^T \text{ and} \tag{3.13}$$

The two individual predictions are combined to obtain the final interaction for drug target pairs with the help of average function:

$$F^* = (F_d^* + F_p^*)/2 \tag{3.14}$$

W_d (W_p) is the drug (target) domain similarity which is a linear combination of chemical structure (genomic sequence) similarity and drug target interaction network similarity of drug compounds (target proteins) respectively. L_d (L_p) is normalized graph Laplacian of drug (target) domain and Y is the drug target interaction network

The authors highlight that the method employed did not use any negative samples to predict the new drug target interactions. They suggest some ideas to improve their model

by including more sophisticated kernels and more informative biological kernels such as drug side effects information.

3.1.5 Yamanishi.Y et al, *Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrate framework*. 2010.

The main objective of this work is to analyze and understand the relationship between chemical space, pharmacological space (drug side effect similarity) and drug target interaction space. The authors based on their results obtained using supervised bipartite method show that the drug target interactions are more related to drug side effect similarity (referred to as pharmacological effect similarity) than chemical structure similarity of drug compounds.

The supervised model implemented in predicting drug target interactions is based on the integration of chemical, pharmacological data of drug compounds and genomic sequence information of protein targets. In order to predict the new interactions for drug compounds which are not yet marketed i.e. (without any side effect information), the authors implement a method to predict pharmacological effect similarity from chemical structure similarity. Then, the predicted pharmacological effect similarity is integrated with chemical structure similarity of drugs and genomic sequence similarity of targets to infer the drug target interaction on a large scale. The proposed method has two steps: 1). Prediction of pharmacological similarity of drugs based on their chemical structure similarity with respect to other drug compounds and 2). Prediction of unknown drug target interactions based on the pharmacological effect similarity of drugs. The authors

claim that this is the first instance where pharmacological effect similarity is integrated with chemical and genomic spaces for inferring unknown drug target interactions.

Pharmacological data is obtained using Japan Pharmaceutical Information Center database which manages the drug package inserts in Japan, approved by Health and Welfare Minister of Japan. The entries of JAPIC database are in XML format with package inserts are categorized using keyword tags and the authors extracted them and based on 'pharmaceutical effect' tag, similarity between two drugs is calculated based on the frequency of the keyword in the data. The similarity matrix obtained is referred as 'pharmacological space' in this work.

In the initial step of the proposed method the prediction of pharmacological effects from chemical structures of drugs, a similarity matrix regression model is formulated based on two similarity matrices **C** and **P** using training set and prediction set:

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C_{tt}} & \boldsymbol{C_{pt}^T} \\ \boldsymbol{C_{pt}} & \boldsymbol{C_{pp}} \end{pmatrix} \tag{3.15}$$

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P_{tt}} & \boldsymbol{P_{pt}^T} \\ \boldsymbol{P_{pt}} & \boldsymbol{P_{pp}} \end{pmatrix} \tag{3.16}$$

In the above equations *p* represents prediction set and *t* represents training set.

$S_{phar}(\boldsymbol{y}, \boldsymbol{y'}) = f(\boldsymbol{x}, \boldsymbol{x`}) + \epsilon = \boldsymbol{u(x)}^T \boldsymbol{u(x`)} + \epsilon$, where (3.17)

$$u(\boldsymbol{x}) = \sum_{j=1}^{n} s_{chem}(\mathbf{x}, \boldsymbol{x_j})\beta_j \tag{3.18}$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots\ldots, \beta_n)^T$ is a weight vector and *n* is the number of compounds in the training set. A detailed derivation of the method can be found in the paper [33].

Therefore, using this method unknown pharmacological effects similarity for drugs are inferred from the chemical structure similarity and the authors mention that even though there exists a relationship between chemical structure similarity and side effect similarity of drug compounds, there are handful of exception cases where drugs with high chemical structure similarity have low pharmaceutical effect similarity. In the second step of the this framework unknown targets of drugs are inferred using supervised bipartite graph model based on distance learning algorithm using pharmacological information of drugs and genomic sequence similarity of targets. A similar method was employed in [32] but with chemical similarity of drugs.

$$g(y, z) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}\, s_{phar}(y_i, y)\, s_{geno}(z_j, z) \tag{3.19}$$

Where *y* represents new compound, *z* represents a target, *n* represents number of drugs, m represents number of targets in the training set, $s_{phar}(\cdot, \cdot)$ represents pharmacological similarity function, $s_{geno}(\cdot, \cdot)$ represents genomic sequence similarity function and α_{ij} are the parameters learned.

The method can be summarized as follows:

1. Known drug target interactions are mapped onto a unified space in which interacting drug target pairs are placed closed to each other.
2. A correlation model is learned between pharmacological effect similarity of drugs, genomic sequence similarity of targets and unified space then new drugs are

mapped onto to the unified space based on pharmacological similarity and new proteins are mapped onto unified space using genomic sequence similarity.

3. If *g(y, z)* is greater than a threshold value in the unified space, then *y* is predicted to be interacting with *z*.

Since, the authors wanted to study the relationship between pharmacological effect similarity of drugs and drug target interactions the method is implemented for new drug candidate compounds. They also note that the drug target interactions are more related to pharmacological similarity of drugs than chemical structure similarity. Finally, the authors conclude by saying that the performance of the model implemented can be improved by the use of more sophisticated kernels designed for chemical similarity and genomic similarity.

3.1.6 Laarhoven et al. *Gaussian interaction profile kernels for drug-target interaction.* 2011.

The authors employ kernel regularized least squares classifier using Gaussian interaction profile kernel for prediction of new interactions for drugs and targets with at least one known interaction. The authors show that network topology of drug target interactions pose to be a very good source for inferring new interactions and propose two variations of RLS method. First, the prediction of new interactions for drug compounds and target proteins is implemented separately using target kernel and drug kernel respectively and the final interaction scores for a pair are computed using an average function. New kernels are employed in both individual predictions which are defined over drug interaction profile vectors and target interaction profile vectors respectively. An

interaction profile is a binary vector with each entry (0 or 1) represents presence or absence of interaction with corresponding drug or target protein. *Y* an adjacency matrix for drug target interaction network is defined in which each row represents target interaction profile vector of a drug and each column represents drug interaction profile vector for corresponding target. A Gaussian kernel is constructed for drugs and proteins from these interaction profile vectors and drug-drug, target-target similarity matrices are obtained. The kernel matrix score for drugs d_i, d_j is calculated as follows:

$$K_{GIP,d}(d_i, d_j) = exp(-\gamma_d \left|\left| y_{d_i} - y_{d_j} \right|\right|^2) \quad (3.20)$$

Here γ_d controls the kernel bandwidth. The parameter is normalized, dividing it by the average number of interactions per drug in the following way: $\gamma_d = \Phi_d / (\frac{1}{n_d} \sum_{i=1}^{n_d} |y_{d_i}|^2)$ where, $\Phi_d = 1$. A kernel for targets is constructed using drug interaction profiles in a similar fashion.

In the next step kernels are constructed from chemical structure similarity matrix S_d and genomic sequence similarity S_g by making them symmetric and adding a small multiple of identity matrix to impose positive definite property and are denoted by $K_{chemical,d}$ and $K_{genomic,t}$ respectively. Further integration of chemical similarity kernel and genomic similarity kernel with respective Gaussian interaction profile kernel(s) is computed using weighted average function as follows:

$$K_d = \alpha_d K_{chemical,d} + (1 - \alpha_d) K_{GIP,d} \quad (3.21)$$

$$K_t = \alpha_t K_{genomic,t} + (1 - \alpha_t) K_{GIP,t} \quad (3.22)$$

Where $\alpha_d = \alpha_t = 0.5$. The above kernels are used in the regularised least squares classifier defined for drugs and targets individually. The RLS classifier for predicting new interactions for drug compounds using target similarity kernel and the RLS classifier for predicting new interactions for targets using drug similarity kernel are combined to obtain final drug-target pair interaction scores. The equation below represents a closed form solution as follows:

$$\hat{Y} = \frac{1}{2}(K_d(K_d + \sigma I)^{-1}Y) + \frac{1}{2}(K_t(K_t + \sigma I)^{-1}Y^T)^T \quad (3.23)$$

Y is the adjacency matrix of drug target interaction network. $\hat{Y}$ is the predicted value(s) and σ = 1. This method is referred as RLS-avg.

Another way of combining drug and target kernels is employed by the authors using the Kronecker product kernel i.e. $K = K_d \otimes K_t$ of drug and target kernels. The prediction of interactions for all pairs is computed in a single step as shown:

$$vec(\hat{Y}^T) = K\,(K + \sigma I)^{-1}\, vec(Y^T) \quad (3.24)$$

Where $vec(Y^T)$ is a vector of all interaction pairs obtained by stacking the columns of Y^T and refer this as RLS-Kron in this study. A better implementation of this method based on Eigen decomposition is presented for the kernel as:

$$K = K_d \otimes K_t = V \wedge V^T \quad (3.25)$$

Where $V = V_d \otimes V_t$ and $\wedge = \wedge_d \otimes \wedge_t$ are the Eigen vectors and Eigen values respectively. Now the closed form solution for prediction is:

$$\hat{Y} = V_d\, Z^T\, V_t^T \quad (3.26)$$

The vector Z is calculated as:

$$vec(Z) = (\wedge_d \otimes \wedge_t)\ (\wedge_d \otimes \wedge_t + \sigma\, I)^{-1}\ vec(V_t^T\ Y^T\ V_d) \quad (3.27)$$

The Eigen decompositions reduce the runtime considerably. With three different kernel combinations possible by altering the α_d [0, 0.5, 1] and α_t [0, 0.5, 1] the two RLS classifiers are implemented accordingly. In the case where $\alpha_d = \alpha_t = 0$ the prediction is obtained based on chemical structure and genomic sequence similarity only. When $\alpha_d = \alpha_t = 1$, the predictions are obtained based on Gaussian interaction profile kernels of drug and targets respectively and when $\alpha_d = \alpha_t = 0.5$, the prediction scores are obtained based on the average of the two types of kernels.

The authors claim that the best results are obtained for both RLS-avg, RLS-Kron classifiers using the combination of GIP kernels with chemical and genomic kernels when compared with predictions based on the individual kernels alone. The authors mention that the information from the known drug-target interaction network is an effective source of information for predicting true drug target interactions. They further indicate that the method implemented in this work is applicable for a drug or a target which has at least one prior known interaction in the drug target network. And also suggest a few improvements to further improvise the current method by using other sources of information about the drugs and targets involved.

3.1.7 Gönen. M. *Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization*. 2012.

The method proposed by the author is the first of its kind employed for inferring drug target interactions based on full probabilistic model. To make the network inference more efficient the method makes use of variational approximation based on covariance calculations. Drug compounds and target proteins are projected onto a unified space based on Bayesian formulation using chemical structure similarities for drugs and genomic sequence similarities for proteins. The method is a combination of non-linear dimensionality reduction based on kernels, matrix factorization and binary classification techniques in order to predict drug target interactions. Integrated Bayesian formulation of projecting drugs and targets into a unified space is the fundamental idea behind the whole approach.

