RANDOM WALK APPLIED TO HETEROGENOUS DRUG-TARGET NETWORKS FOR PREDICTING BIOLOGICAL OUTCOMES

Abhik Seal

Submitted to the faculty of the University
Graduate School in partial fulfillment of the
requirements
for the degree
Doctor of
Philosophy
in the School of Informatics and
Computing, Indiana University
January 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Dr. David J Wild

_____

Dr. Yong Yeol Ahn

_____

Dr. Ying Ding

_____

Dr. Sriraam Natarajan

Defense Date : 11 december 2015

ACKNOWLEDGMENTS

First, I would like to thank Prof. David Wild for advising my study and research in the last four years. David gave me all the freedom and flexibility to explore my research interests and encouraged me to pursue them. With the freedom and I had the opportunities to learn various new research topics including Network Science, Machine Learning, Pharmaceutical data mining and investigate real scientific questions Pharmaceutical industry. The dissertation would not be finished without the collaboration with Prof. Yong Yeol Ahn. Prof. Ahn guided me to this exciting research area of Network Science. I appreciate his guidance in the study and who have helped me to explore and understand new methods in network based prediction. I also would like to thank my research committee members for their comments on the dissertation. All the members from David's group, Ahn's group and Ying's group are thanked, especially, David's old students Jeremy Yang , Dr. Bin Chen and Dr. Jae Hong Shin and OSDD team lead by Dr. Anshu Bharadwaj, Dr. Abdul Jaleel and Dr. Yogeswari Perumal (BITS India Hyderabad) from India. I had a great pleasure of working with them. One invaluable and extraordinary experience during my graduate study is that I got four internship trainings from Academic and Pharmaceutical Industry. These internships not only led to the establishment of my technical expertise but also helped build my career. Other than David who generously allowed me spending all the summers outside, I would thank all my mentors and managers. They are Rishi Raj Gupta (Abbvie), Derek Debe (Abbvie), Dirk Tomandl (Dow Agrosciences), John Overington (EMBL) , George Papadatos (EMBL) and Mark Davies(EMBL.) I would like to thank Linda Hostetter, as "the mom of informatics", she really helped me a lot on the tedious work outside of research, from various form submissions to course registrations. I would also like to thank SOIC help desk team who has helped me in installing various tools in our server. Finally, I would like to express my gratitude to my mom and brother for their mental support in completion of my degree.

Abhik Seal

RANDOM WALK APPLIED TO HETEROGENOUS DRUG-TARGET NETWORKS FOR PREDICTING BIOLOGICAL OUTCOMES

Prediction of unknown drug target interactions from bioassay data is critical not only for the understanding of various interactions but also crucial for the development of new drugs and repurposing of old ones. Conventional methods for prediction of such interactions can be divided into 2D based and 3D based methods. 3D methods are more CPU expensive and require more manual interpretation whereas 2D methods are actually fast methods like machine learning and similarity search which use chemical fingerprints. One of the problems of using traditional machine learning based method to predict drug-target pairs is that it requires a labeled information of true and false interactions. One of the major problems of supervised learning methods is selection on negative samples. Unknown drug target interactions are regarded as false interactions, which may influence the predictive accuracy of the model. To overcome this problem network based methods has become an effective tool in predicting the drug target interactions overcoming the negative sampling problem.

In this dissertation study, I will describe traditional machine learning methods and 3D methods of pharmacophore modeling for drug target prediction and will show how these methods work in a drug discovery scenario. I will then introduce a new framework for drug target prediction based on bipartite networks of drug target relations known as Random Walk with Restart (RWR). RWR integrates various networks including drug–drug similarity networks, protein-protein similarity networks and drug-target interaction networks into a heterogeneous network that is capable of predicting novel drug-target relations. I will describe how chemical features for measuring drug-drug similarity do not affect performance in predicting interactions and further show the performance of RWR using an external dataset from ChEMBL database. I will describe about further implementations of RWR approach into multilayered networks consisting of biological data like diseases, tissue based gene expression data, protein-complexes and metabolic pathways to predict associations between

human diseases and metabolic pathways which are very crucial in drug discovery. I have further developed a software tool package netpredictor in R (standalone and the web) for unipartite and bipartite networks and implemented network-based predictive algorithms and network properties for drug-target prediction. This package will be described.

_____

Dr. David J Wild

_____

Dr. Yong Yeol Ahn

_____

Dr. Ying Ding

_____

Dr. Sriraam Natarajan

CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Despite many advances in the past decades, drug discovery is still a costly and time-consuming process. In recent years the rate of successful drug developed has decreased [1, 2] and in this light new indications for existing and abandoned drugs showing some promise [3]. Such a new strategy is called drug-repurposing [4]. Drug repositioning is also promising for shelved compounds because they failed in clinical trials and were not further investigated. These drugs could be quickly marketed for new indications [5, 6], thus reducing the attrition rates. These new interactions can also be useful for understanding causes of adverse effects of existing drugs. Traditional drug discovery follows a reductionist's approach, where a large complex system is divided into multiple parts. For example, a medicinal chemist assumes that ligands and their structure have sufficient information to provide an understanding of the behavior of target interaction and pharmacology. Connecting cellular components to tissue or organ-level based on gene expression data helps to identify new targets. Similarly identifying disease-based pathways requires gene expression data from tissues/organs, which forms proteins complexes, which

connects metabolic pathway. This is because different diseases like alzheimer's, ulcers, ischemic heart disease and cirrhosis occurs which occurs in different tissues like brain, stomach, heart and liver have different level of protein expression respectively. Actions of drug are changing the way it was use to described earlier in figure 1.1. Today it is clear that an adverse event or a pharmacological event occurs due to coordination of number of biological components in the system describe in figure 1.2. Interactions between the different components and influences from the environment, give rise to network behavior, which are absent in the isolated components [7].

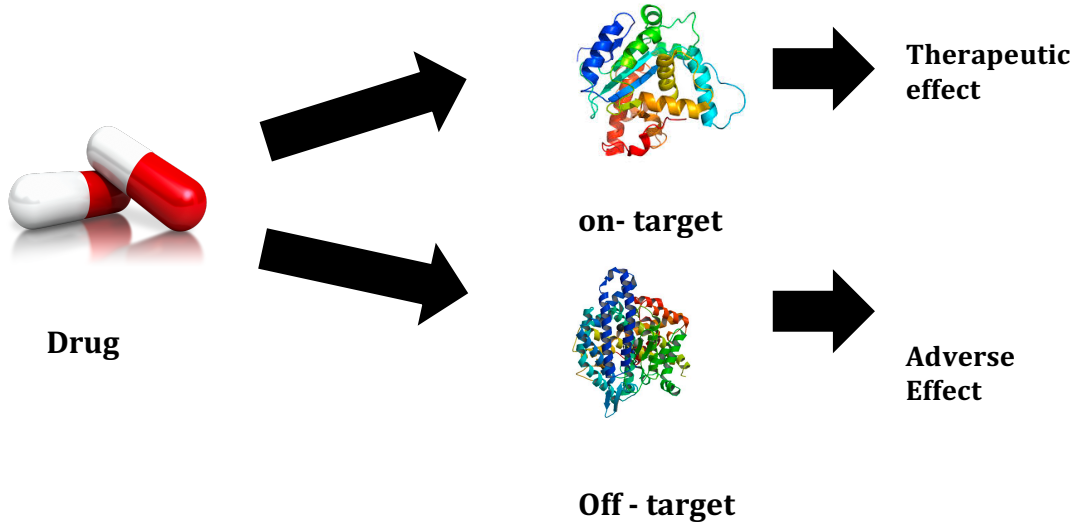

Figure 1.1 Classic drug target pharmacology [95]

Traditional approaches for drug target interaction prediction are generally based on virtual screening. Virtual screening or insilco screening is the use of high performance computing environments to screen compounds for drug candidates, which is classified into ligand-based and target-based approaches described in the next section.

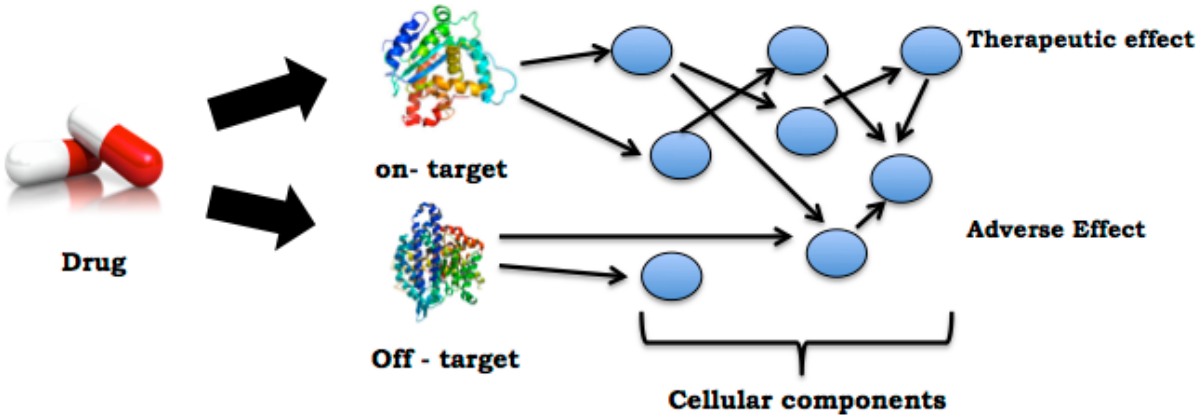

Figure 1.2 Network view of drug action idea taken from Berger and Iyenger [95]

Ligand-based approaches screen candidate compounds or ligands to predict whether they interact with a given target based on the assumption that similar drugs interact with the same target. The similarity of two drugs is measured in different ways with respect to different aspects. Other than comparing drugs according to their chemical structures [8], side effect has also been used to measure the similarity between drugs [9]. Assuming that similar targets bind to the same ligand, target-based approaches, on the other hand, compare proteins to predict whether they bind to the given ligand, or whether they are the targets of the given drug or compound. More specifically, for a given drug, new targets are identified by comparing candidate proteins to the known targets of this drug with respect to certain descriptors such as amino acid sequence, binding sites, or ligands that bind to them. Supervised machine learning using ligand based methods [10–13] drug-

target pairs are labeled as positive or negative samples according to confirmed interaction between corresponding drug and target pairs. The selection of negative samples is a common problem of all the supervised learning methods, as unknown drug–target interactions have been assumed as negative samples in the supervised learning methods. Selection of negative samples largely influences the predictive accuracy. It's difficult to decide the correct combination of datasets and fingerprints to prioritize targets. Selection of molecular descriptors plays a crucial role in drug discovery. Several fingerprints such as path based fingerprints (TT ad AP), substructure based fingerprints (MACCS and pubchem), circular fingerprints (ECFP, FCFP, PHFP) are used for prioritization of compounds. Performance of the fingerprints depends on different datasets. Mostly circular and path based fingerprints have high performance in retrieving active compounds than other types of fingerprints [95-98]. The random walk method with proper optimization of parameters discussed in chapter 2 of the dissertation will show how one can use any kind chemical fingerprints to get similar results for target prioritization. The random walk based method described in this dissertation overcomes the limitation for negative sampling problem and also it doesn't require the labeled information of drug–target interaction. One can give unlabeled information of drugs and after computation the algorithm predicts the interactions. In chapter four, another problem we are trying to address if the method of random walk with restart can be implemented using multipartite networks, which integrates a heterogeneous network structure contains more 3 kinds of networks. Chapter 5 discusses about the netpredictor standalone and web software which computes the network

properties and computes four different algorithms HeatS, network based inference (NBI), random walk with restart (RWR) and netcombo (NBI+RWR). The method, which is implemented in this dissertation, falls under ligand-based approach.