Computational approaches such as docking simulations are downplayed by the author as they require 3D structural information of targets to infer interactions. Ligand based methods according to authors do not perform up to the mark when a target with very limited ligand information is queried, literature text mining techniques are based on keywords and are affected with redundant names due to non-standardized naming is also noted by the author and finally, the author(s) point out that the general approach of binary classification for inferring networks with the use of pair wise kernels is computationally complex when a large number of drug target pairs are involved.

Bayesian matrix factorization is improvised by formulating a full conjugate probabilistic model and the method uses a deterministic variational approximation technique to infer

interactions in the network which is based on the fundamental idea of projecting drugs and proteins onto a unified space with the help of chemical similarity and genomic similarity kernels.

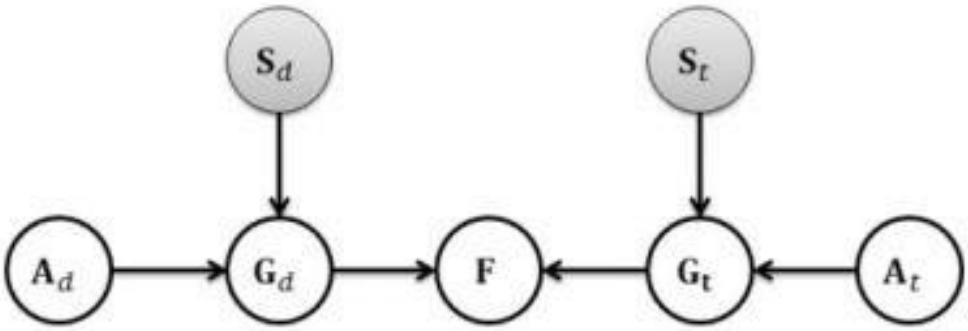


Figure 3.2 Representation of Bayesian matrix factorization method

The method proposed is represented by the figure-3.2 [7]. Which shows A_d and A_t parameter matrices used to project drug similarity kernel matrix S_d and genomic sequence similarity kernel matrix of target proteins S_t into corresponding low dimensional spaces G_d and G_t respectively to estimate an interaction matrix Y generated from score matrix F. Normal and gamma distributions are used in the probabilistic model to improve the efficiency because of conjugacy between them and by considering random variables as deterministic values, the score matrix is decomposed as follows:

$$F = G_d^T G_t = (A_d^T K_d)^T (A_t^T K_t) = K_d^T A_d A_t^T K_t \quad (3.28)$$

In here K_d represents drug similarity kernel which is nothing but S_d and K_t represents target similarity kernel which is S_t , and we can notice that by parameterizing the projected G_d and G_t low dimensional matrices in terms of kernel matrices K_d and K_t, the factorization of F is carried out enabling to make predictions for out-of-sample points.

The prediction is done based for three scenarios and in the first case, inferring interactions for a given new drug d_* for the set of targets X_t , the initial step is to compute the similarities between $d_*$and the drug in the set X_d : $k_{d,*} = [k_d(d_*, d_1)\ k_d(d_*, d_2)\ \dots\ k_d(d_*, d_{N_d})]^T$ where N_d is the number of drugs in set X_d.

The interaction scores are computed as follows:

$$f^* = (\tilde{A}_d^T k_{d,*})^T (\tilde{A}_t^T K_t) = k_{d,*}^T \tilde{A}_d \tilde{A}_t^T K_t \quad (3.29)$$

Where positive values are predicted to be interacting targets with d_*.

In the second case, new interactions are predicted for a new protein $t_*$from the set of drug compounds X_d, the similarities between t_* and set of target proteins X_t is obtained as following: $k_{t,*} = [k_t(t_*, t_1)\ k_t(t_*, t_2)\ \dots\ k_t(t_*, t_{N_t})]^T$

And the interaction scores are calculated as shown:

$$f_* = (\tilde{A}_d^T K_d)^T (\tilde{A}_t^T k_{t,*}) = K_d^T \tilde{A}_d \tilde{A}_t^T k_{t,*} \quad (3.30)$$

Where positive values are predicted to be interacting drugs with protein t_*.

In the third scenario, prediction is done for a new drug target pair which can be thought of as a combination of the first two methods and to infer an interaction score between d_* a new drug and a new target t_* and is computed as follows:

$$f_*^* = (\tilde{A}_d^T k_{d,*})^T (\tilde{A}_t^T k_{t,*}) = k_{d,*}^T \tilde{A}_d \tilde{A}_t^T k_{t,*} \quad (3.31)$$

Where a positive value is voted as a presence of an interaction between the new drug target pair.

The problem with the method proposed is in estimating three matrices A_d, A_t and F using an iterative process initialized to random variables which tend to be inefficient while dealing with larger data sets. The authors point out that the method can be further improvised by introducing other kernel functions defined over drug-drug similarity or protein-protein similarity such as side effect similarity for drugs and also the method can be extended to use multiple kernel learning.

In the above paragraphs, we have discussed existing machine learning methods proposed to solve the problem of inferring true drug target interactions. All the methods work on the same fundamental assumption that similar drugs tend to target similar set of target proteins and similar proteins tend to be targeted by similar set of drug compounds. The similarity between any two given drugs can be obtained by comparing those two compounds with respect to their chemical structures, side effects, number of carbon bonds, molecular and other functional information. Along the same line two given proteins can be compared for similarity with respect to their genomic sequences, gene ontology annotations and protein-protein interaction network distances etc. The success of proposed methods and techniques in this area has unearthed many new interactions and has taken drug design, drug discovery to a whole new level but there is still a lot left to be done to design efficient machine learning techniques which can detect true drug target interactions with high accuracy and continual progress to employ new approaches or improvise existing methods is required. Our work makes an effort to modify and extend the method proposed in [27]. The major issue with the method is that it is not applicable for new drugs or targets with no prior interaction information. So, we extend the RLS algorithm to weighted profile method. To be clear we make three major improvements to

the method: i).We use more sophisticated kernel functions defined over drug-drug similarities and target-target similarities ii). We integrate the drug kernel K_d with more effective pharmacological effect similarity of drug compounds and iii). We extend the result of RLS algorithm to weighted profile method thus enabling us to infer novel interactions for new drugs or targets with no known interactions.

3.2 Background Concepts

The following section prepare the reader to understand the propose method more clearly.

Supervised machine learning prediction methods involve designing a model which can predict an outcome based on the previous knowledge about the data. The model is represented in terms of a response variable and independent variable and the task is to predict the outcome of response variable based on the independent variable. A variable can be numeric/continuous or discrete valued and there are two types of modeling techniques based on the type of response variable, i). Classification: a model in which the goal is to predict a discrete valued output i.e. the process where the task is to separate the data into classes pertaining to the response variable in typical settings there are two classes 0 or 1 ii). Regression: a model in which the goal is to predict a continuous or numeric valued output. By applying a threshold function a regression model can be modified to represent a classification model. Unsupervised learning does not rely on the labels of known data samples (training data) instead by discovering the inherent patterns in the data tries to predict the output value for unknown data instances. Different types of classification models are which include, linear classification models, quadratic classifiers, decision trees, neural networks and support vector machines etc. The models based on

kernels functions produce real valued prediction scores and they make use of classifiers for example, binary classifier such as decision tree to infer the class labels.

In the following sections we discuss the models based on nearest neighbor classification and least squares method as these form the basis of the proposed model.

3.2.1 *k*- Nearest neighbor classification method

The nearest neighbor algorithm is one of the most used classification algorithm mainly due to its ease of use and works based on distance measure. It stores all the available class labels and when a new data point is queried it classifies the query point based on the majority vote of its *k* nearest neighbors measured with a distance function (or similarity). An important thing to note is that in the case of distance, the smallest value in the training set corresponds to the nearest neighbor of the unknown data sample and when dealing with similarities, the highest value in the training sample corresponds to the nearest neighbor of the queried sample in the data.

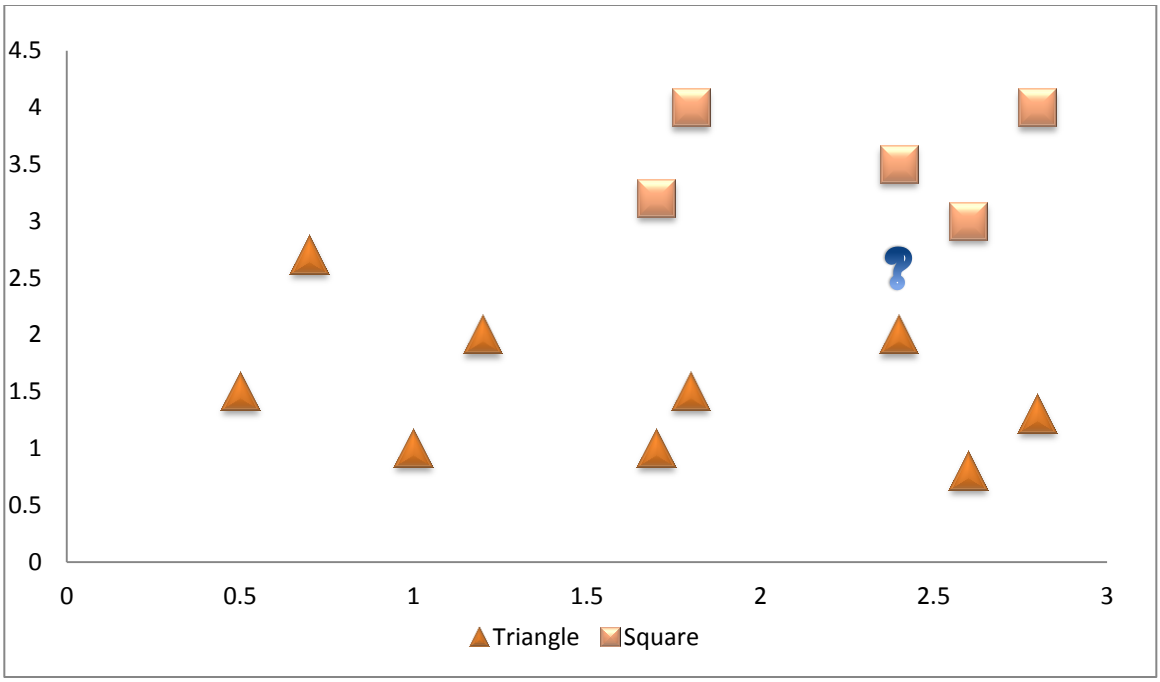


Figure 3.3 Representation of *k* – nearest neighbor method

In the above shown figure there are two classes based on the shapes triangle and square. Let the question mark point be the query data point and the task is to predict or classify to which class it belongs based on the *k*-nearest neighbor rule. Depending on the value of *k* we may get different output. In the figure when *k* = 1, the point will be classified as triangle and on the other hand when *k* = 3, the query point will be classified as a square. So, the value of *k* is an important parameter of *k*-NN algorithm.

A formulation of the *k*-nearest neighbor method:

Let the instances be a m-dimensional feature vector $y = (y_1, \ldots, y_m)$ and a mapping function $f(y_i)$.

During the training phase it stores all the known data samples $<y_i , f(y_i)>$ in the memory and in the classification (regression) phase, given a query point y_p, it first locates nearest data point in the sample training data y_m and estimate $f^{|}(y_p) \leftarrow f(y_m)$ when *k* = 1 and

When $k > 1$ given y_m, take a majority vote based on its k nearest neighbors if it is a discrete valued function or take the mean of f values of k nearest neighbors if it is a continuous valued function (regression) as follows:

$$f^{|}(y_p) \leftarrow \frac{\sum_{i=1}^{k} f(y_i)}{k} \tag{3.32}$$

The main aspects of k-nearest neighbor algorithm are the distance (similarity) function and the value of k. The k-NN classification is also known as lazy classification method.

3.2.2 Weighted *k*-Nearest neighbor method

This method is similar to the above approach and is used when we want to weight the neighbors of the query point more heavily.

In the training phase it stores all the known data samples $<y_i , f(y_i)>$ in the memory and in the classification (regression) phase for categorical valued function $f : \mathfrak{R}^n \rightarrow$ A, given a query data point y_p and its nearest neighbors denoted by $y_1 , \ldots , y_k$ we get:

$$f^{|}(y_p) \leftarrow \arg max_{a \in A} \sum_{i=1}^{k} w_i \, \beta(a, f(y_i)) \tag{3.33}$$

Where $\beta(x, y) = 1$ if $x = y$ or $\beta(x, y) = 0$ if $x \neq y$, $w_i = \frac{1}{d(y_p, y_i)^2}$ and

For continuous valued function $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$, given the query point y_p to be regressed:

$$f^{|}(y_p) \leftarrow \frac{\sum_{i=1}^{k} w_i \, f(y_i)}{\sum_{i=1}^{k} w_i} \tag{3.34}$$

Where $d(y_p , y_i)$ is the distance between y_p and y_i . In this case we can use all the training samples i.e. $k = m$.

3.2.3 Introduction to kernels and kernel based methods

Machine learning methods dealing with classification and predictions require data to be represented as feature vectors but kernel methods makes use of kernel functions, which deal with similarities of data points. Kernel functions enable us to operate in a high dimensional feature space without the need for calculating the coordinates of data points in that space instead by computing the inner products of projections of all data pairs in the high dimensional feature space. This is called kernel trick and using this any linear model can be transformed into a non-linear model where features are replaced by kernel function. A function that returns the dot product between the projections of two data points is called a kernel function and can be formulated from a feature map as $\phi: B \rightarrow F$, where *F* is a Hilbert space known as the Feature space. Feature map defines a kernel as:

$$\langle b1, b2\rangle \leftarrow K(b1, b2) = \langle \phi(b_1), \phi(b_2)\rangle \quad (3.35)$$

Where ϕ is mapping function.

Some examples of kernel functions are given below:

i. Linear kernel $K(b_1,b_2) = \langle b_1,b_2\rangle$

ii. Polynomial kernel $K(b_1,b_2) = (\beta (b_1,b_2) + 1)^d$, $\beta \in \Re$, $d \in N$

iii. Radial basis function kernels: This kernels satisfy $K(b_1,b_2) = K(||b_1 - b_2||)$

   i. Gaussian kernel: $K(b_1,b_2) = e^{\frac{||b_1-b_2||^2}{2\mu^2}}$, $\mu > 0$

   ii. Laplacian kernel: $K(b_1,b_2) = e^{\frac{||b_1-b_2||}{\mu}}$, $\mu > 0$

Using available kernel(s) we can make new kernels or combine two individual kernels to yield a new integrated kernel. The set of kernels have a closure property under certain operations. For example, given two kernels K_1 and K_2:

$K = K_1 + K_2$ is a kernel

$K = c\,K_1$ *or* $K = c\,K_2$ is a kernel for $c > 0$

$K = a\,K_1 + b\,K_2$ is a kernel provided $a, b > 0$ and many more

We can make different sophisticated kernels using simple kernels but modularity is important in this process. Kernel methods stores a training example by earning its corresponding weight and prediction for test set inputs are inferred by using a similarity function (kernel) between the test instance and all the training inputs.

A kernel is a similarity function $k(b_1 , b_2) > 0$ is the similarity of b_1, $b_2 \in \chi$ and feature representation f, $f(b) = (f_1(b), \dots , f_m(b))$ is a feature vector which defines a kernel according to mercer's theorem as follows:

$$k(b_1 , b_2) = f(b_1) \cdot f(b_2) = \sum_{i=1}^{m} f_i\,(b_1) f_i\,(b_2) \qquad (3.36)$$

Feature based methods and kernel based methods are interchangeable mathematically as feature and kernel representations are duals. Features based learning algorithms map features into feature space and learn the stats of features whereas kernel based learning algorithms use similarities between test set and training set and learn stats of training data.

Given below is an example for a kernelized binary classifier which computes a weighted sum of similarities as follows:

$$\hat{a} = sgn \sum_{i=1}^{n} w_i a_i k(b_i, b') \quad (3.37)$$

Here (b_i, a_i) is the i^{th} training example, $b'$is the unlabeled test input, $\hat{a} \in$ {-1, +1} is the prediction label of binary classifier for the unlabeled input b', $k : \chi$ x $\chi \rightarrow$ R is a kernel function measuring similarity between any given input pair b $\in \chi$ and $b' \in \chi$.

$S = \{(b_i, a_i)\}_{i=1}^{n}$, is the training set where $a_i \in$ {-1, +1} and $w_i \in$ R are the weights of corresponding training examples. The *sgn* function determines the outcome to be positive or negative.

3.2.4 Least squares method

A regression problem is to find a function that is best fit of given labeled data and use the fitted function to predict the value of test data, in a geometric sense it refers to hyper-plane fitting the given data points and in the case of linear regression the task is to find a linear function that best fits the data. The optimal solution for this problem known as least squares is to find a line that minimises the sum of squares of distances from training data points. And regression methods can also be used for solving classification problems by introducing a threshold function.

Let us assume that a training set B with $b_i \in \chi \subseteq R^n$, $a_i \in A \subseteq R$, We have to find a function *f* that interpolates the data

$$a = f(\boldsymbol{b}) = \langle \boldsymbol{w} \cdot \boldsymbol{b} \rangle + c \quad (3.38)$$

According to the least squares method we have to choose the parameters (**w**, c) such that it minimizes the sum of squared deviations of the data calculated using the square loss function *L* as follows:

$$L(\boldsymbol{w}, c) = \sum_{i=1}^{m}(a_i - < \boldsymbol{w} \,.\, \boldsymbol{b_i} > - c)^2 \quad (3.39)$$

Loss for a particular choice of parameters by sum of squares is obtained and the amount of loss associated can be computed using a variety of loss functions. Given, $\hat{a}$ $\hat{\boldsymbol{w}}$ = (w', c') and $\hat{B} = \begin{pmatrix} \hat{b}_1' \\ \vdots \\ \hat{b}_m' \end{pmatrix}$ where $\hat{b}_i' = (b_i', 1')$, the loss function *L* is minimised by differentiating it with respect to parameters (**w**, c) and equating the resulting expressions to zero as shown:

The loss function *L* is represented as:

$$L(\hat{\boldsymbol{w}}) = (a - \hat{B}\hat{\boldsymbol{w}}')\,(a - \hat{B}\hat{\boldsymbol{w}}) \quad (3.40)$$

By equating the derivative of loss to zero we get the following normal equation

$$\hat{\boldsymbol{B}}' \,\hat{\boldsymbol{B}}\, \hat{\boldsymbol{w}} = \hat{\boldsymbol{B}}' \,\boldsymbol{a} \quad (3.41)$$

By taking the inverse of $\hat{\boldsymbol{B}}' \,\hat{\boldsymbol{B}}$, we have the solution as:

$$\hat{\boldsymbol{w}} = (\hat{\boldsymbol{B}}' \,\hat{\boldsymbol{B}})^{-1} \,\hat{\boldsymbol{B}}' \,\boldsymbol{a} \quad (3.42)$$

The above solution can also be applied for ridge regression and we obtain the following solution:

$$\hat{\boldsymbol{w}} = (\hat{\boldsymbol{B}}' \,\hat{\boldsymbol{B}} + \sigma \boldsymbol{I})^{-1} \,\hat{\boldsymbol{B}}' \,\boldsymbol{a} \quad (3.43)$$

A dual representation of ridge regression leads us to a classification case where the prediction function is obtained as shown:

$$f(b) = \acute{\boldsymbol{a}} \, (\sigma \boldsymbol{I} + \boldsymbol{G})^{-1} \boldsymbol{k} \qquad (3.44)$$

Where $G = \boldsymbol{B}\acute{\boldsymbol{B}}$ is gram matrix of inner products of training data points, here K can be a similarity matrix kernel between all data points in B and $k_i = \langle \mathbf{b} \cdot \mathbf{b}_i \rangle$.

CHAPTER 4

4.1 Preface

Biochemical databases such as KEGG [12], PubChem [28], SuperTarget [8], Matador [8] are full of information about drug compounds and proteins but information regarding drug target interactions and the amount of validated drug-protein interactions are very few in number when compared with the number of drugs or number proteins in a given database. So, this motivated researchers to employ computational approaches to predict the drug target interactions. The prediction of drug target interactions is a very crucial step in drug design and drug discovery which not only leads us to discover novel drug for new target proteins but also help us to discover many hidden off target interactions for marketed drugs. In silico methods can be thought of as a precursor to lab experiments to validate drug target interactions which are time consuming and expensive. Traditional approaches involve docking simulations which rely on 3D target protein structures which are rarely available hence, machine learning techniques which are more efficient than docking simulations are being employed. To obtain better results we need to employ robust statistical learning methods which are efficient and accurate. Text mining approaches [36] which use information from biomedical documents are good but not efficient enough because of literature redundancy. Many research scholars have proposed successful machine learning methods based on drug-drug similarities and target-target similarities (supervised and semi-supervised) for predicting drug target interactions from heterogeneous data sources such as chemical structure similarities, side effect similarities of drug compounds and genomic sequence similarities and other functional information. The underlying assumption in all these methods is that similar drugs target similar

proteins and vice versa. Various methods proposed include kernel regression (supervised bipartite graph inference) [32, 33], pairwise kernel method [23], Laplacian regularized least squares method [31], Gaussian interaction profile kernels method [27] and kernelized Bayesian matrix factorization method [7] and so on. These models have two steps for predicting drug target interactions first step is training phase, where a model is learned based on the available drug target interactions represented in the form of a bipartite graph and the next step is to predicting new interactions based on the trained model. In the approaches based on kernels, set of data points for example drugs and target proteins are represented in a Hilbert space by a set of points which defines an elementary system where the relative positions of the data points in the Hilbert space refer to the interactions between drugs and proteins. Kernels methods are quite effective to build a model for inferring the interactions between data points. In the case of drug target prediction the there are two classes of data points one represent drug compounds and the other represent target proteins and using similarity kernels a relationship is learned between drugs and target proteins. Prediction of drug target interactions can be done in four scenarios: First, predicting interaction for known drug and known target i.e. drugs (targets) with at least one known interaction(s), second case involves predicting new targets for drugs with known interaction(s). Third scenario is to predict new drugs for target proteins with at least one known interaction(s) and in the final fourth predicting interactions for a new drug and a new target candidate for which we do not have any prior interaction information. In the current chapter we propose a method for predicting drug target interactions based on the work in [27]. We modify the RLS algorithm which uses chemical structure similarity of drugs and genomic sequence similarity of proteins in two

ways: first we introduce more useful pharmacological effect similarity based on side effect keywords [33] of drug compounds and we design more sophisticated kernels for chemical structure similarity, pharmacological effect similarity of drugs and genomic sequence similarity of targets. And show that these simple yet effective changes to the existing method have better results. This method predicts interaction for drug and targets with at least one known interaction in the dataset and thus not possible to detect interaction for novel drugs (targets). Hence, we extend the results obtained from the modified RLS (from here on we refer to the proposed method as KRLS) method to a simple straight forward weighted profile method which can infer interactions for new drugs or proteins with no prior information about their interactions based on their similarity with neighboring drugs or proteins and their interaction profiles.