The next section describes about the traditional types of virtual screening.

1.2 VIRTUAL SCREENING STRATEGIES

Computational ligand (compound) design is divided into two strategies, ligand based and target based drug design that can be used together or independently. Ligand based drug design relies on set of active and inactive ligands where no 3D structural information of the target is available. Structure based drug design is applicable when the target's 3D information is available along with the binding site information. Figure 1.3 shows how virtual screening in classified.

Depending upon structural and Bioactvity data available:

- One or more actives molecule known perform similarity searching.
- Several active known try to identify a common 3D pharmacophore and then do 3D database search.
- Reasonable number of active and inactive known train a machine learning model.
- 3D structure of protein known use protein ligand docking.

People working in pharmaceutical industry does not follow a specific route whatever information they have follow a hybrid of methods. Figure 1.4 shows how both the ligand based and structured based design can be integrated. While designing new ligands one starts with a database of chemical structures

and performs a search for similar structures based on 3D shape or 2D chemical fingerprints. Then if a 3D crystallographic protein structure is avail- able with a ligand attached then one can design a structure-based pharmacophore [14, 15]. If the 3D protein structure is missing one can go for the ligand based pharmacophore and then select some potential lead compounds. If there exists a 3D structure then one can use 3D docking tool to find the pattern of interactions.

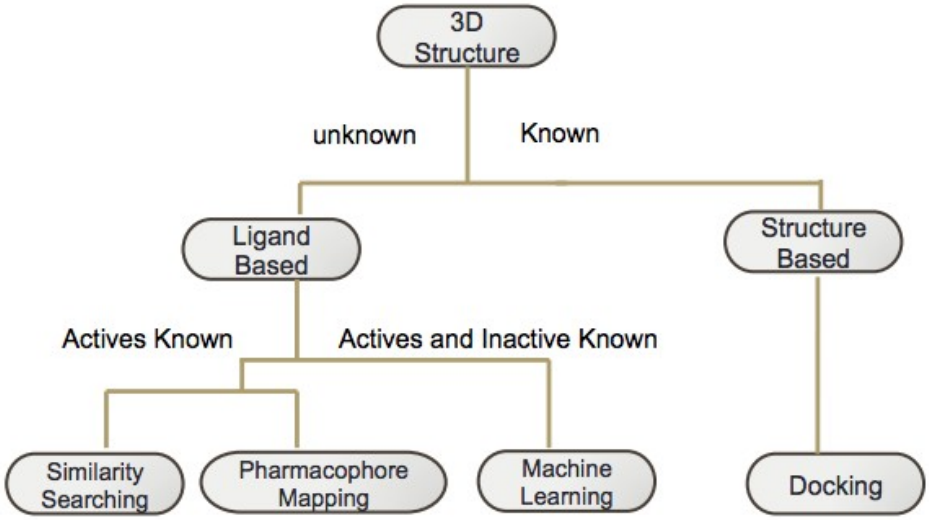

Figure 1.3: Diagram showing Virtual screening pathways.

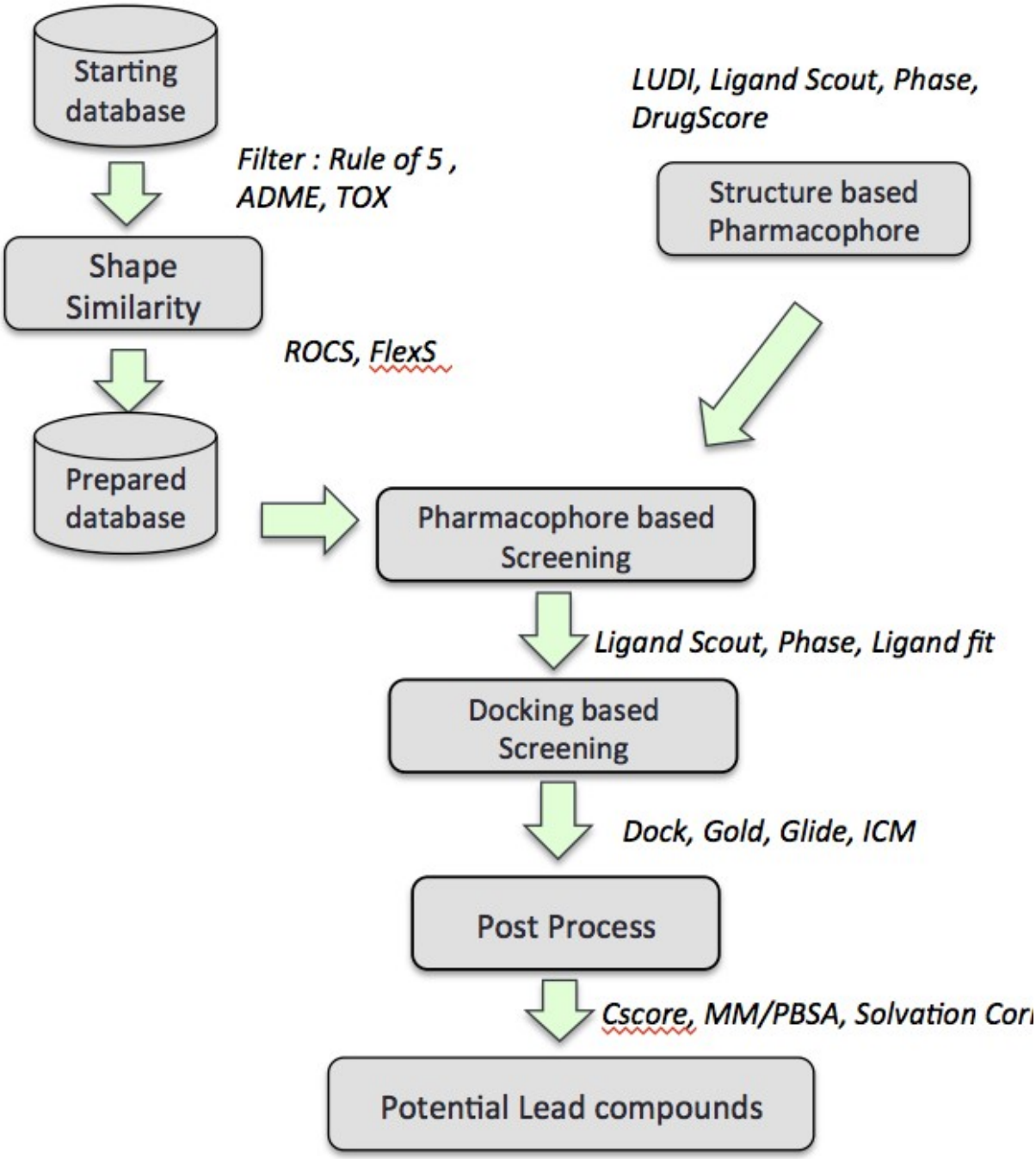

Figure 1.4: Diagram showing hybrid Virtual screening

1.3 OVERVIEW OF THE STUDY

1.3.1 RANDOM WALK WITH RESTART ON DRUG TARGET HETEROGENOUS NETWORK

Predicting novel drug–target associations is important not only for developing new drugs, but also for furthering biological knowledge by understanding how drugs work and their modes of action. Network based description and analysis not only give a systems-level understanding of drug action and disease complexity, but can also help to improve the efficiency of drug design. As more data about drugs, targets, and their interactions becomes available, computational approaches have become an indispensible part of drug target association discovery. In this chapter we apply random walk with restart (RWR) method to a heterogeneous network of drugs and targets compiled from drugbank database and investigate the performance of the methods under parameter variation and choice of chemical fingerprint methods.

Random walk is a useful mathematical framework that provides a systematic way to measure importance of nodes in a network. The most widely known is the PageRank algorithm [29]. PageRank, developed for ranking web pages, measures page clicks of hypothetical web surfers who randomly click hyperlinks in the network of webpages. Since it is possible for the surfer to be trapped in a dead-end webpage that does not have any connection to the main network, at each time step the surfer may jump to a random webpage with a probability *c*. Interestingly, this

formulation also provides a simple way to define a random walk-based "distance" from a node a (or a set of nodes) to every other node, namely by allowing the random walkers to jump only to the source node a (or the source set of nodes) and restart from there. As a result, it is more likely to find the random walker at the vicinity of the source node than at a distant part of the network, and thus we are able to estimate the relevance (closeness) of each node with respect to the source node. The prediction method applies this idea to identify drugs and targets that are relevant to a set given set of drugs and targets.

Consider an undirected, unweighted network G = (V, E), where V is the set of nodes and E is the set of links. For each pair of nodes a,b$\in$V we can assign a proximity score by executing the following procedure:

– we start a random walker from a.

– At each time step, with the probability 1 - c, the walker walks to one of the neighbors, b, according to the transition value matrix $W_{ab} = \frac{S_{ab}}{K_a}$ where S_{ab} is the adjacency matrix of the network and (S_{ab} equals 1 if node a and b are connected, 0 otherwise) K_a denotes the degree of a.

– With probability c , the walker goes back to a.

– After many time steps the probability of finding the random walker at node x converges to the steady-state probability which is our proximity score $S_{a\to x}$.

The random walk with restart, whose updating equation is shown as follows:

$$p_{t+1} = (1-c)W^T p_t + cp_0 \qquad (1.1)$$

Keep updating p until convergence; the stationary distribution vector p can meet,

$$pt = (1 - c)(I - cW^T)^{-1} p_0 \qquad (1.2)$$

We show that choice of chemical fingerprint does not affect the performance of the method when the parameters are tuned to optimal values. We use a subset of the ChEMBL15 dataset that contains 2,763 associations between 544 drugs and 467 target proteins to evaluate our method, and we extracted datasets of bioactivity ≤ 1 and ≤ 10 μM activity cutoff. For 1 μM bioactivity cutoff, we find that our method can correctly predict nearly 47, 55, 60% of the given drug–target interactions in the test dataset having more than 0, 1, 2 drug target relations for ChEMBL 1 μM dataset in top 50 rank positions. For 10 μM bioactivity cutoff, we find that our method can correctly predict nearly 32.4, 34.8, 35.3% of the given drug–target interactions in the test dataset having more than 0, 1, 2 drug target relations for ChEMBL 1 μM dataset in top 50 rank positions. We further examine the associations between 110 popular top selling drugs in 2012 and 3,519 targets and find the top ten targets for each drug. We demonstrate the effectiveness and promise of the approach—RWR on heterogeneous networks using chemical features—for identifying novel drug target interactions and investigate the performance.

1.3.2 DISEASE TO PATHWAY PREDICTION USING RANDOM WALK WITH RESTART

Besides identifying individual disease related genes, associating pathways to human inherited diseases is of great importance, because the disease conditions arise from the cooperative behavior of multiple proteins in protein interaction network which forms protein complexes and plays an important role in disease pathways. Integrating pathway level data would play a key role in understanding mechanism of action of diseases. It is well known that genes within a cell do not function alone. They interact with each other to form complexes and form pathways to carry out biological functions. Also identifying the pathways could help in design and repurpose drugs of similar or unknown diseases with similar symptoms. Here in this chapter I have designed a system, which is composed, of disease data connected to 60 different tissues [30] based protein interaction network which is developed from the expression profiles of human proteome, and associate them with protein complexes to biological pathways. We propose a random walk based model to query a specific disease, which then loads the disease tissue, based protein-protein interaction network and its pro- teins complexes and identifies biological pathways associated with the disease. With several leave-one out validations we optimized the network to achieve best results. The results can be used to predict unknown pathways associated with the disease and would help in drug repurposing related to those pathways.

1.3.3 NETPREDICTOR SOFTWARE FOR NETWORK BASED PREDICTION

Searching missing associations between drug and targets is valuable to understand polypharmacology and as well as understand off-target mediated effects of chemical compounds in biological systems. Traditional machine learning algorithms like Naive Bayes, SVM and Random Forest have been successfully applied to predict drug target relations. However, using supervised machine learning method we need to label the drug-target pairs with negative and positive samples to understand the known relation between drug and target is known or not. Therefore, unknown drug target relations are regarded as negative and im proper negative sample selection can largely affect the predictive accuracy. Network based models tries to avoid these kind of issues on negative sampling biases. Cheng [31] developed a technique based on Network Based Inference (NBI) and developed three supervised methods on drug similarity, target similarity and network topology and showed superior performance of network topology based method. Alaimo [32] have extended Cheng's method of network model to integrate Chemical and target similarity into account to show that the performance of the method superior to Cheng's model. Chen [33] and Seal [27] have used random walk with restart (RWR) based method to predict drug target interactions on a heterogeneous network made up of drug-drug similarity, protein-protein similarity and bipartite graph between drugs and targets. Seal [27] have extended the method by optimizing a parameter η, which showed that the

performance of RWR is independent of the choice of using Chemical fingerprint features. The netpredictor package provides the implements the Random walk with Restart algorithm in a biparite and unipartie network and also implements the Alaimo's algorithm of network-based inference. The algorithm also implements method to predict the unknown relations and the relations and perform permutations tests on the given network for predicted values.

1.4 RELATED WORK

Random walk with restart is being applied to many bioinformatics related problems in prioritization of diseases based on protein - protein interaction network [34] non-coding RNAs [35]. Campillos etal. [36] established a method using drug side-effect similarity to find diverse compounds binding a similar target. Cheng etal [37] developed three super- vised inference models namely drug-based similarity inference (DBSI), target-based similarity inference (TBSI) and network-based inference (NBI) to predict drug target interactions. DBSI and TBSI depends on chemical structure similarity and target sequence similarity, respectively, whereas NBI is only based on drug–target bipartite network topology similarity. Yamanishi [38] proposed a bipartite graph learning method to predict drug-target interactions by integrating the chemical structure information, the sequence similarity information and known drug-target information into a supervised kernel-regression method to predict new drug-target interactions. He further proposed a pharmacological similarity based network [39] learning method where he

integrated pharmacological similarity into supervised bipartite graph model to identify new drug-target interactions.

1.5 CONCLUSION

In this dissertation we are using network based random walked with restart to predict biological outcomes. One of the outcomes is prediction of drug target interactions and another one is prediction of metabolic pathways of diseases using tissue-based protein – protein interaction network and protein complexes information network. The paper from Chen etal described the way to performing RWR in bipartite network with chem-biological data. However there were certain limitations on that work. The paper focused on prediction of drug targets based on certain groups of proteins like GPCR, Ion Channels, Enzymes and Nuclear Receptor. However, it didn't consider the off class interactions of drugs. In order to check the off class cross prediction we created a full data of 3519 proteins and then we predicted the off class interactions with relative good performance. Testing drug target interactions at different activity endpoints is also very crucial in understanding lead compounds. We tested our methods with ChEMBL data at 10000 μM and 1000 μM and it showed very good performance. Another important fact that came out of this dissertation is we know chemical fingerprints plays a crucial role in prioritization of targets. We have optimized parameter η to 0.01, which controls the importance of two kinds of nodes, i.e. drug node and target node. When this parameter is optimized we can control the use of chemical similarity matrices. It means that whatever similarity matrices you use one can exactly same prioritization results. This will help in using open public version of fingerprints than commercial versions. Also Chen

etal didn't release any kind of codes for there computation we have developed a standalone and web based application of the random walk with restart and network based inference methods for prioritization of targets and is freely available.

Other than prioritization of drug targets, with the random walk framework can also be used to integrate multi partite networks and predict outcomes. We wanted to predict disease based metabolic pathways. Diseases occur in different tissues and organs also depend upon the protein expression levels. Diseaes such as Alzheimer's, ulcer which occurs in brain and stomach respectively, the gene expression levels would also vary . In order to predict the disease based metabolic pathways we used four-layered network consisting of diseases, tissue based protein expression data, protein complex information and protein pathways network. The method can predict the biological pathways, the proteins which are involved and it can help to prioritize complex disease based pathways.

CHAPTER 2

RANDOM WALK APPLIED TO DRUG TARGET BIPARTITE NETWORK

2.1 INTRODUCTION

Recent work has demonstrated the power of network-based approaches in drug discovery [33, 40, 41]. We have shown previously that a large semantic network of drug–target interactions provides a powerful framework for predicting new associations [42] and that an algorithm that predict drug-target associations by using this network performs surprisingly well, even without training datasets or incorporating target preference [43]. zthis chapter, we apply a random walk-based link prediction algorithm based on Chen et al. [33] to a more extensive drug–target network from drug-bank and evaluated its performance using an external bioactivity dataset from ChEMBL 15 database. We combine three networks drug-drug, target-target, and drug-target to construct a heterogeneous network of drugs and targets. The links between drugs are obtained by quantifying molecular similarity with chemical fingerprints and examining the shared targets. The links between targets are obtained by calculating local sequence similarity between proteins and again examining the links between shared drugs.

2.2 METHODS

We apply the RWR algorithm to a drug–target network and use an external dataset extracted from ChEMBL 15 (544 drugs and 467 proteins) at bioactivity cutoff points of 10μ M and 1 μ M to quantitatively evaluate the performance and robustness of the approach.

2.2.1 DATASETS

For the drug dataset, I compiled a set of approved drugs from DrugBank database (Version 3.0) [44], consisting of 727 compounds and 3519 protein targets. To construct the network between drugs, we incorporate two types of similarity measures: chemical (structural) similarity and target similarity. We calculate chemical similarity be- tween drugs by using the Jaccard Index (Tanimoto Coefficient) between their chemical fingerprints. The Jaccard Index is defined as the size of the intersection of two sets di- vided by the size of the union of the sets, ranging between 0 and 1. For binary vectors like chemical fingerprints, it is defined as C/(A + B-C) where C is the number of bits in common, A is the number of bits in one of the fingerprints, and B is the number of bits in the other fingerprint. We use four types of chemical features namely, MDL MACCS166 keys (fragmental descriptors) [45], ECFP6 fingerprints (extended connectivity fingerprint path 6) [46], 2D Pharmacophore fingerprints (PHFP4) [47] and ROCS program which uses Tanimoto combo similarity—which combines shape and color measures of a compound, we calculate them with ROCS program [48].

ECFP (extended connectivity fingerprint) encodes information on atom-centered fragments that is derived from the variant of the Morgan algorithm [49]. ECFPs are generated using the neighborhood of each non-hydrogen atom into multiple circular layers up to a given diameter. These atom-centric substructural features are then mapped into integer codes using a hashing procedure, which constitute

the extended-connectivity fingerprint.

ECFP can, for instance, represent a very large number of features (over 4 billion), do not rely on predefined dictionary of features, can represent stereochemical information, and can be interpreted as the presence of particular substructures. 2D pharmacophore fingerprints are calculated using topological (bond) distances.

Pharmacophore fingerprints consist of pairs, triplets, or quartets of molecular features and the corresponding bond distances among them. We use PHFP 4 (quartets which includes number of bonds in the shortest path between the features) fingerprints for the calculation. The feature vectors of quartets involve four pharmacophoric features, six Euclidean distances separating those features, and an indication of chirality. For 3D alignment and similarity we used ROCS 3.2, which is a shape-similarity method based on the Tanimoto-like overlap of volumes. The alignment was developed using the Combo score, which combines the Tanimoto shape score with the color score that added the score for the appropriate overlap of groups with similar properties (donor, acceptor, hydrophobe, cation, anion, and ring) [http://docs.eyesopen.com /rocs/shapetheory.html] defined by SMARTS. Conformers for the data set is created using OMEGA [50], about 250 conformers with RMSD threshold of 0.6 is generated. ROCS performs shape-based overlay of conformers as atom-centered Gaussian functions. ROCS score performed in color optimization mode where it optimizes the molecular overlay to maximize both the shape overlap and the color overlap obtained by aligning

groups with the same properties that are contained in the color force field file. This overlay is then subsequently scored using the sum of shape Tanimoto for the overlay and the color score called Tanimoto combo score. We use C_S to refer the N-by-N chemical compounds similarity matrix. For the 727 drugs we used different chemical descriptors to calculate the Tanimoto similarity distribution to create a view of how similar the drugs look like. The distributions of different similarities Figure. 2.1 shows that for four finger- prints (166 MACCS Keys, PHFP4, 3D ROCS, and ECFP6), 0.56% had a similarity above 0.7 for the MACCS keys, 0.31% had similarity above 0.4 for PHFP4, 0.88% had similarity above 1.2 Tanimoto Combo score for ROCS, 0.24% had similarity above 0.3 for ECFP6. The mean similarity is 0.346, 0.019, 0.742, and 0.063 for MACCS, PHFP4, ROCS, ECFP6 fingerprints, respectively. This indicates how diverse chemical structures are in the drug dataset.