4.2 Problem framework

First we define a problem framework and use the following notations in this and subsequent chapters. Let $D = \{d_1, d_2, d_3, \ldots, d_{m_d}\}$ be set of drugs in the dataset (let the number of drugs for which interactions are known in the data is m) and $T = \{t_1, t_2, t_3, \ldots, t_{n_t}\}$ (let the number of proteins for which interactions are known is n) be set of targets in the dataset. Let S_c be the chemical similarity matrix, S_p be the pharmacological effect similarity matrix for drugs where the $(i, j)^{th}$ element of a given matrix $S_c(d_i,d_j)$ or $S_p(d_i,d_j)$ denote the similarity score for a pair of drugs d_i, d_j ($i = j = 1$ to m_d) and S_g be the genomic sequence similarity matrix for target proteins where $(i, j)^{th}$ element of $S_g(t_i, t_j)$ denote the genomic sequence similarity score for a pair of proteins t_i and t_j ($i = j = 1$ to n_t). Assume A to be the binary adjacency matrix of known drug target interactions (m x n matrix represents the bipartite graph). The $(i, j)^{th}$ element of A denoted by $A(d_i, t_j) = 1$ if the drug

d_i interacts with target t_j or else $A(d_i, t_j) = 0$ (i = 1 to m_d and j = 1 to n_t) a_i denotes interaction profile vector of drug d_i and a_j^T denotes interaction profile vector of protein t_j. This dataset of known interactions is used as a gold standard data for evaluating the performance of the proposed method in the cross validation stage.

4.3 Proposed Method

In the process of inferring interactions for a protein target t_{new} whose drug interactions are unknown using genomic sequence similarity and interactions of its neighbors, our proposed method can be divided into two steps as depicted in figure-4.1:

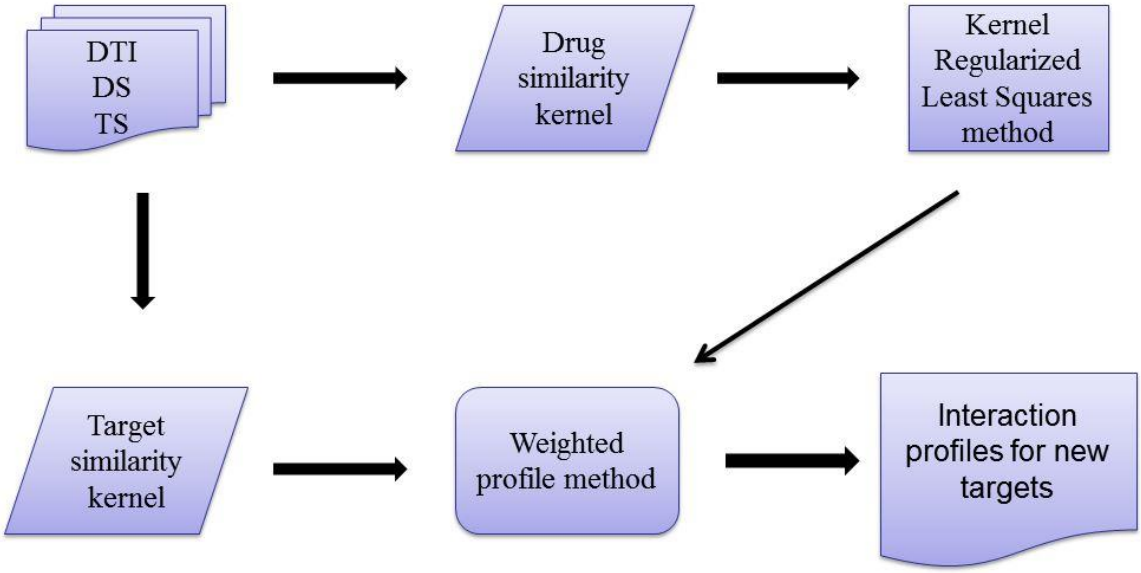


Figure 4.1: Proposed Model for Drug Target Interaction Prediction

Step 1: We infer novel interactions for target proteins present in A^T using Kernel Regularized Least Squares algorithm (KRLS) with drug kernel calculated from a weighted combination of chemical structure similarity kernel denoted by K_c and pharmacological effect similarity kernel denoted by K_p. The chemical structure similarity kernel is computed from chemical structure similarity matrix S_c of drugs and the

pharmacological effect similarity kernel is computed from pharmacological effect similarity matrix S_p. Hence, we get a new drug target interaction matrix $\hat{A}^T$ which contains newly predicted real valued drug interaction scores for protein targets with one or more known interactions in A^T (this step is a small and effective improvement of the method proposed in [27]). This step of the model is depicted in the figure-4.2.

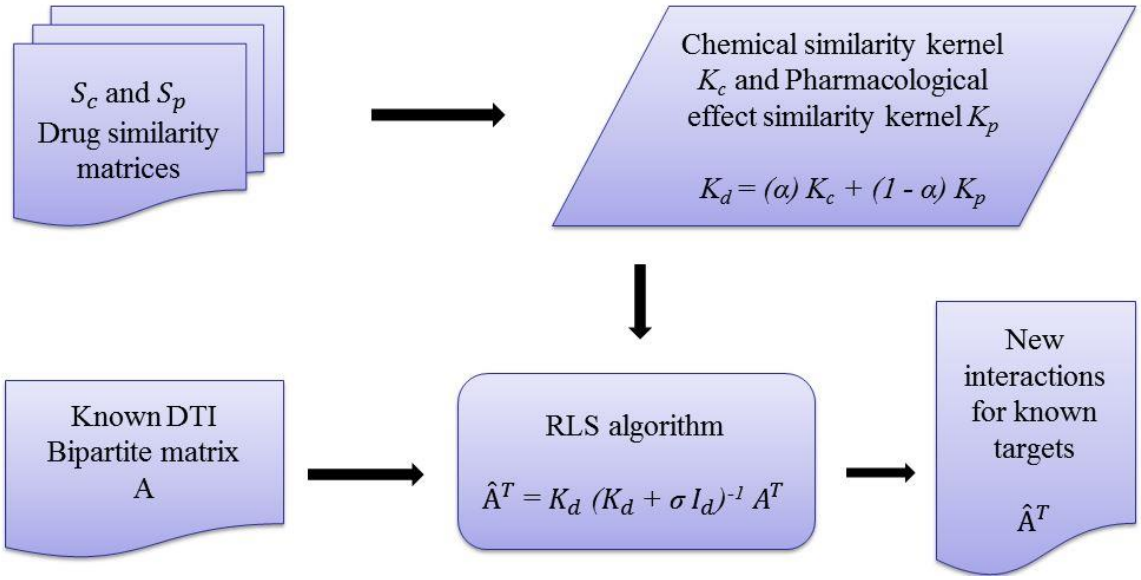


Figure 4.2 Proposed Model: Modifying the RLS Algorithm

ii). In the second step, interaction score vector of new target protein t_{new} for which all of its interactions are unknown are predicted using weighted profile method based on genomic sequence similarity kernel denoted by K_g which is computed from genomic sequence similarity matrix S_g of protein targets and the new real valued interactions score matrix $\hat{A}$ which contains (newly predicted real valued) interactions of neighbors of t_{new}. We get a vector $\hat{a}^T_{t_{new}}$ which contains interaction scores of t_{new} with respect to all the drugs (d_i to d_{m_d}) where high scoring entries indicate the presence of interaction with individual drug compound (this step extends the RLS algorithm proposed in [27] to infer interactions for new protein targets). This step is depicted in the figure 4.3.

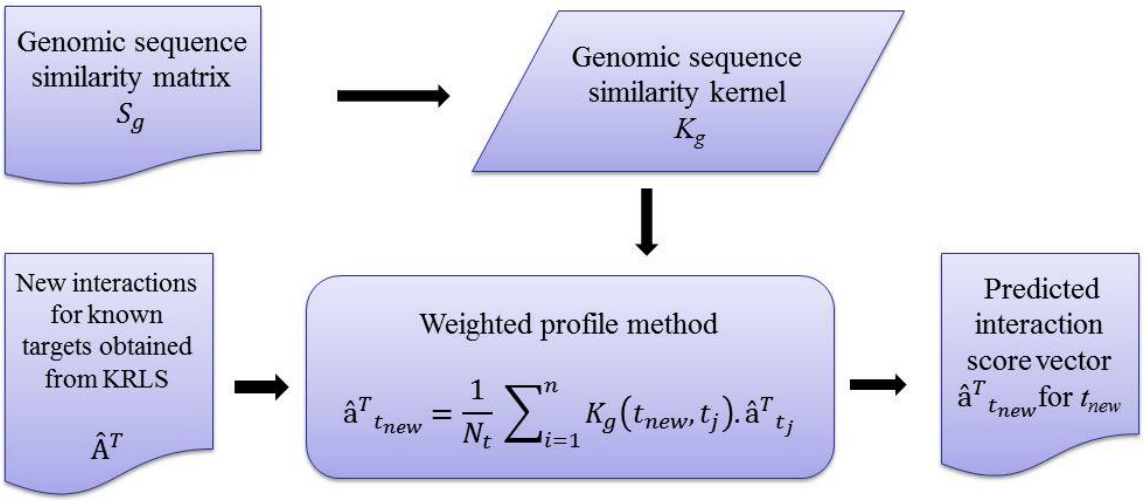


Figure 4.3 Proposed Model: Extending the RLS Algorithm to Weighted Profile Method

From drug point of view, interaction score vector $\hat{a}_{d_{new}}$ for a new drug compound d_{new} is inferred in a similar way. First, novel interactions for drug compounds in matrix A are predicted using KRLS method based on genomic similarity kernel K_g and a new real valued interaction score matrix $\hat{A}$ is obtained which is carried to second step where using weighted profile method based on drug kernel K_d and $\hat{A}$ obtained in the initial step, interaction score vector of new drug d_{new} is computed where high scoring entries in the vector indicate a presence of interaction with corresponding protein target.