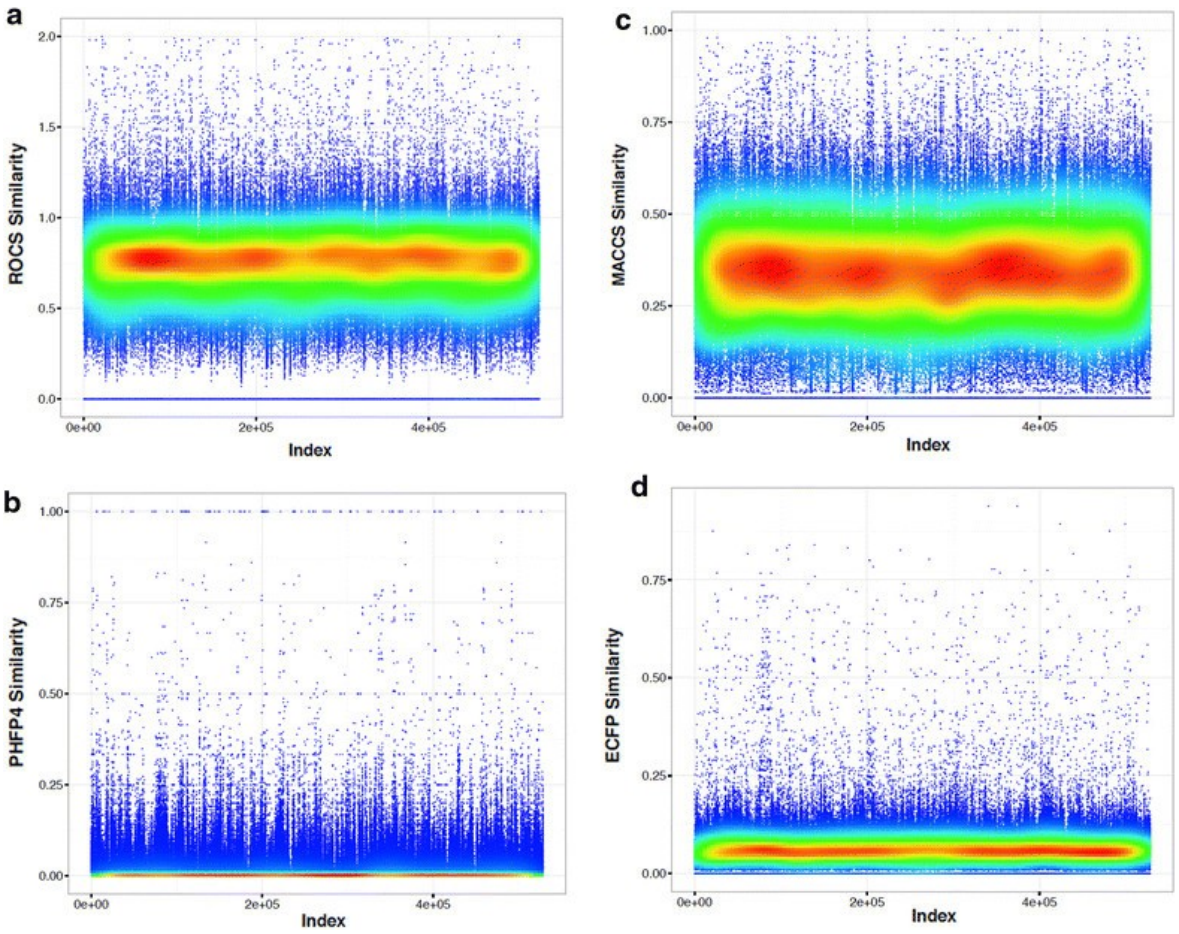

Figure 2.1: Diagram showing distribution of chemical fingerprints.

For the protein dataset, I extracted 3,519 target proteins across all available species and their sequences from the DrugBank database. As proteins in other species may provide useful information in our network-based approach, we keep all the proteins regardless of species. Note that, human proteins still dominate the dataset. We calculate the sequence similarity matrix T_S by using the R biostrings package and the normalization procedure proposed by Bleakley and Yamanishi [41].

$$T_s = \frac{SW(g,g')}{\sqrt{SW(g,g)}\sqrt{SW(g',g')}}, \quad (2.1)$$

where SW (·, ·) means the original Smith–Waterman similarity score.

We constructed the drug-target relationship matrix A whose element A(i,j) is 1 if drug *i* interacts with target *j*, otherwise 0. The matrix is sparse; the total number of connections among the drugs and targets is only 2,557, with 687 drugs having at least one known target and with 628 proteins having at least one drug. There are 73 connected components in the whole drug target network dataset. The largest connected component in this bipartite graph has 498 drugs and 279 proteins. The connections are concentrated to a small number of drugs (see Fig. 2.2a) that affect nervous systems mostly psychoanaleptics and psycholeptics have the largest number of interactions. As most drugs are metabolized by cytochrome p450, which serves as an important protein target and enzyme for the drugs, the interaction between important enzymes CYP3A4, CYP2D6 and CYP3A5 are not considered on the drug target interaction matrix except for the drug paliperidone, which has interactions to all the three cytochromes targets mentioned above.

Figure 2.2b exhibits the targets that interact with most number of drugs. The top frequent targets are Muscarinic receptor (ACM1), Adrenoreceptor alpha 1A (ADA1A), Histamine receptors (5HT2A), and dopamine receptors (DRD2). In addition to the drug–drug similarity matrix C_S (based on chemical similarity) and target–target similarity matrix T_S (based on

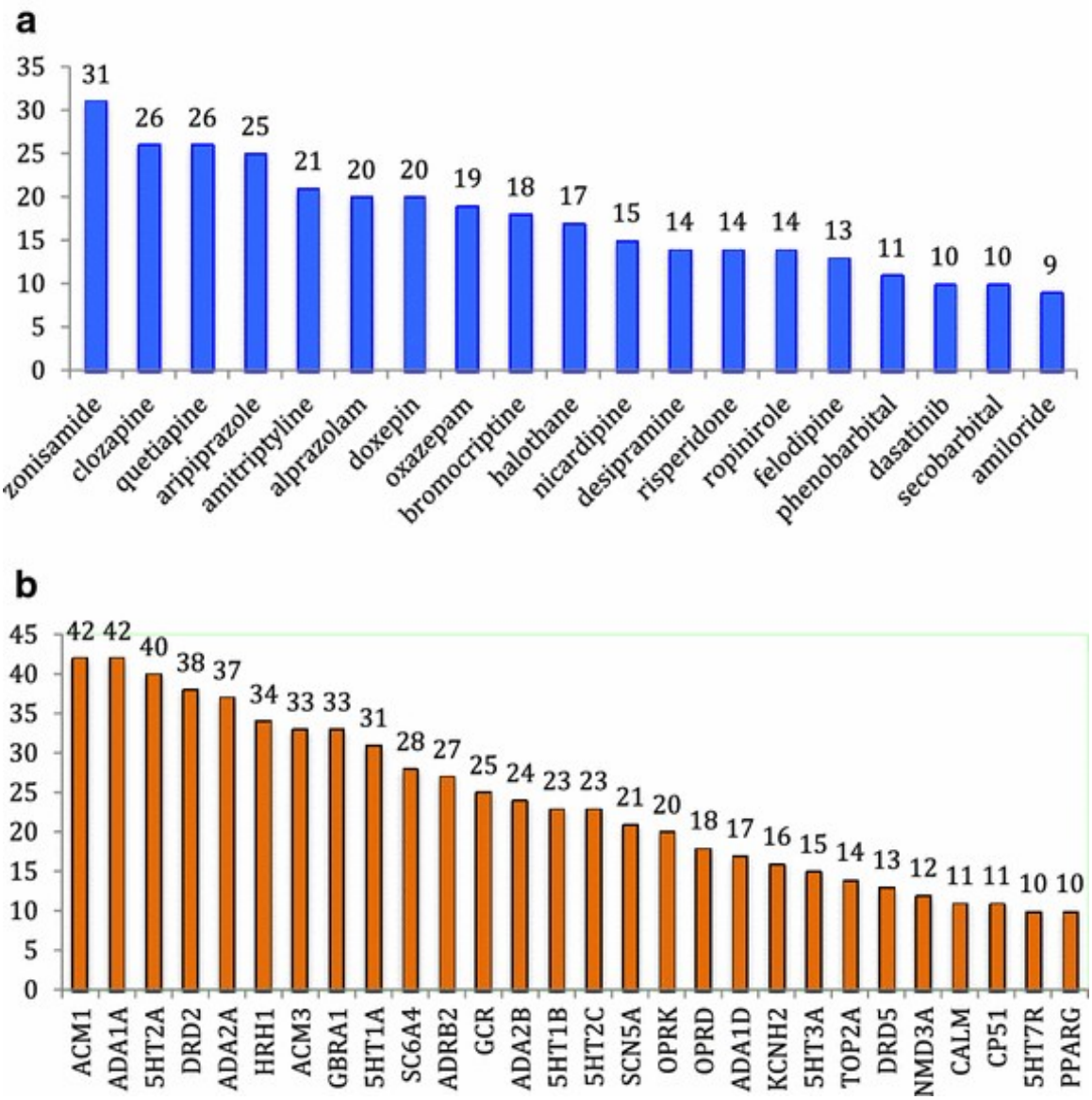

Figure 2.2: Diagram showing the distribution of compounds and targets in drugbank dataset.

sequence similarity), we introduce additional measure of drug–drug and target–target similarities based on the network structure. C_s^n is a drug–drug similarity matrix based on the number of shared targets between drugs; $T_s^n$is a target–target similarity ma- trix based on the shared drugs. The similarity between two drugs di and dj is quantified by Jaccard

coefficient, which is defined by:

$$C_s^n(d_i, d_j) = \frac{M_l(i,j)}{M_l(i,i)+M_l(j,j)-M_l(i,j)}, \quad (2.2)$$

where, M_l is the inner product of the drug-target interaction matrix. The similarity between targets is defined in the same manner. We define the final drug-drug similarity matrix S_d by taking a linear combination of the chemical similarity matrix (C_s) and target sharing similarity matrix (C_s^n). Similarly, the final target-target similarity matrix S_t is calculated using the sequence similarity matrix (T_s) and drug sharing similarity matrix (T_s^n).

2.2.2 RANDOM WALK WITH RESTART IMPLEMENTATION

We combined drug-drug, drug-target, and target-target networks into a undirected heterogeneous network. Many nodes have connections to both drugs and targets and we call them *bridge nodes*. At a bridge node, a random walker may jump to a node with the other type or to a node with the same type. The probability to do so is λ and 1-λ respectively. For instance, if a random walker is at a drug node, it can jump to one of the connected target nodes with the probability λ, or jump to connected drug nodes with the probability 1- λ. We call the parameter λ the *jumping probability*. If λ is 0, a random walker will explore only one type of networks. Most importantly, the probability $p_\infty(i)$ is the probability of finding the random walker at node i in the steady state. It gives a measure of probability of source and target node (proximity) between node i and the source nodes where the random walks restarts.

The transition matrix is represented by,

$$W = \begin{bmatrix} W_{TT} & W_{TD} \\ W_{TD} & W_{DD} \end{bmatrix}$$

Here W_{TT} *is* the target to target transition matrix, W_{DD} is the drug to drug transition matrix, W_{DT} is drug to target transition matrix and W_{TD} is target to drug transition matrix. The calculation of each of the transition matrix in discussed in Chen *et al* [3]. The calculation of each of the transition matrix in discussed in Chen [33] given below in equation 2.3, 2.4, 2.5 and 2.6. The random walk is implemented on the heterogeneous network using the Eq. 2.7 given below,

The transition values of target vertexes from ti to tj is defined as

$$W_{TT}(i,j) = \begin{cases} \frac{S_t(i,j)}{\sum_j s_t(i,j)} & if \ \sum_j A(i,j) = 0 \\ \frac{(1-\lambda)S_t(i,j)}{\sum_j s_t(i,j)} & otherwise \end{cases} \quad (2.3)$$

The transition values of drug vertexes from di to dj is defined as

$$W_{DD}(i,j) = \begin{cases} \frac{S_d(i,j)}{\sum_j s_t(i,j)} & if \ \sum_j A(i,j) = 0 \\ \frac{(1-\lambda)S_d(i,j)}{\sum_j s_d(i,j)} & otherwise \end{cases} \quad (2.4)$$

The transition values from target vertex ti to drug vertex dj is defined as

$$W_{TD}(i,j) = \begin{cases} \frac{\lambda A(i,j)}{\sum_j A(i,j)} & if \ \sum_j A(i,j) \neq 0 \\ 0 & otherwise \end{cases} \quad (2.5)$$

The transition values from drug vertex di to target vertex tj is defined as

$$W_{DT}(i,j) = \begin{cases} \frac{\lambda A(j,i)}{\sum_j A(j,i)} & if \ \sum_j A(j,i) \neq 0 \\ 0 & otherwise \end{cases} \quad (2.6)$$

$$p_{t+1} = (1-c)W^T p_t + cp_0 \quad (2.7)$$

p_t is a vector in which ith elements holds the probability of finding the random walker at node i at time step t. Initial probability vector p_0 controls the restart probability c.

$$p_0 = \begin{bmatrix} (1-\eta)u_0 \\ \eta v_0 \end{bmatrix} \text{ (initial probability matrix)}$$

u_0 and v_0 be the initial probability vectors for target network and drug network, respectively. Parameter η controls the importance of two kinds of seed nodes, i.e. drug node and target node. We tested the importance parameter η for different values ranging from 0 to 1.

After a number of iteration steps, the p_t converges to a steady-state probability vector p_∞, where $p_\infty = \begin{bmatrix} u_\infty \\ v_\infty \end{bmatrix}$. In practice, we consider $p_t = p_\infty$ if the change between p_t and p_{t+1} (measure by the Frobenius norm) is less than 10^{-10}.

For finding novel targets for a given drug, we set the drug and the targets that are directly connected to the drug as our seed nodes. Suppose that there are six targets $T_1,\ldots,T_6$ and four drugs D_1, D_2, D_3, and D_4. We focus on drug D_3 and tries to find novel targets for D3. We already know that D_3 interacts with T_2 and T_3. Then T_1, T_4, and T_5 are candidate targets for drug D_3. We set T_2, T_3, and D_3 as the source nodes, namely

$$u_0 = [0,1,1,0,0,0]^T \quad \text{and} \quad v_0=[0,0,1,0]^T$$

The stationary probability p_∞ represents the expected relevance of each drugs and targets regarding the source node set *T2, T3* and *D3*. For instance, if the value for *T1* is the largest among *T1*, *T4* and *T5*, then we expect that *T1* is most likely to interact with *D3*.

2.3 RESULTS AND DISCUSSION

2.3.1 EVALUATING PREDICTION PERFORMANCE USING LINK PERTURBATION

The network-based method aims to predict new targets for a given drug. We evaluated our approach using a perturbed network where we have removed some links to measure how well our approach re-identifies those removed links. There are five parameter to explore: the restart probability c, the jumping probability λ, the relative importance η, which controls the relative importance between two types of seeds, w_d and w_t that weigh the drug and target similarity matrices and network based similarity measure of the drugs and proteins, respectively. Among these five parameters, we have tested η because, to our knowledge, the restart probability *c,* jumping probability λ and w_d and w_t are not likely to affect the results in a significant way. First, it is known that in most cases the choice of restart probability *c* does not affect performance of PageRank algorithm and other PageRank based algorithms. For instance, the results of PageRank are highly insensitive to the choice of restart probability [28, 51] It has been shown that the prediction results from RWR are also robust [52, 53]. Because of these evidences, we

simply adopt the previously used value of 0.3 [33]. Second, the robustness of λ (jumping probability) has already been discussed [52–54]. It has been shown that the weight parameters w_d and wt are robust among the prediction results [33].

In our drug target network 684 (94%) drugs have at least one target. I prepare a test network of 684 drugs where I remove one links from 684 drugs with a total of 684 drug–target interactions. The links include drugs which has only one target in order to see if the method able to predict single known interaction. We checked how many missing links are in top N of the ranked list. We divided the number of actual targets that are in the top N lists by the number of tests (684) and call the fraction as 'recovered fraction'. I also used a random set to calculate the statistics with same parameters and found that the results are way better than random set. I tested our results with different values of w_d and wt ranging from 0 to 1 and found that at extreme point like 0 and 1 the performances drops radically but the performance gets best on values of w_d and wt of 0.5 given in Additional file 3: Sheet 3 (http://www.jcheminf.com/content/7/1/40#sec5). We tested different values of η for the four different chemical fingerprints to identify the optimal value of η and the right of chemical features. We observed that the prediction performance becomes optimal when η is small but not 0. I found optimal performance at η=0.01. For all the other values of η(0.1-0.9) the prediction rate for all fingerprints is equal. We found nearly 28% of the true interactions out of 684 can be retrieved at the top 10 rank positions and more than 38% of the interactions can be retrieved at the

top 50 rank positions. We also prepare 10 test networks of drugs that have more than two targets links, where we randomly remove 100–1,000 links. Using the 10 test networks we predicted the removed links. We repeat this process, from preparing a test network to calculating the recovered fraction, 50 times to obtain the 'average recovered fraction'. From Table 1 we can see that if we remove 100 links it gave us the best prediction rates and as we increase the number of removed links to 1,000 the prediction rates falls. From Table 2 shows the recovered fraction rates for top 10, 25, 50, 100, 200, 500, 1,000 retrieved targets we also find almost 32% of the true interactions can be retrieved at the top 10 rank positions for each of the test networks and more than 75% of the true interactions can be retrieved at the top 50 rank positions. This indicates that the method performs well if I remove links from drugs which are having at least two or more known interactions, since it uses the given interaction information in the network. I also measured the area under accumulation curve, area under ROC curve AUC (Top 10%), BEDROC and enrichment factor given in Table 1. The area under the receiver operating characteristic (ROC) curve (AUC) is widely used to evaluate the performance of the ranking method. The advantage of using AUC is, the value ranges from 0 to 1 with 0.5 corresponding to randomness. Another key criterion for measuring the success of ranking prediction is the enrichment of annotated associations among top ranking associations. The higher the percentage of annotated associations among the top ranking associations, the better the performance of the prediction. The enrichment criterion is evaluated by

enrichment factor (EF) [52, 53]. EF reflects the capability of a screening application to detect true links (true positives) compared to random selection. Thus, its value should always be greater than 1 and the higher it is, the better the enrichment performance. When we are predicting links it should rank true links in the top-ranking list. Metric likes ROC not sensitive to early recognition for example considering cases like where (1) true links are retrieved at beginning of a rank ordered list, (2) where true links are randomly distributed and (3) where true links, which are retrieved in the middle of the rank, ordered list. In all of the above cases ROC is 0.5 but in terms of early recognition we see that case (1) is better than (2) and (3). To overcome these limitations methods such as RIE and BEDROC have been proposed. By changing the tuning parameter, α, one can test whether the method is able to rank true links early or not.

Number of links removed	AUAC	AUC	BEDROC	EF	AUC(top 10%)
100	0.947	0.991	0.833	9.23	0.867
200	0.938	0.995	0.827	9.100	0.857
300	0.930	0.995	0.818	8.95	0.845
400	0.920	0.991	0.805	8.79	0.830
500	0.916	0.997	0.801	8.71	0.824
600	0.908	0.995	0.789	8.56	0.812
700	0.899	0.981	0.780	8.42	0.802
800	0.885	0.997	0.761	8.20	0.783
900	0.869	0.955	0.741	7.91	0.765
1000	0.854	0.956	0.715	7.62	0.741

Table 2.1: Shows the statistical metrics with the number of links removed

We found that the performance of the algorithm for ranking the targets by different chemical features is approximately same which indicates using this approach a user can identify protein targets with any one set of chemical features. We used public 166 MACCS keys, ECFP6, PHFP4 and 3D ROCS to perform the analysis and it is surprising that the commercial programs feature performance is same as the 166 public MACCS keys.

As a baseline, we test how RWR results differ from the results of random set of interactions. We randomized the interactions and similarity matrices and performed RWR and found the random set prediction rate was way below our original prediction rate as given in Additional file 3: Sheet 1.

(http://www.jcheminf.com/content/7/1/40#sec5)

# of links	TOP 10	TOP25	TOP50	TOP100	TOP200	TOP500	TOP1000
100	32.24	78.24	87.76	90.74	91.92	93.22	93.88
200	31.92	77.95	87.26	89.86	91.15	92.37	93.12
300	32.14	78.31	86.82	89.48	90.68	91.8	92.63
400	32.04	77.4	85.34	88.07	89.24	90.33	91.45
500	32.62	77.39	85.04	87.56	88.7	89.95	91.1
600	32.53	76.21	83.68	86.23	87.54	88.86	90.16
700	32.5	75.64	82.69	85.18	86.57	87.89	89.33
800	33.06	74.13	80.88	83.45	84.86	86.35	87.97
900	33.58	72.14	78.49	81.04	82.77	84.57	86.38
1000	33.71	69.81	76.008	78.31	80.22	82.12	84.42

Table 2.2: Shows the recovered fraction rates values with the number of links removed.

2.3.2 EVALUATING PREDICTION PERFORMANCE USING EXTERNAL DATASET FROM CHEMBL

In addition to the internal evaluation using link perturbation approach, we evaluate the performance of our method using an external dataset, namely ChEMBL version 15 database. From ChEMBL 15 data we extract all the drugs and targets that have activity values not more than 1 μM additional file 3 sheet (http://www.jcheminf.com/content/7/1/40#sec5) and and 10 μM additional file 3 sheet 4 (http://www.jcheminf.com/content/7/1/40#sec5) with units IC50, Ki, Kd, EC50, AC50, LC50, and GI50. Our training model is based on DrugBank and UniProt database so we mapped the drugs and targets ChEMBL ids with the DrugBank ids and UniProt ids. We used pubchem mapping tool (http://pubchem.ncbi.nlm.nih,gov/idexchange) to map ChEMBL ids to DrugBank ids and the UniProt mapping tool (http://www.uniprot.org/?tab=mapping) to map target ChEMBL ids to uniprot ids. It gives us 544 drugs and 467 protein targets, with 3,463 and 564 drug target interactions those are below 10 and 1 μM, respectively. Naturally, there are lots of interactions that are present in both DrugBank and ChEMBL. We tested performance of parameter η at different values on ChEMBL 1 μM set and 10 μM having which have more than 0, 1 and 2 target relations. Figures 2.3 and 2.4 shows the recovered fractions against the rank with different η (eta) values for ChEMBL data at 1 and 10 μM cutoff with different fingerprints respectively.