Considering the interaction prediction problem for a given drug target pair (d_{new}, t_{new}) we compute the interaction score by taking a weighted average of each individual scores obtained one from $\hat{a}^T{}_{t_{new}}$ and the other from $\hat{a}_{d_{new}}$. Therefore, we now can predict the high scoring values to be presence of interaction between d_{new} and t_{new}. The entire idea discussed above is as follows:

To predict the interaction score vector $\hat{a}^T{}_{t_{new}}$ for a new target protein t_{new}, first we compute the drug kernel K_d as represented in the figure-4.4:

$K_d = (\alpha)\ K_c + (1 - \alpha)\ K_p$, where α is a parameter (4.1)

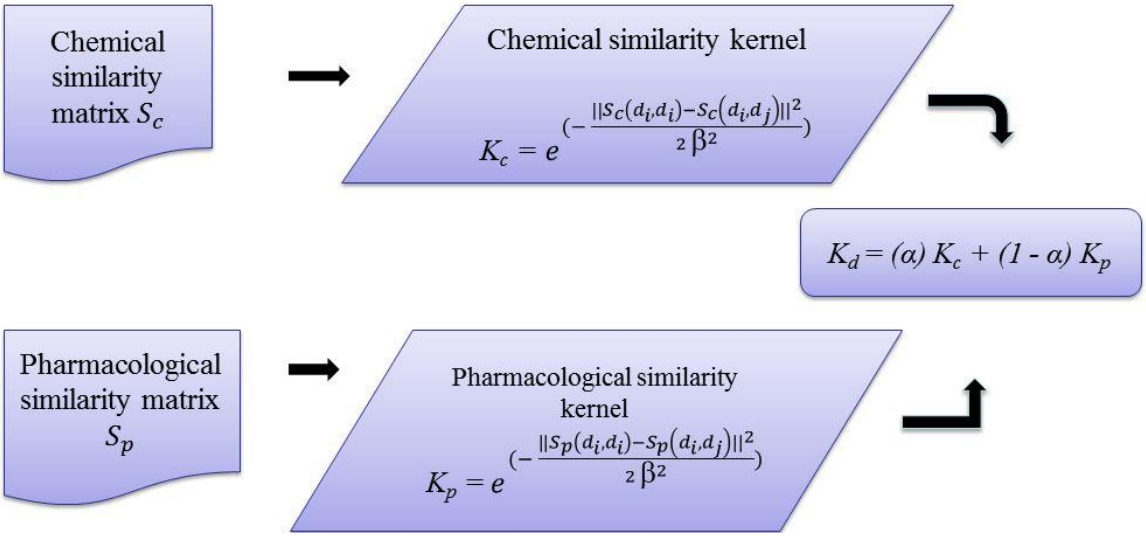


Figure 4.4 Proposed Model: Computing Weighted Drug Similarity Kernel

K_c and K_p are the drug similarity kernels obtained from chemical structure similarity and pharmacological effect similarity of drug compounds respectively which are computed using the radial basis function as shown:

$$K_c = e^{(-\frac{||S_c(d_i,d_i)-S_c(d_i,d_j)||^2}{2\beta^2})} \quad (4.2)$$

Where S_c is the chemical structure similarity matrix of drug compounds and $S_c(d_i, d_j)$ is the chemical structure similarity score of drug compounds $d_i\ and\ d_j$ respectively. (i = j = 1 to m_d, m_d is the total number of drugs in the dataset).

$$K_p = e^{(-\frac{||S_p(d_i,d_i)-S_p(d_i,d_j)||^2}{2\beta^2})} \quad (4.3)$$

Where S_p is the pharmacological effect similarity matrix of drug compounds and $S_p(d_i, d_j)$ is the pharmacological effect similarity score of drug compounds $d_i\ and\ d_j$ respectively. (i = j = 1 to m_d, m_d is the total number of drugs in the dataset) and $\beta = 1$.

Now the novel interactions are predicted for target proteins in *A* using KRLS method and we have a closed form solution as shown:

$$\hat{A}^T = K_d \, (K_d + \sigma \, I_d)^{-1} \, A^T \qquad (4.4)$$

Where I_d is m_d x m_d identity matrix and we have used $\sigma = 1$.

$\hat{A}^T$ is real valued matrix in which $\hat{a}^T{}_{t_j}$ represent the real valued interaction score vector of target protein t_j with each drug d_i (i = 1 to m_d) and is used in weighted profile method to infer the interaction score vector for a new target t_{new} based on genomic sequence similarity kernel K_g.

K_g the protein similarity kernel is computed from S_g the genomic sequence similarity matrix using Gaussian kernel function as shown in the figure-4.5:

$$K_g = e^{\left(-\frac{||S_g(t_i,t_i) - S_g(t_i,t_j)||^2}{2\beta^2}\right)} \qquad (4.5)$$

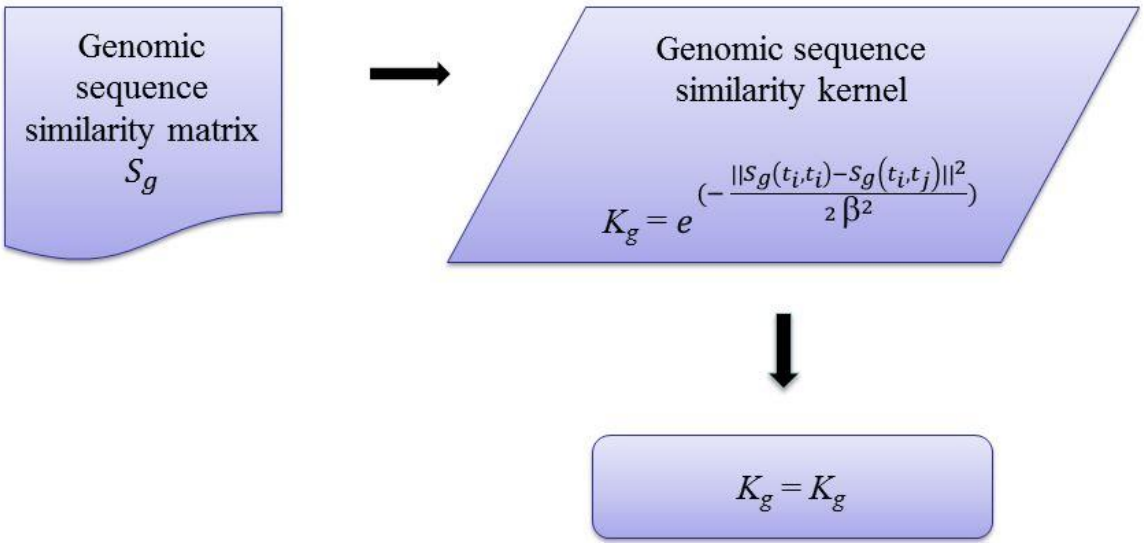


Figure 4.5 Proposed Model: Computing Target Similarity Kernel

Where S_g is the genomic sequence similarity matrix of targets and $S_g(t_i, t_j)$ is the genomic sequence similarity score of protein targets t_i *and* t_j respectively. (i = j = 1 to n_t, where n_t is the total number of proteins in the dataset) and β =1.

In the next the interaction score vector $\hat{a}^T{}_{t_{new}}$ for t_{new} is inferred as shown:

$$\hat{a}^T{}_{t_{new}} = \frac{1}{N_t} \sum_{i=1}^{n} K_g(t_{new}, t_j). \hat{a}^T{}_{t_j} \quad (4.6)$$

Where N_t is a normalizing factor calculated as:

$$N_t = \sum_{j=1}^{n} K_g(t_{new}, t_j) \quad (4.7)$$

And finally high scoring entries in the 1 x m_d, $\hat{a}^T{}_{t_{new}}$ are predicted to be interacting with target protein t_{new}.

Similarly the interaction score vector for drug d_{new} is also inferred; first we predict novel interactions for drugs in binary matrix A using KRLS method based on genomic kernel K_g we have a closed form solution as shown:

$$\hat{A} = K_g (K_g + \sigma I_t) A \quad (4.8)$$

Where I_t is n_t x n_t Identity matrix and we have used σ = 1.

$\hat{A}$ is real valued matrix in which $\hat{a}_{d_i}$ represent the real valued interaction score vector of drug compound d_i with each target t_j (i = 1 to m_d and i = 1 to n_t) and is used in weighted profile method to infer the interaction score vector for a new drug compound d_{new} based on drug similarity kernel K_d.

The interaction score vector $\hat{a}_{d_{new}}$ for d_{new} is inferred as shown:

$$\hat{a}_{d_{new}} = \frac{1}{N_d} \sum_{i=1}^{m} K_d(d_{new}, d_i).\hat{a}_{d_i} \quad (4.9)$$

Where N_t is a normalizing factor calculated as:

$$N_d = \sum_{i=1}^{m} K_d(d_{new}, d_i) \quad (4.10)$$

And finally high scoring entries in the 1 x n_t, $\hat{a}_{d_{new}}$ are predicted to be interacting with drug compound d_{new}.

To infer an interaction score for drug target pair (d_{new}, t_{new}) we take the weighted average of $\hat{A}^T(d_{new}, t_{new})$ and $\hat{A}(d_{new}, t_{new})$ as shown:

$$Score\ (d_{new}, t_{new}) = \frac{\hat{A}^T(d_{new},t_{new}) + \hat{A}(d_{new},t_{new})}{2} \quad (4.11)$$

High scored values are predicted to be interacting pairs.

Hence, by extending the result of KRLS method to weighted profile method we are able to predict the interaction for protein targets (drugs) for which we do not have any interaction information which was not possible by the method proposed in [27].

4.4 Similarity Measures

This section provides a summary of how the authors in the paper [33] computed drug-drug and target-target similarity measures from drug data and target protein data extracted from chemical and biological databases.

4.4.1 Computing Chemical structure similarity measure for drug compounds

The chemical structure similarity of drug compounds is computed using a program called SIMCOMP [10] which is based on the graph alignment algorithm finds common sub structures between a given pair of drugs and gives out a global similarity score. The similarity between two compounds structures x_{d_1} and x_{d_2} is calculated using Tanimoto coefficient as follows:

$$S_{chem}(x_{d_1}, x_{d_2}) = |x_{d_1} \cap x_{d_2}| / |x_{d_1} \cup x_{d_2}| \quad (4.12)$$

The similarity score obtained is referred to as 'chemical structure similarity' and applying this to all compounds in a dataset, a similarity matrix S_c is obtained which represents chemical space. The chemical structures of drug compounds are retrieved from KEGG DRUG and KEGG LIGAND databases [12].

4.4.2 Computing Pharmacological effect similarity measure for drug compounds

Approved by Health and Welfare Minister of Japan, Japan Pharmaceutical Information Center (JAPIC) maintains all information of pharmaceutical products in JAPAN and from drug package inserts pharmacological effect keywords are obtained which were in Japanese language and they were analyzed with the help of MeCab (http://lsd.pharm.kyoto-u.ac.jp/en/index.html) to nouns and phrases.

The obtained information is translated to English and with the help of life science dictionary synonyms were unified. XML format was used to describe the JAPIC entries and using various tags different set of profiles for drugs were generated. Similar tags were used to form groups and they obtained five tag groups with a specific number of keywords for each group, the table-4.1 summarizes the XML tag names, number of corresponding keywords and a small description provided by the authors in [33].