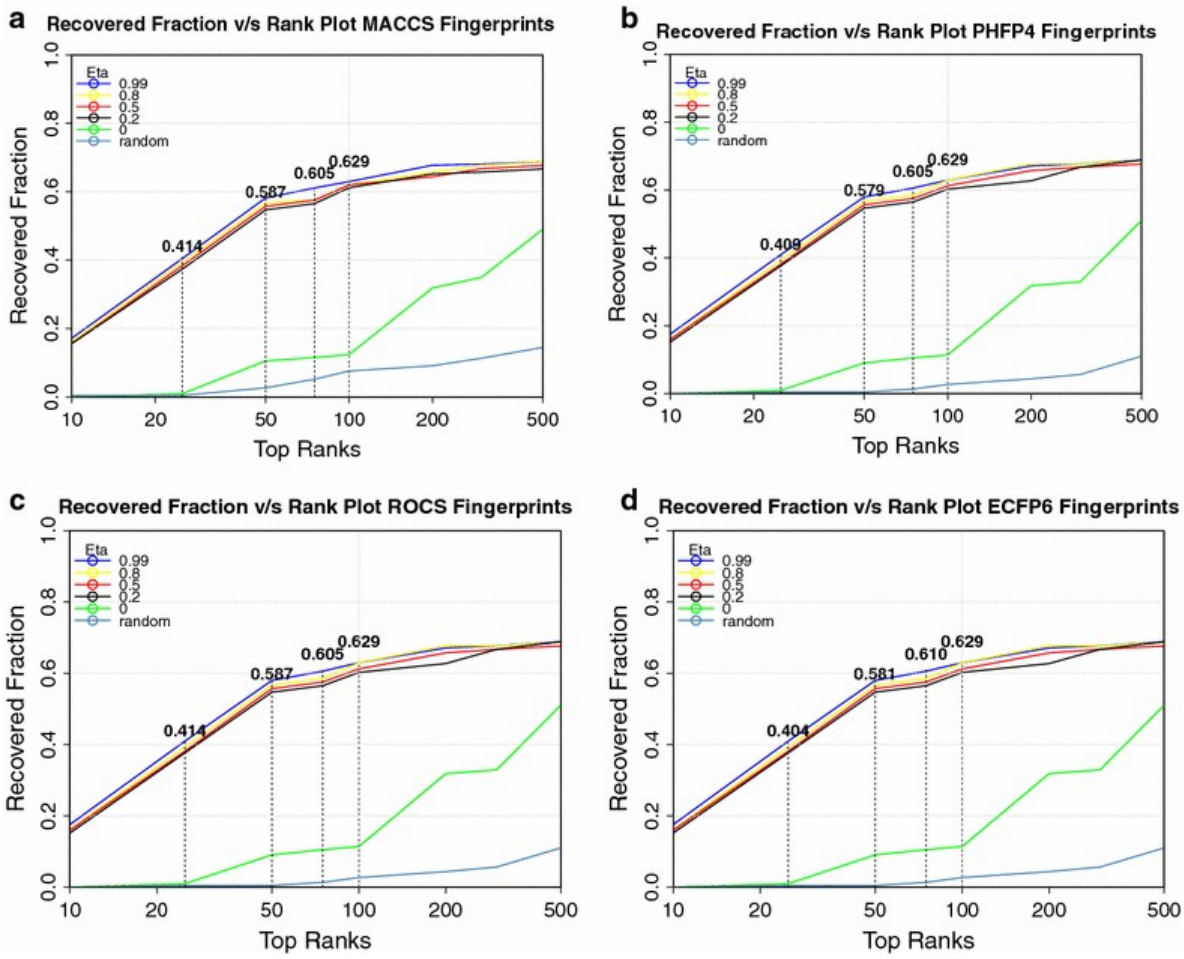

Figure 2.3: Showing the recovered fractions against the rank with different η (eta) values for ChEMBL datat at 1 μM cutoff.

From Tables 2.3 and 2.4 we observe that RWR performance is better for 1 μM target than 10 μM because at 10 μM we have lots off targets from different classes and as a result of that the prediction rate falls. For ChEMBL 1 μM dataset, drugs having more than 0, 1 and 2 targets we achieve BEDROC score of 0.433, 0.553 and 0.611, respectively, which is much better than a random set of interactions. To test whether random walk performs better than just a simple sequence similarity search we took the approved drugs and it's known targets from the ChEMBL 10

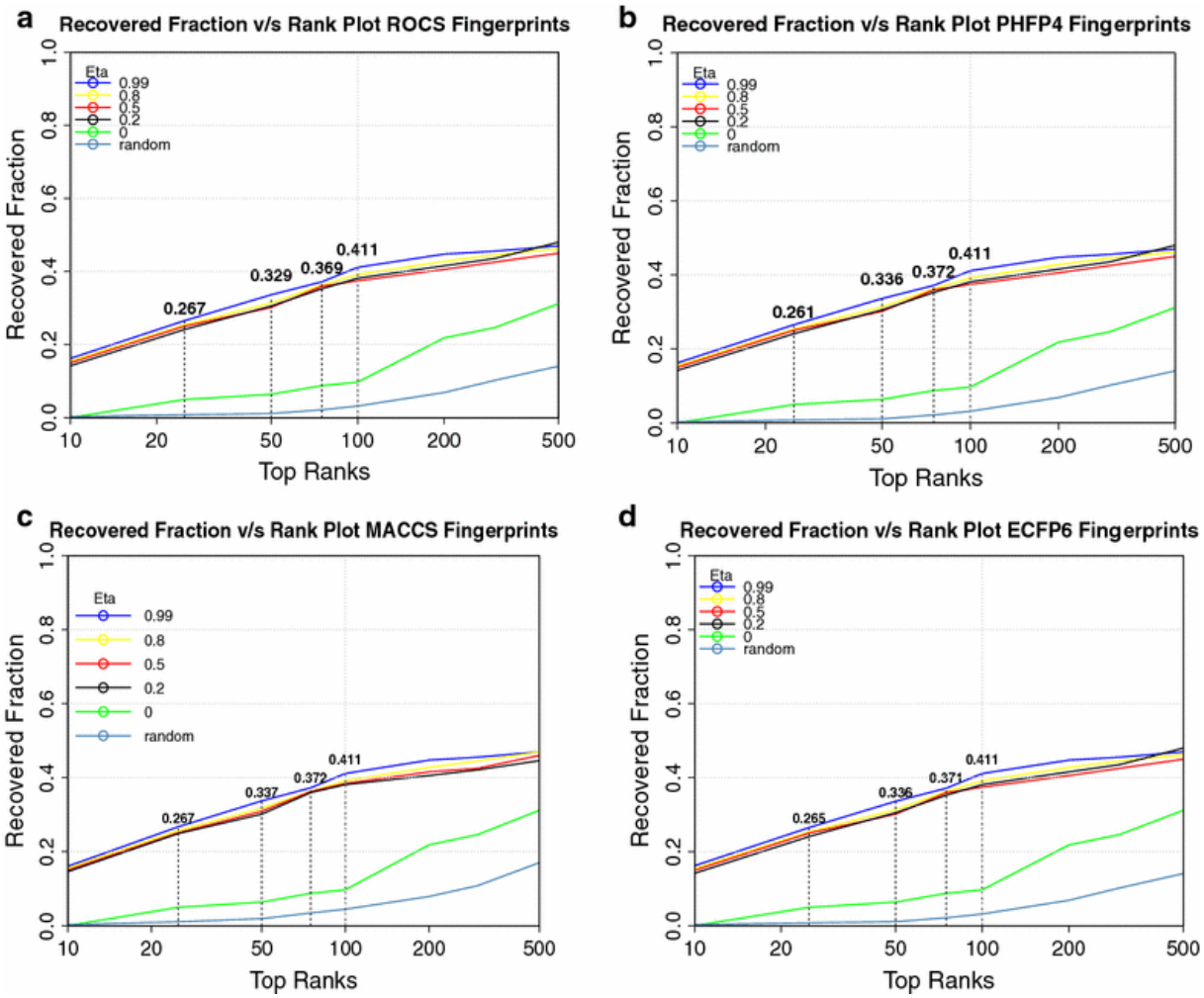

Figure 2.4 Showing the recovered fractions against the rank with different η (eta)

values for ChEMBL datat at 10 μM cutoff.

μM dataset and performed sequence similarity based such against 3,519 targets and ranked them. We found RWR performance is way better in ranking targets than performing simple sequence based search. The results are shown on Tables 3 and 4. This is the first time that the random walk-based method is evaluated using a binding assay dataset.

Data types	Number of targets	AUAC	AUC	BEDROC	EF	AUC(Top
ChEMBL 1 uM (RWR)	> 0	0.709	0.995	0.433	5.058	0.455
ChEMBL 1 uM (Seq)	> 0	0.67	0.67	0.396	4.48	0.414
ChEMBL 1 uM (random RWR)	> 0	0.494	0.493	0.075	1.09	0.079
ChEMBL 10 uM (RWR)	> 0	0.596	0.837	0.323	3.865	0.351
ChEMBL 10 uM (Seq)	> 0	0.518	0.518	0.237	2.641	0.2555
ChEMBL 10 uM (random RWR)	> 0	0.394	0.364	0.036	0.954	0.029
ChEMBL 1 uM (RWR)	> 1	0.784	0.784	0.553	6.286	0.569
ChEMBL 1 uM (Seq)	> 1	0.652	0.651	0.39	4.507	0.412
ChEMBL 1 uM (random RWR)	> 1	0.483	0.483	0.081	1.29	0.083
ChEMBL 10 uM(RWR)	> 1	0.613	0.61	0.353	4.091	0.378
ChEMBL 10 uM (Seq)	> 1	0.551	0.552	0.279	3.084	0.3
ChEMBL 10 uM (random RWR)	> 1	0.514	0.514	0.075	1.244	0.088
ChEMBL 1 uM(RWR)	> 2	0.823	0.824	0.611	6.866	0.631
ChEMBL 1 uM (Seq)	> 2	0.701	0.705	0.513	5.109	0.469
ChEMBL 1 uM	> 2	0.533	0.533	0.0671	1.465	0.065
ChEMBL 10 uM(RWR)	> 2	0.632	0.633	0.399	4.569	0.422
ChEMBL 10 uM (Seq)	> 2	0.569	0.569	0.298	3.03	0.315
ChEMBL 10 uM (random RWR)	> 2	0.521	0.521	0.262	1.95	0.125

Table 2.3: Shows the types of data we used the drug target interaction having more than 1 and 2 drug interactions.

2.4 CASE STUDY: PROFILING TOP SELLING DRUGS

Here, as a case study we investigate the target profiles of the popular top selling drugs in 2012 [54]. First, we consider u_∞, the steady-state probability vector for the targets in our framework, as 'target profile' of a drug. Then we examine the top 10 predicted targets for the top selling drugs. We find that some targets are associated with many drugs (see Table 2.5). For instance, adrenoceptor alpha 1A appears in 60% of drug's top 10 target association lists; serotonin receptor 5HT2A appear in 43%; and adrenoceptor alpha 1B in 35%. Most drugs shown on the

Table 2.5 mostly belong to the rhodopsin class of GPCR's. In Additional file 4, (http://www.jcheminf.com/content/7/1/40#sec5) predictions are provided for 110 drugs with 3,519 targets and Fig. 2.5 shows a bipartite network of 110 drugs with top 10 predicted targets for each drug. We took some random drugs and tried to find known binding associations to protein targets. We searched three databases ChEMBL [55], PDSP [56], and Pubchem [57] using the binding coefficients like IC50 and Ki. Table 2.6 lists the 10 predicted drug–target associations that we have identified evidence of binding interaction in other databases. These findings suggest that these targets may have many undiscovered interactions with existing drugs. Further investigation may have significant values on understanding side effects of existing drugs as well as repurposing them.

Data types	Number of targets	Top 10	Top 25	Top 50	Top 100	Top 200
ChEMBL 1 uM (RWR)	> 0	0.144	0.342	0.47	0.532	0.607
ChEMBL 1 uM (Seq)	> 0	0.164	0.315	0.394	0.42	0.43
ChEMBL1uM(random RWR)	> 0	0.002	0.013	0.018	0.036	0.021
ChEMBL 10 uM (RWR)	> 0	0.11	0.247	0.324	0.386	0.409
ChEMBL 10 uM (Seq)	> 0	0.122	0.183	0.234	0.249	0.254
ChEMBL 10 M (random RWR)	> 0	0.014	0.023	0.035	0.048	0.079
ChEMBL 1 uM (RWR)	> 1	0.274	0.477	0.55	0.58	0.614
ChEMBL 1 uM (seq)	> 1	0.189	0.35	0.428	0.472	0.513
ChEMBL 1 uM ((random RWR)	> 1	0.007	0.023	0.038	0.076	0.091
ChEMBL10uM (RWR)	> 1	0.22	0.277	0.348	0.417	0.446
ChEMBL 10 uM (seq)	> 1	0.13	0.212	0.276	0.296	0.302
ChEMBL 10 uM (Random RWR)	> 1	0.014	0.023	0.035	0.048	0.079
ChEMBL 1 uM (RWR)	> 2	0.271	0.518	0.598	0.634	0.677
ChEMBL 1 uM (seq)	> 2	0.19	0.393	0.53	0.56	0.598
ChEMBL 1 uM	> 2	0.006	0.018	0.034	0.055	0.08
ChEMBL 10 uM (RWR)	> 2	0.233	0.297	0.353	0.4299	0.472
ChEMBL 10 uM (seq)	> 2	0.13	0.22	0.295	0.316	0.324
ChEMBL 10 uM ((Random RWR)	> 2	0.012	0.028	0.04	0.057	0.093

Table 2.4: Table shows the hit rate for drugs having more than 1 and 2 drug Interactions

Targets	% of drugs associated with the targets	% of drug associations appearing in prediction.
ADA1A	7.27%	60%
5HT2A	4.54	43.63%
ADA1B	7.27%	35.45%
5HT1A	4.54%	33.63%
ADRB1	5.45%	31.81%
5HT1B	5.45%	30.90%
5HT2C	3.63%	30%
ACM2	9.09%	26.36%
5HT3A	4.54%	25.45%
5HT1D	5.45%	23.63%
ACM3	9.09%	21.81%
5HT7R	4.54%	18.18%

Table 2.5: Table shows the hit rate for drugs having more than 1 and 2 drug interactions well as repurposing them.

Finally, let us summarize the contributions of this paper. First, we offer a general approach that takes the whole drug target network into account without separating protein categories, in contrast to the previous study [33]. The following estimation corroborates our approach. Our drug-target dataset contains 727 drugs and 3,519 proteins. The number of interactions between drugs and targets is 2,557, which makes 684 drugs to have at least one known target and 457 drugs to have two or more interactions. The proteins in the dataset are grouped under 15 different categories according to ChEMBL target classifications (https://www.ebi.ac.uk/chembl /target/browser). Out of 3,519 proteins, 1,386 proteins belong to one of the categories and other proteins do not have category information. The number of drugs that have at least two interactions with proteins that are categorized is 412. Among these 412 drugs, the number of drugs that have interactions with proteins from multiple groups is 169. In other words, we estimate that about 40% of drugs have interactions across multiple groups according ChEMBL dataset. Therefore, it is more reasonable to consider all proteins together, rather than running the prediction model separately for each category.

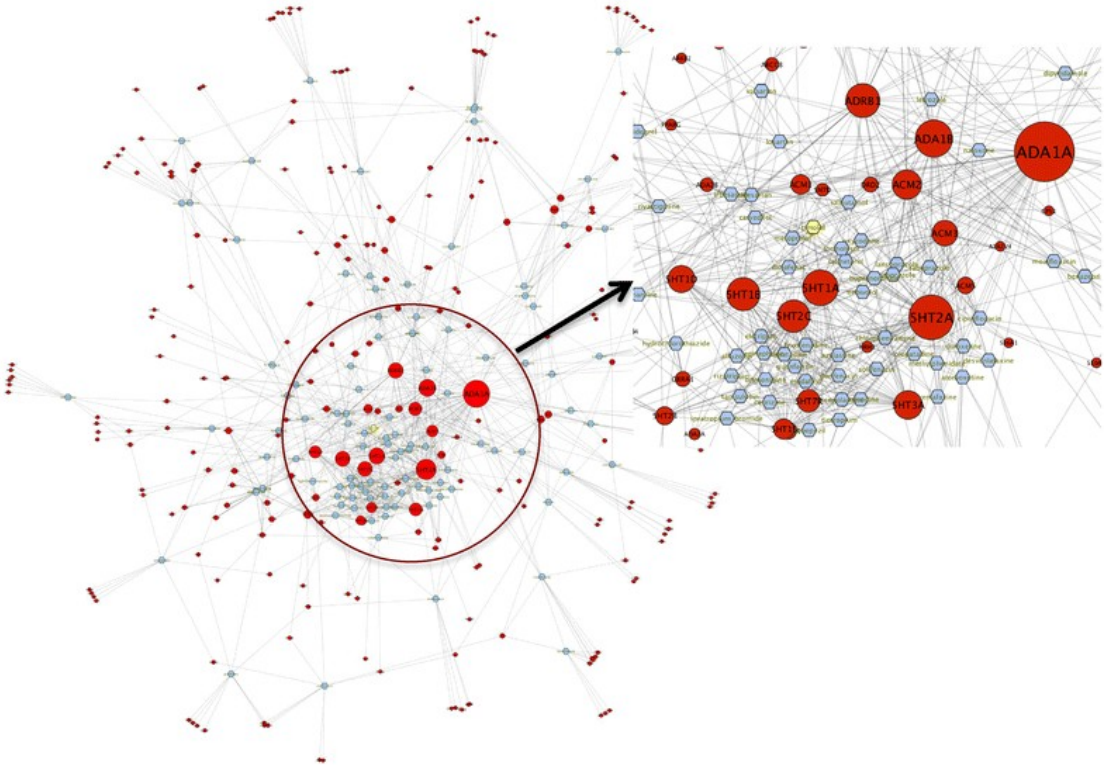

Figure 2.5: Shows the network of the top 10 predicted targets of 110 drugs

Protein	Drug	Source	Activity Tpe	Activity M
5HT2A	Carvedilol	Pubchem AID 625192	IC50	0.41
5HT2A	Desloratadine	Pubchem AID 625192	IC50	0.033
KCNH2	Lidocaine	ChEMBL	IC50	263.02
ADRB1	Salmetorol	ChEMBL	IC50	0.501
5HT1A	Amphetamine	PDSP database	Ki	6.6
HDAC2	Atorvastatin	ChEMBL	IC50	22.5
ADA1A	Duloxentine	PDSP	Ki	10
ACM1	Montelukast	Pubchem AID 625153	IC50	8.045
SC6A4	Quetiapine	PDSP	Ki	10

Table 2.6: Drug target interactions with association values from different Databases

Second, we further investigate the methodology by presenting a benchmark of a parameter η in conjunction with the four chemical fingerprint types: MACCS 166 keys, ECFP6 fingerprints, PHFP4 fingerprints, and ROCS. In the previous study, the parameter space of η is not explored below 0.1, but we find that we can improve the performance by de- creasing η below 0.1. We also find that the performance is robust under the choice of chemical fingerprinting method, particularly when η is around the optimum (0.01). Very small η means the walk in the target network is much more important than the walk on the drug–drug network. In a sense, it indicates that drug network add some information but only marginally. And also the drug network is not very useful in prioritizing targets.

2.5 CONCLUSION

We have demonstrated that RWR approach provides a powerful way of predicting of drug–target interactions. There are two significant benefits of the approach. First, it provides a natural way to integrate multiple types of information such as drug–drug similarity, target–target similarity, and existing drug–target interactions into a coherent framework. Second, in contrast to other approaches like short-path-based methods, the random walk framework incorporates the network structure around a single or multiple points of interests extensively, taking into account not only the closeness of targets, but also the multitude of the paths to the targets. These properties allow us to predict novel targets even for the drugs that have no known target, by

connecting such drugs to the network through the drug–drug similarity. Still, the performance of RWR could be further improved by incorporating more known drug–target interactions. We have studied the performance of the method under the variations of η parameter and the choice of fingerprints methods, showing that while training the model one can use any of the chemical features as similarity matrix with parameter η=0.01 to obtain the predicted results, without significantly affecting the outcomes.

CHAPTER 3

SYSTEMIC IDENTIFICATION OF DISEASE ASSOCIATED PATHWAYS BY RANDOM WALK WITH RESTART

3.1 INTRODUCTION

One of the fundamental challenges in human health in elucidating the molecular basis of hereditary diseases. Pioneering studies by Goh [58, 59] have resulted in the definition of Human Disease Network (HDN) which helped relate these diseases through shared genes, shared proteins, regulatory proteins, shared pathways and similar gene expression profile. Recent advances in genomics, molecular and cell biology, biochemistry have allowed us to visualize the organization of disease complexity at multiple scales involving - molecular data, proteins, tissues, pathways and organ level data. Understanding the relationships between different scales of organization will allows us to study drug interactions and its effects at molecular levels and organismal effects [60]. The biological function or the pharmacological effect can be studied based on an organ tissue based system where proteins are interacting with each other by non covalent interactions and forming complexes which takes part in biological pathways. If genetic variants occur in these protein complexes it can alter the function of entire complex and may alter the normal pathway to cause a disease. A pathway consists of series of related reactions, whereby the reactions are linked through common compounds (metabolic pathways) or through the common

macromolecular complexes (protein-protein interaction pathways). To understand the molecular basis of disease it is important to understand the biological pathways and how the proteins are coordinating among themselves in a network of protein-protein interactions. However, the interactions are condition and tissue specific i.e. for a particular tissue and disease some proteins are differentially expressed and have varied interactions among the proteins to activate various pathways [61].

It is common for the diseases to get associated with a specific tissue types. For example for metabolism of drugs, genes like Cytochrome P450 would differentially express in liver than in brain, thyroid, tonsil or skin. Similarly, genes for Parkinson's disease would differentially expressed in brain rather than in liver or rectum. The idea here is to find associations of tissue specific disease pathways. This would empower us to select pathways where the disease genes are cooperating among themselves. Biological pathways represent biological reactions and its interaction network within a cell. Each reaction is identified by an enzyme, which is coded by a gene. Various Gene prioritization algorithms use protein interaction networks, ontologies, gene expression data to prioritize candidate genes for diseases [28, 62, 63]. However using only protein-protein interaction data or gene expression data it is not possible to detect the molecular basis of the disease and how symptoms are raised. It is very important to map expression of genes to the tissues and the tissues to diseases [64]. But to understand the molecular basis of disease we should emphasize in understanding the biological pathways and how the proteins are coordinating among

themselves inside a specific tissue.