XML tag	Number of keywords	Description
Caution	16849	Adverse events such as caution for application, overdose and warning
Interaction	14223	Combined used of drugs
Patient	16362	Types of patients based on gender, age or disease
Pharmaceutical effect*	17109	Efficacy, usage and pharmacology
Property	17142	Properties such as melting point, partition coefficient, pharmacokinetics and solubility

Table 4.1 Summary of XML tag names and side effect keywords

Now to calculate the pharmacological similarity score for the drug compounds, first every drug in the data set is represented using a binary profile vector $y = (y_1, y_2, \ldots, y_k)^T$ where a k is the number of keywords and pharmacological keyword is coded 0 or 1, across 17109 keywords. Using weighted cosine correlation coefficient, similarity score between two given drug compounds y and y' is obtained as follows:

$$S_{phar}(y, y') = \frac{\sum_{k=1}^{K} w_k y_k y'_k}{\sqrt{\sum_{k=1}^{K} w_k {y_k}^2} \sqrt{\sum_{k=1}^{K} w_k {y'_k}^2}} \quad (4.13)$$

Where w_k is the weight function for the k^{th} keyword defined as $w_k = e^{(-{d_k}^2/\gamma^2 h^2)}$, k = 1, 2, …, K, d_k is the frequency of the k^{th} keyword in the data and K is the total number of keywords in the data and γ is the standard deviation of $\{ d_k \}_{k=1}^{K}$ and h is a parameter.

Applying this operation for all drug-drug pairs in the data we obtain a similarity matrix S_p and it represents '*pharmacological space'*. The weight function is used to give more importance to less frequent words found in the package inserts.

4.4.3 Computing genomic sequence similarity measure for proteins

Using normalized version of Smith-Waterman scores, genomic sequence similarity between two given proteins z and z' is calculated as follows:

$$S_{genomic}\,(z, z') = SW(z, z') / \sqrt{sw(z, z)}\,\sqrt{sw(z',\ z')} \qquad (4.14)$$

Where $SW(\cdot\ ,\ \cdot)$ is the original Smith-Waterman score, and amino acid sequences were retrieved from KEGG GENES [12] database. Applying this to all the target proteins in the dataset a similarity matrix S_g is obtained and is used to denote '*genomic space*'.

4.5 Performance evaluation methods

4.5.1 Cross validation technique

Cross validation is a technique used to measure the performance of learning algorithms, the idea behind this technique is to calculate the predictive performance of the model learned from the given data, can also be used to compare two or more different models in order to see which model best suits the given data. It is done by dividing data into two parts:

i). Training set: used for learn/training the model

ii). Test set: used for validating the model

Generally the training set and test set are chosen in a way that they successively cross over such that every data instance gets an equal chance of being validated. Typical form used across the machine learning community is *k*-fold cross-validation. Mostly other forms of this technique are special case of *k*-fold cross validation which includes leave-one-out cross-validation and hold-out validation etc.

4.5.1.1 *k* - Fold Cross-Validation

In this technique, the data is randomly split into *k* segments of roughly equal size, one of the partitioned subsample is used as test data for validating the model and the remaining *k* – 1 segments are used for training the model for prediction. This process is repeated *k* times where each of the *k* subsamples is being used as the test data while the rest used as training set. By doing this we get *k* results one from each fold of *k*-fold cross validation and by taking the average of these *k* results we obtain a single measure of the performance of the trained model for the given data set. During the whole process all the instances are used for validation and training where each observation is used as test set exactly once. The value of *k* is set by the user and can take the values *k* = 1,2,3,4.., N.

We have used 5-fold cross-validation in this work. In this technique first we divide the given dataset into five equal subsamples, any one of the five subsamples is used as test data and rest four are used for training the classifier and the prediction of drug target interactions is done for the test set. After obtaining the new interactions we calculate the performance of the trained model. We end up with five scores of Area under the Receiver Operating Characteristic curve for each fold of 5-fold cross-validation and then we

calculate the average of these five scores to obtain the final AUC score. The figure-5.1 illustrates the 5-fold cross validation procedure.

4.5.2 ROC Analysis

Performance of a binary classifier can be measured using Receiver operating characteristic by varying the threshold. The ROC curve is plotted with the fraction of false positives out of total actual negatives (false positive rate) on *x*-axis and the fraction of true positives out of total actual positives (true positive rate) on *y*-axis obtained at different threshold values. True positive rate is also known as sensitivity and False positive rate is obtained by 1 - specificity. ROC analysis enables us in choosing optimal models or parameters for classification. One point in the ROC space corresponds to an individual result of the classifier. The figure-4.6 depicts a ROC curve. The dotted line indicates an instance where the classifier performance is average and the curve closer to left top corner indicates a better performance of the a given classifier.

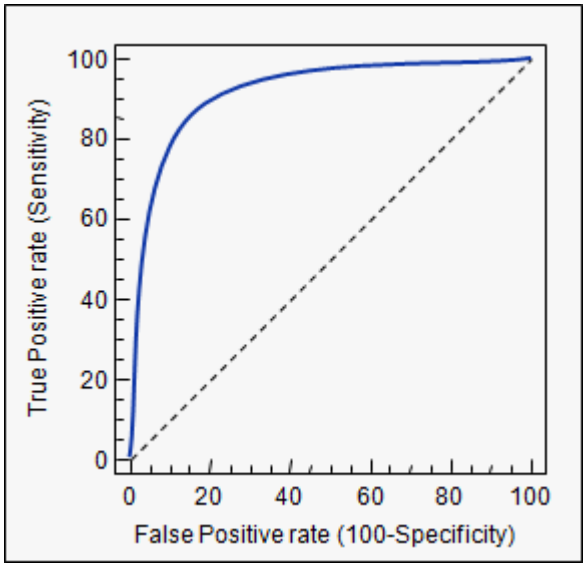


Figure 4.6 Receiver Operating Characteristic Curve: an example

In general there are two types of classifiers, one which predicts the class label directly and the other type produces continuous output to which by applying different thresholds we predict the class label. Assuming the value 1 to be an interaction and 0 to be a non-interaction, the outcome of the classifier has four possible cases, if the original value is 1 and the predicted value is also 1 we count it as a true positive, but if the predicted value is 0 then, we count that as false negative and if the original value is 0 and the predicted value is 1, we count it as false positive, but if the predicted value is also 0 then, we count it as true negative. So, we can construct a two-by-two confusion matrix which is shown in the table-4.2:

		Classified as	
		1	0
Really is	1	TP	FN
	0	FP	TN
Metrics		**Positive predicted value (PPV) Precision** TP/(TP+FP)	**Negative predictive value(NPV)** TN/(TN+FN)

Table 4.2 Confusion Matrix Table

Where TP, FP, TN and FN refer to true positives, false positives, true negatives and false negatives respectively and we can calculate different metrics using the confusion matrix above as follows:

Total positive results P

$$P = TP + FN \tag{5.4}$$

Total negative results N

$$N = TN + FP \tag{5.5}$$

True positive rate (TPR), sensitivity

$$TPR = TP / P = TP / (TP + FN) \tag{5.6}$$

False positive rate (FPR)

$$FPR = FP / N = FP / (TN + FP) \tag{5.7}$$

Specificity (SPC) or True Negative Rate

$$SPC = TN / N = TN / (FP + TN) = 1 - FPR \tag{5.8}$$

ROC graphs can be obtained by plotting false positive rate on *x*-axis and true positive rate on *y*-axis which depicts relative trade-offs between false positives and true positives. For a given classifier, by varying thresholds we obtain different (fp rate, tp rate) pairs which correspond to several points in the ROC space.

4.5.2.1 Area under ROC curve

To compare two or more different classifiers or model we calculate area under the ROC curve, which can be interpreted as "the probability that the classifier will rank a randomly chosen positive instance than a randomly chosen negative instance" [6]. This is similar to the Wilcoxon test of ranks. Generally while comparing two classifiers using area under the ROC curve, higher values of AUC score corresponds to better the performance of the

classifier. The figure-4.7 shows the ROC curve obtained for classifier B by varying threshold values and AUC for classifier A for a single threshold value.

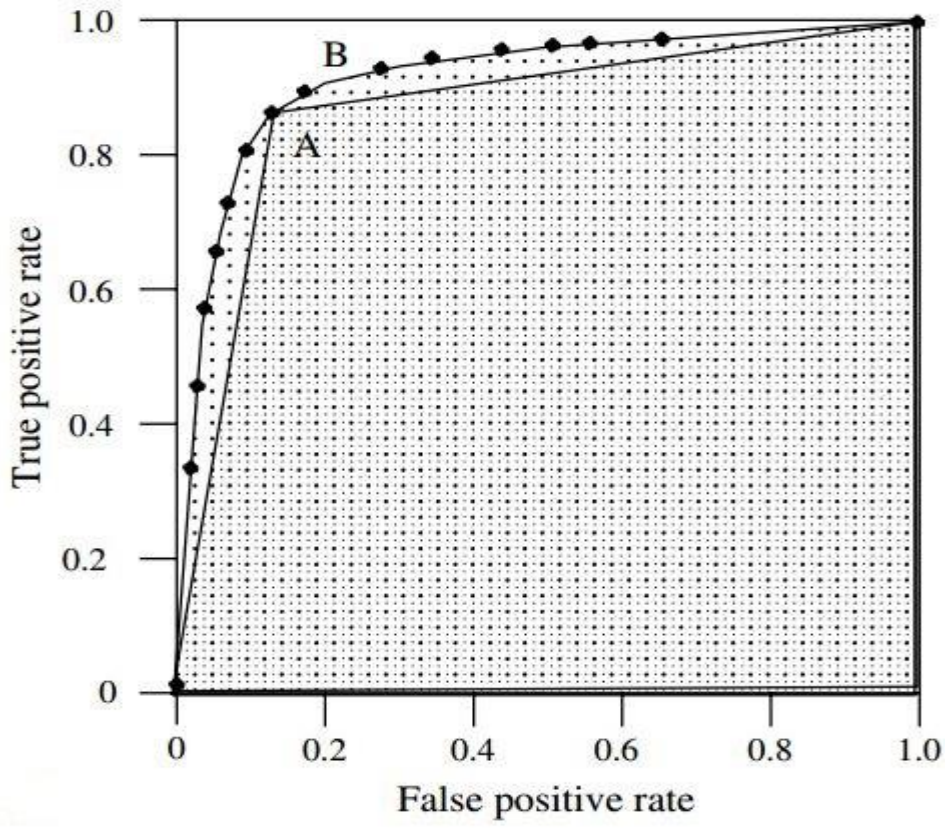


Figure 4.7 Area under ROC curve: an example [6]

CHAPTER 5

5.1 Preface

In this section we discuss the experiments performed and the performance evaluation of the proposed method in terms of Area under Receiver Operating Characteristic curve. Initially we explain the data sets used in this work like how they are represented to fit the proposed model. Then, we show that using the proposed method we have obtained higher AUC (area under roc curve) when compared with state of art methods and our results indicate that the proposed method can predict true drug target interactions accurately.

5.2 Data set

We have used the same dataset provided in [33]

http://cbio.ensmp.fr/˜yyamanishi/pharmaco/

The data set contains four different protein classes of drug-target interactions networks in human beings which involve enzymes, ion-channels, G-protein-coupled receptors and nuclear receptors. Table-5.1 presents some statistical information regarding the four different classes.

Class	Number of drugs	Number of targets	Number of drug-target interactions
Enzyme	212	664	1515
Ion-channel	99	204	776
GCPR	105	95	314
Nuclear receptor	27	26	44

Table 5.1Statistical Information on Dataset

5.2.1 Drug target interaction data

The drug target interactions are obtained from the following databases KEGG BRITE [12], BRENDA [22], SuperTarget [8] and DrugBank [29].

The known drug target interaction data is considered as gold standard data and it is being used to train the classifier to infer unknown drug target interactions and also during cross validation in order to evaluate the performance our proposed method.

5.3 File format

For each protein family there are four different files we have used and the table-5.2 gives an idea about the files and their format for a dataset.

File (all are text files)	Format	Description
Adjacency matrix of gold standard drug-target interaction data	$\begin{bmatrix} d_1t_1 & \cdots & d_1t_n \\ \vdots & \ddots & \vdots \\ d_mt_1 & \cdots & d_mt_n \end{bmatrix}$	A binary matrix where an entry $d_it_j = 1$ indicates presence of interaction between the drug-target pair. m is number of drugs and n is number of targets in the dataset.
Chemical structure similarity matrix	$\begin{bmatrix} d_1d_1 & \cdots & d_1d_m \\ \vdots & \ddots & \vdots \\ d_md_1 & \cdots & d_md_m \end{bmatrix}$	A real valued matrix where $0 < d_id_j < 1$ the higher the value the more similar the given drug-drug pair and the diagonal is 1.
Pharmacological similarity matrix	$\begin{bmatrix} d_1d_1 & \cdots & d_1d_m \\ \vdots & \ddots & \vdots \\ d_md_1 & \cdots & d_md_m \end{bmatrix}$	A real valued matrix where $0 < d_id_j < 1$ the higher the value the more similar the given drug-drug pair and the diagonal is 1.

Genomic sequence similarity matrix	$\begin{bmatrix} t_1t_1 & \cdots & t_1t_n \\ \vdots & \ddots & \vdots \\ t_nt_1 & \cdots & t_nt_n \end{bmatrix}$	A real valued matrix where $0 < t_it_j < 1$ the higher the value the more similar the given target-target pair and the diagonal is 1.

Table 5.2 Explains Format of Data files

5.4 Evaluation

We discuss the evaluation technique used to assess the performance of our proposed method in the section 4.5 of the thesis. The adjacency matrix of drug target interactions (bipartite graph) provided in the paper [33] is used as gold standard data and is used for evaluation. We employ 5-fold cross validation and calculate AUC score (area under receiver operating curve) for each protein family of drug target interactions. In this section we compare the result of the proposed model to state-of-art methods based on AUC measure.

5.4.1 Performance evaluation for the proposed method

We employ the same procedure followed by the authors in the work [33]. We explain in the context of predicting interactions for new proteins and a similar procedure is followed for drug compounds. So, initially for each dataset we split the target proteins into five subsets of equal size. Then, each subset is used in turn as the test set and classifier is trained on the remaining four subsets. This process is repeated five times. For a protein target in test set, we assume all of its interactions are unknown and are labeled as 0. Performance of the classifier is assessed using area under the ROC curve. The ROC curve is plotted using (tpr, fpr) pair obtained for different thresholds. Here upper one percentile

in the prediction score is chosen as threshold as prediction scores are interpreted as confidence. Since, in our work we did not validate new interactions, what we do is assume the target proteins in the test set (present in gold standard data) to be new proteins i.e. their interactions with all the drugs in the dataset to be unknown and thus we train the classifier on the dataset with interaction profiles of remaining proteins in the training set (gold standard data) and predict the interactions of the proteins in the test set. The results in the table-5.3 indicate that our method has slightly better AUC (average) on all four protein families.

	Proposed method KRLS	KR-CG	KR-TP	KR-PP	RLS-GIP-AVG-avg	RLS-GIP-KRON-avg
Enzyme	**95.7**	82.1	89.2	84.5	93.7	93.4
Ion-Channel	**96.9**	69.2	81.2	73.1	94.7	94.9
GCPR	**94.6**	81.1	82.7	81.2	91.3	91.7
Nuclear Receptor	**93.0**	81.4	83.5	83.0	90.8	90.2

Table 5.3 Comparison Table of AUC Scores

Here, we compare with previous state of art methods [27, 33] employed in the prediction of drug target interactions. KR-CG refers to the Kernel regression method proposed in [32] which is a supervised bipartite graph inference based on chemical structure similarity of drugs, KR-TP and KR-PP refers to kernel regression methods proposed in [33] where TP indicates they have used true pharmacological similarity and PP indicates they have used predicted pharmacological similarity by assuming the drugs in test do not have any pharmacological similarity data and it is inferred using chemical similarity by a similarity regression method. RLS-GIP method is proposed by the authors in [27] where RLS-AVG

refers to the method in which predictions are made for drugs (using genomic kernel) and target proteins (using drug kernel) separately and are combined using an average function, RLS-KRON refers to the method where the authors have used Kronecker product kernel for drugs and proteins to infer drug target interactions. It should be noted that the result obtained from this method corresponds to the scenario where we have at least one known interaction for the drugs (proteins) in test set.

The results in the above table indicate that in our method by improvising the RLS method with kernels defined over drug-drug and target-target similarities and extending it to weighted profile method is effective in predicting true drug target interactions.

5.4.2 Relevance of Kernels on drug target interaction prediction

We have also tested the model by varying the value of the parameterα_d $[0, 0.25, 0.5, 0.75, 1]$, by varying the values between 0 (drug kernel) $K_d = K_p$ (pharmacological similarity kernel) and 1 (drug kernel) $K_d = K_c$ (chemical similarity kernel). It should be noted that altering the parameter $\alpha_d$changes the *KRLS* method while inferring interactions for new proteins and whereas altering the parameter α_d changes the weighted profile method when inferring interactions for new drug compounds. As the value of α_d increases, the chemical similarity kernel has more impact on the predicted interactions and as the value of α_d decreases predicted interactions are influenced pharmacological similarities of drug compounds.

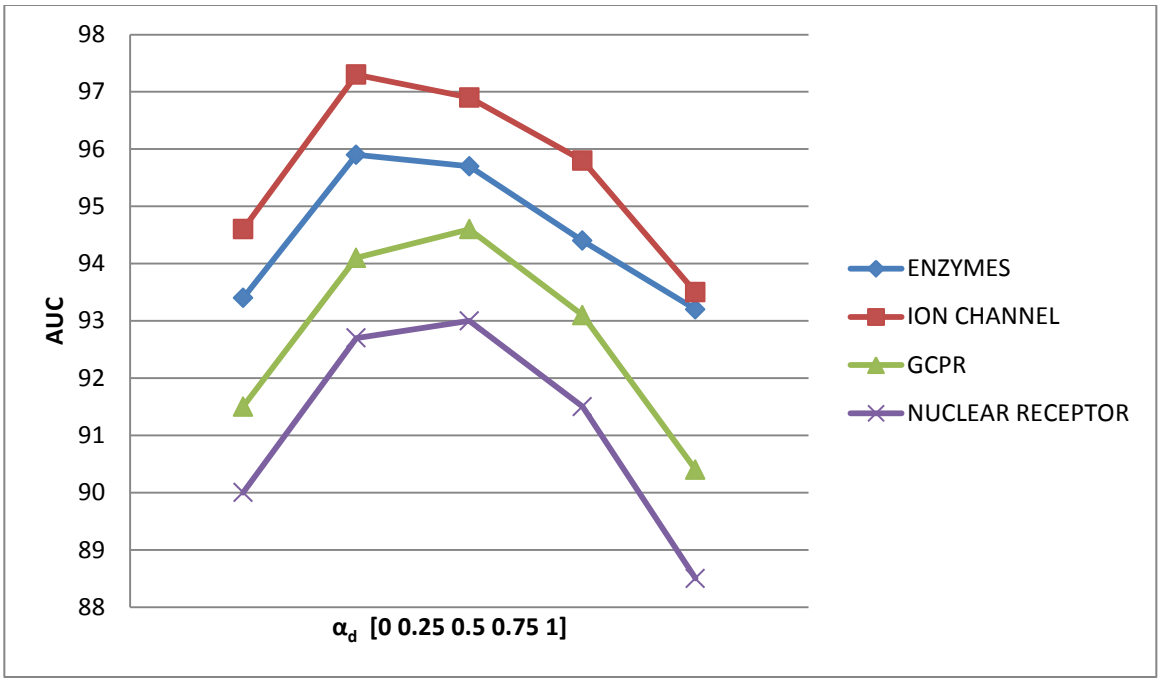


Figure 5.1 Relevance of drug similarity kernels (AUC for different values of α_d)

In the figure-5.1, as we move from left to right the value of α_d $[0, 0.25, 0.5, 0.75, 1]$ increases indicating that the effect of pharmacological similarity of drugs on the drug target interaction inference is decreasing while the effect of chemical similarity of the drug compounds is increasing. It can be observed that the drug target interactions are more related to the pharmacological similarity of drug compounds than chemical similarity of drug compounds which is also observed by the authors in [33]. Even though drug target interactions are correlated to pharmacological similarity, we have achieved a good result where combined average of both the drug similarity kernels is used. In the case of bigger datasets such as Enzymes, it can be seen that the chemical similarity kernel is a bit uninformative in detecting the true drug target interactions.

Finally, as mentioned by the authors in [31] that there is a wealth of information available in the drug target interaction network, we have tested the result of including this

information in the to see its effect on the performance of method proposed by calculating n_t x n_t interaction similarity kernel represented by K_I for all the proteins in the bipartite graph A, where the entry K_I (t_i, t_j) is the number of common drugs shared by the proteins t_i and t_j. We obtain K_I in a similar way employed by the authors in [31] but since the bipartite graph is different from the one used in [31], we had to re-compute the similarity matrix. And now this similarity kernel matrix K_I is integrated with the genomic sequence similarity kernel K_g.

The chart below shows the results obtained after integrating the genomic kernel with new target protein kernel (drug target network information) and it can be observed that using the information from the drug target interaction network does prove to be effective to some extent in inferring the unknown drug target interactions.

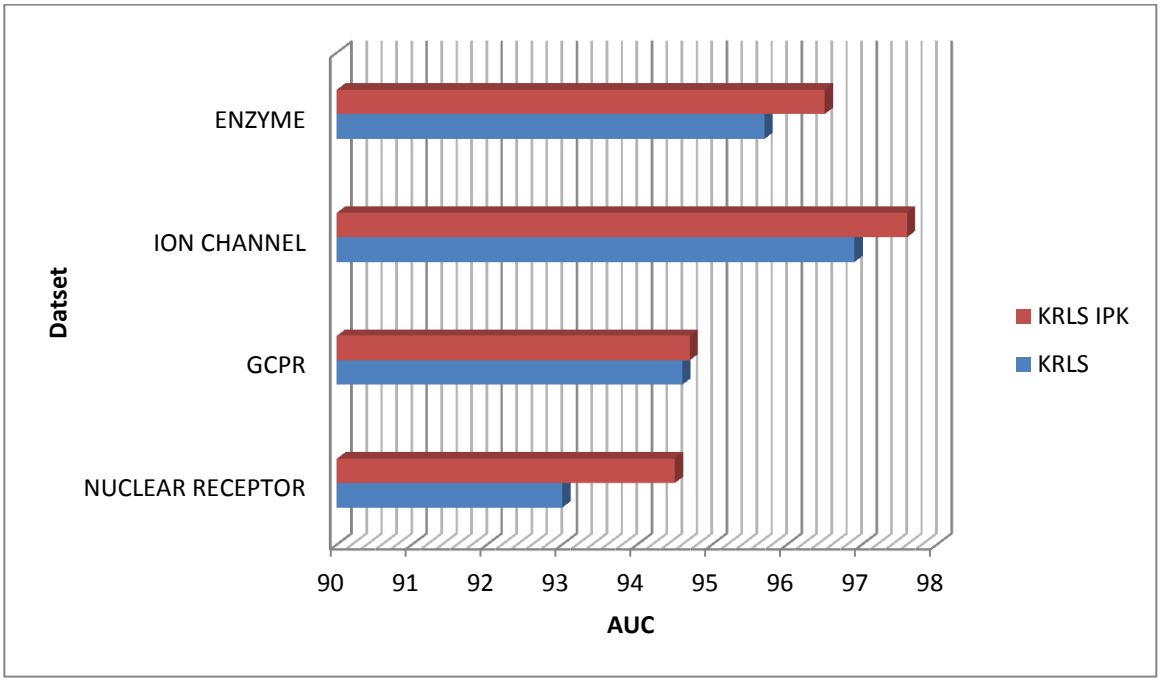


Figure 5.2 AUC scores after integrating DTN information into target protein kernel

As we can see data sets with higher rate of known interactions are benefited more with the inclusion of information on drug target interaction network topology.

CHAPTER 6

6.1 Conclusion

In this work, we have proposed an algorithm that modifies and extends a previously implemented method [27] to predict new drug target interactions by integrating the weaker drug kernel with more informative pharmacological effect similarity used in [33] and also as the method in [27] cannot be applied to infer interactions for new drugs(proteins) i.e. drugs or protein targets in test set with no known interaction, we extend the method to more simple yet effective weighted profile method thus enabling us to predict interactions for new drugs (proteins). The novelty of the proposed method can be described in two ways:

i). We improvise the RLS method employed in [27] by introducing kernels defined on drug-drug and target-target similarities using radial basis function, then the drug kernel is altered by integrating a similarity measure for drugs which is more correlated to drug-target interactions (pharmacological effect similarity of drugs with chemical similarity) to obtain more effective interaction predictions.

ii). As the RLS method proposed in [26] cannot predict interactions for proteins or drugs with unknown interaction profiles we extend the result of KRLS method to weighted profile method which infers interaction for new proteins (drugs) using the target similarity kernel (drug similarity kernel) based on the interaction profiles of its neighbors. The real valued prediction scores obtained from KRLS method has better information on interaction profiles of neighboring drugs (proteins) corresponding to new drugs (protein

targets) and achieves a better prediction result than weighted profile method which uses a binary matrix of drug target interactions[32].

To evaluate the performance of the prediction, we have calculated area under the ROC curve measure, which measures sensitivity as the function of 1-specificity. Sensitivity (TPR) defines the number of correctly predicted positive interactions among all positive interactions available while testing the model and 1 - specificity (FPR) defines the number of incorrectly predicted positive interactions among all unknown interactions (negative samples) available during the test. Obtained results with high AUC scores indicate that our method can predict true drug target interactions and also integrating the drug kernel with pharmacological effect similarity further improved the performance of prediction.

6.2 Future Work

A simple linear combination of drug based prediction scores and target based predictions scores is being employed in the current model and we can improve our model to a more sophisticated method of combining these individual predictions. And furthermore we believe that integrating more similarity measures of drugs such as drug-drug interaction closeness, gene expression based similarity and for proteins such as protein-protein interaction closeness, gene ontology based similarity can prove to be effective in detecting many unknown and useful off targets for drugs. We used a weighted combination of drug-drug similarity kernels in this work and techniques based on feature selection can be next step of improvement to make better informed predictions. The performance of our method has been evaluated based on the gold standard data provided in [33] and the practical applicability of our method can be investigated further for

example, in identifying targets of drug compounds for drug design and experimental validations. On a concluding note we can say that any work of research can always be improved upon and we are working towards it.

REFERENCES

[1] K. Bleakley and Y. Yamanishi, “Supervised prediction of drug-target interactions using bipartite local models.,” *Bioinformatics*, vol. 25, no. 18, pp. 2397–403, Sep. 2009.

[2] 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. 263–6, Jul. 2008.

[3] D. Chen, “Potential drug targets prediction for H1N1 influenza A based on protein-protein interaction networks,” *African J. Pharm. Pharmacol.*, vol. 6, no. 42, pp. 2950–2955, Nov. 2012.

[4] M. C. Cobanoglu, C. Liu, F. Hu, Z. N. Oltvai, and I. Bahar, “Predicting drug-target interactions using probabilistic matrix factorization.,” *J. Chem. Inf. Model.*, vol. 53, no. 12, pp. 3399–409, Dec. 2013.

[5] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, “Similarity-based machine learning methods for predicting drug-target interactions: a brief review.,” *Brief. Bioinform.*, Aug. 2013.

[6] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[7] M. Gönen, “Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization.,” *Bioinformatics*, vol. 28, no. 18, pp. 2304–10, Sep.2012.

[8] S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, and R. Preissner, "SuperTarget and Matador: resources for exploring drug-target relationships.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D919–22, Jan. 2008.

[9] J. Hainmueller and C. Hazlett, "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach," *Polit. Anal.*, vol. 22, no. 2, pp. 143–168, Oct. 2013.

[10] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways.," *J. Am. Chem. Soc.*, vol. 125, no. 39, pp. 11853–65, Oct. 2003.

[11] L. Jacob and J.-P. Vert, "Protein-ligand interaction prediction: an improved chemogenomics approach.," *Bioinformatics*, vol. 24, no. 19, pp. 2149–56, Oct. 2008.

[12] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "KEGG for linking genomes to life and the environment.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D480–4, Jan. 2008.

[13] M. Kuhn, M. Campillos, P. González, L. J. Jensen, and P. Bork, "Large-scale prediction of drug-target relationships.," *FEBS Lett.*, vol. 582, no. 8, pp. 1283–90, Apr.2008.

[14] M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. J. Jensen, A. Beyer, and P. Bork, "STITCH 2: an interaction network database for small molecules and proteins.," *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D552–6, Jan. 2010.

[15] A. Masoudi-Nejad, Z. Mousavian, and J. H. Bozorgmehr, "Drug-target and disease networks: polypharmacology in the post-genomic era," *Silico Pharmacol.*, vol. 1, no. 1, p. 17, 2013.

[16] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Globalized bipartite local model for drug-target interaction prediction," *Proc. 11th Int. Work. Data Min. Bioinforma. - BIOKDD '12*, pp. 8–14, 2012.

[17] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Drug-target interaction prediction by learning from local information and neighbors.," *Bioinformatics*, vol. 29, no. 2, pp. 238–45, Jan. 2013.

[18] E. Pauwels, V. Stoven, and Y. Yamanishi, "Predicting drug side-effect profiles: a chemical fragment-based approach.," *BMC Bioinformatics*, vol. 12, no. 1, p. 169, Jan. 2011.

[19] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, "Combining drug and gene similarity measures for drug-target elucidation.," *J. Comput. Biol.*, vol. 18, no. 2, pp. 133–45, Feb. 2011.

[20] M. Planck, T. Hofmann, B. Sch, and A. J. Smola, “A Review of Kernel Methods in Machine Learning A Review of Kernel Methods in Machine Learning,” no. 156, 2006.

[21] C. Richard and J.-C. M. Bermudez, “Closed-form conditions for convergence of the Gaussian kernel-least-mean-square algorithm,” *2012 Conf. Rec. Forty Sixth Asilomar Conf. Signals, Syst. Comput.*, no. 1, pp. 1797–1801, Nov. 2012.

[22] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, “BRENDA, the enzyme database: updates and major new developments.,” *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D431–3, Jan. 2004.

[23] A. Schuffenhauer, P. Floersheim, P. Acklin, and E. Jacoby, “Similarity metrics for ligands reflecting the similarity of the target proteins.,” *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 391–405, 2003.

[24] Z. Spiro, I. A. Kovacs, and P. Csermely, “Drug-therapy networks and the predictions of novel drug targets,” vol. 20, pp. 1–8, 2008.

[25] Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, and Y. Yamanishi, “Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers.,” *Bioinformatics*, vol. 28, no. 18, pp. i487–i494, Sep. 2012.

[26] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, and Y. Yamanishi, "Drug target prediction using adverse event report systems: a pharmacogenomic approach.," *Bioinformatics*, vol. 28, no. 18, pp. i611–i618, Sep. 2012.

[27] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction.," *Bioinformatics*, vol. 27, no. 21, pp. 3036–43, Nov. 2011.

[28] D. L. Wheeler, T. Barrett, D. a Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. a Tatusova, L. Wagner, and E. Yaschenko, "Database resources of the National Center for Biotechnology Information.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D173–80, Jan. 2006.

[29] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D901–6, Jan. 2008.

[30] J. Wixon and D. Kell, "The Kyoto encyclopedia of genes and genomes--KEGG.," *Yeast*, vol. 17, no. 1, pp. 48–55, Apr. 2000.

[31] Z. Xia, "Semi-supervised Drug-Protein Interaction Prediction from Heterogeneous Spaces,"no.10631070,pp.123–131,2009.

[32] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces.," *Bioinformatics*, vol. 24, no. 13, pp. i232–40, Jul. 2008.

[33] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework.," *Bioinformatics*, vol. 26, no. 12, pp. i246–54, Jun. 2010.

[34] M. a Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network.," *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1119–26, Oct. 2007.

[35] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '13*, p. 1025, 2013.

[36] S. Zhu, Y. Okuno, G. Tsujimoto, and H. Mamitsuka, "A probabilistic model for mining implicit 'chemical compound-gene' relations from literature.," *Bioinformatics*, vol. 21 Suppl 2, pp. ii245–51, Sep. 2005.

VITA AUCTORIS

NAME:	Bharadwaja Allapalli
PLACE OF BIRTH:	Hyderabad, INDIA
YEAR OF BIRTH:	1988
EDUCATION:	University of Windsor, M.Sc., Windsor, ON, 2014
	Jawaharlal Nehru Technological University Hyderabad, B.Tech., Andhra Pradesh, INDIA, 2011