In the past methods were developed to map protein complexes to specific diseases and then use such information to facilitate prediction of disease genes. For example, Lage etal. [65] Identified aggregates of proteins connected to candidate protein in a PPI network as a protein complex infer association between the candidate protein and a query disease based on members of the protein complexes [66]. Vanunu etal. [67] proposed a random walk method where the edge weights are normalized by degree of the targets to prioritize protein complexes associated with the disease. Magger etal. [30] used 60 different tissue specific networks for finding disease related genes and found high precision in finding disease related genes. Tissue-specific networks can reflect the related diseases better than using normal global ppi networks. Most of them worked with disease – disease network, ppi network and a bipartite disease – protein network. None of them extended their algorithms beyond 2 layers of networks. This work attempts to extend the network layers to four layers namely disease, proteins, protein complexes and biological pathways.

In this chapter we propose an approach for identification of biological pathways that are related to a query disease via a random walk model on a large heterogeneous network that is composed of disease-disease similarity layer, a tissue based protein-protein interaction layer, a protein complex layer and finally complex and pathway membership layer. Association between protein complexes and pathways are made based on interdependency measure between complex-proteins and protein-pathways, and is described in detail in methods section. Starting from

the query at the disease layer the random walker travels in a four layered network and scores a biological pathway using the probability that the walker stays in the pathway layer at steady state and ranks the pathways according to the probability scores. We validate our method by cross validation technique in which from a query disease we removed all the associated proteins and tried to predict whether the given disease-pathway relation can be predicted using our method.

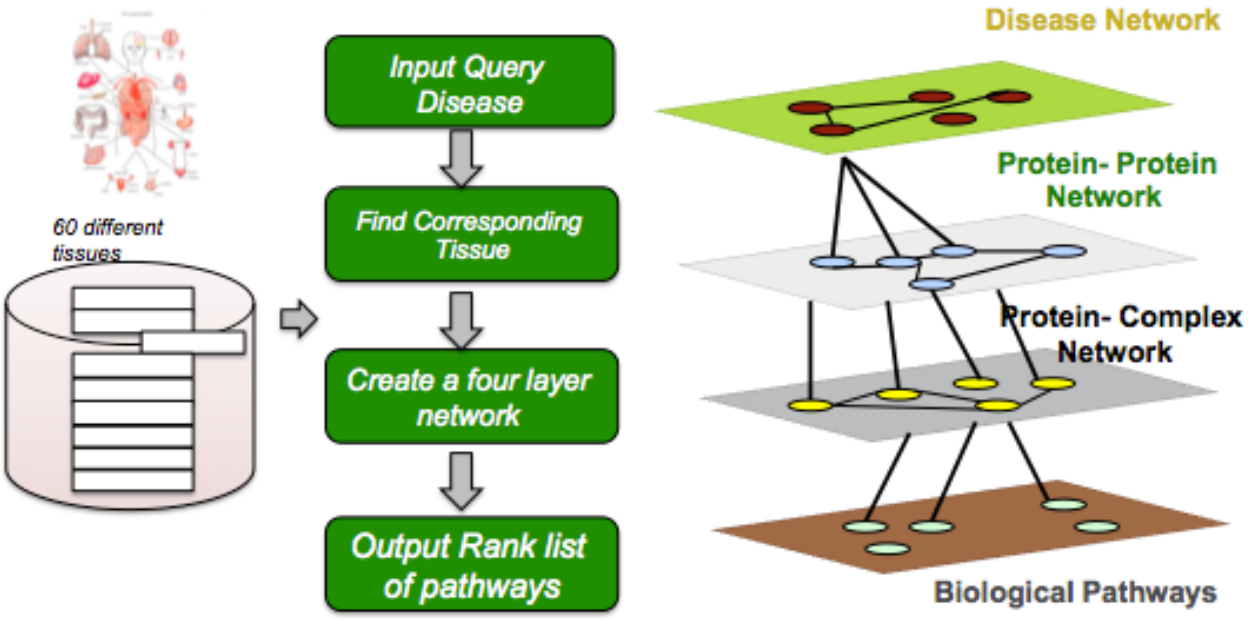

Figure 3.1: Diagram showing the workflow of disease to pathway prediction.

3.2 METHODS

3.2.1 DATASETS AND PRE-PROCESSING

In this chapter we are dealing with four different types datasets namely diseases, protein- protein interactions, protein complexes and pathway information. We describe below how we use them in our model generation. For our work, we used 926 diseases and 60 tissue associations (DisT), which is provided by Jacquemin etal. [68] . Van Driel etal [69] used MeSH vocabulary to create a weighted vector composed of phenotype terms and quantified the similarity between the diseases based on the cosine similarity scores for 5080 diseases. We got the tissue specific PPI network data from Magger etal [30]. The PPI data consist of 9998 proteins with 41,049 interactions. Magger used two types of networks one generated using the edge reweighted strategy and other node removal strategy [70]. In this chapter we used edge reweighted PPI network rather than the node removed network, because node removed network is a very strict method of eliminating unexpressed proteins, in which nodes are removed from the network if the proteins are not expressed in the relevant tissue. This changes the topological property of the network consisting of multiple components. Edge reweight method assigns a continuous value for the interaction based on the expression of the two interacting proteins. It uses a penalty factor, which is multiplied to the original PPI network such that when it is 0 then we have node-removed network and when it is 1 we have

the full original PPI network. Magger used a network with penalty factor 0.1 which achieved maximum precision in identifying disease genes and retaining its full network topology. Since the Node removed network is likely to be less robust to noisy data such as gene expression and hence we selected the edge reweighted PPI network in the current study. In order to see, weather using tissue based PPI network increases precision of disease pathways associations we created a non-tissue based PPI network by calculating the average of the edge weights of 60 different tissue based PPI networks. We extracted disease-protein associations using Biomart tool [71], obtaining a total of 6,015 associations between 4,085 diseases and 3,418 proteins. We mapped the disease and proteins ids to the OMIM and PPI network proteins, which resulted in 1,670 diseases, and 1,338 proteins with 2,091 associations in all. We used protein complex information from the CORUM database [72] accessed (January 4th 2015) and extracted 1826 human protein complexes, which has complex names and one of proteins in the complex can be mapped to PPI network of proteins. We made a protein complex matrix containing 8268 binary associations. For the complex pathway relationship we downloaded all the pathways information from ConsensusPathDB database [73]. This database not only integrates information from KEGG [74] but also from several other resources like PharmGKB [75], SMPDB [76], Reactome [77], Wikipathways [78], HumanCyc [79], Biocarta [80], Netpath [81], and EHMN [82]. I had 2145 proteins ids linked with 2531 pathways with 45,669 associations. For Complex pathway association we used an interdependency measure between the protein–complex and protein–pathways and associations

between them that are significantly interdependent were identified and kept in the complex pathway association contingency table. To compute the pathway and protein complex associations we first create a contingency table, which is shown in table 3.1 of M rows and N columns. In the table, O_{ij} denotes the number of occurrences of proteins that are shared by complexes and pathways.

	$pathway_1$	$pathway_2$	$pathway_3$	....	$pathway_j$
$Pcomplex_1$	O_{11}	O_{12}	O_{13}	...	O_{1j}
$Pcomplex_2$	O_{21}	O_{22}	O_{23}	...	O_{2j}
$Pcomplex_3$	O_{31}	O_{32}	O_{33}	...	O_{31}
$Pcomplex_4$	O_{41}	O_{42}	O_{43}	...	O_{4j}
$Pcomplex_i$	O_{i1}	O_{i2}	O_{i3}	...	O_{ij}

Table 3.1: Protein Complexes and Pathway occurrence matrix

Let , $exp_{ij} = \frac{O_{i+}\, O_{+j}}{T}$ be the expected number of occurrence's of O_{ij} ,where $O_{i+} = \sum_{k=1}^{N} O_{ik}$ and $O_{+j} = \sum_{k=1}^{M} O_{kj}$ and $T = \sum_{l,k} O_{lk}$. An interdependency relationship is considered to exist if O_{ij} is significantly different from exp_{ij}. To calculate the significance we calculate an adjusted residual test statistic as discussed in Jong et al. [83] given below,

$$ad_{ij} = \frac{z_{ij}}{\sqrt{(1-\frac{O_{i+}}{T})}\sqrt{(1-\frac{O_{+j}}{T})}} \qquad (3.1)$$

where,

$$Z_{ij} = \frac{O_{ij} - exp_{ij}}{\sqrt{exp_{ij}}} \quad (3.2)$$

and $\sqrt{(1 - \frac{O_{i+}}{T})}\sqrt{(1 - \frac{O_{+j}}{T})}$ is the maximal likelihood of Z_{ij} .

ad_{ij} has an approximate normal distribution of with mean zero and variance of approximately one. So if the absolute value exceeds 1.96 then it would be considered significant at alpha = 0.05. Based on equation (3.1) we can compute the interdependency relationship between complex and pathways and prepare a adjusted contingency table. There we convert all the absolute values < 1.96 to 0 in the current adjusted contingency tables and create the final complex-pathway association matrix.

3.2.2 OVERVIEW OF THE RWR METHOD

We modeled pathways associated with the diseases as a random walk based prioritization method [84–87], in which given a query disease and a set of predefined pathways as seeds, we first identify the tissue to which the disease is most likely related and then get the associated network for that tissue. Once this step is done then we construct a tissue–specific disease-protein-complex–pathway network, which is a heterogeneous network composed of 4 layers. Then I apply the random walk with restart algorithm to this network to calculate the score for each pathway and rank the candidate pathways. The network I constructed consist of four different layers the top layer consisting of disease similarity $(DDij)_{lxl}$,

where l is number of diseases We used two different methods to create similarity network first, with a K-nearest neighbor (KNN) strategy where we use 15, 20, 25, 30 nearest neighbors to build 4 different types KNN disease networks. Second, with a threshold cutoff of 0.3, 0.4, and 0.5 on the disease similarity network we build three types of threshold networks. In both the strategy we further consider two variations to keep the weights as unweighted. In both the cases we choose normalize the edge weights by degree of the nodes. For that we define a diagonal matrix L such that $L(i, i)$ is a sum of row of i of similarity matrix S we set $S' = L^{-1/2} S L^{-1/2}$ which gives us a symmetric matrix $S'_{ij} = \frac{S_{ij}}{\sqrt{D(i,i)D(j,j)}}$. S'_{ij} is also known as the normalized laplacian.

We connect the disease layer to the protein-protein network layer $(PP_{ij})_{mxm}$ using the disease–protein association matrix $(DP_{ij.})_{lxm}$ where, m is total number of proteins. For each of the query disease we load the corresponding PPI network and normalize it by node degree.

The next we connect the proteins from the PPI network to the protein complexes by using undirected edges to form a matrix $(PC_{ij})_{mxc}$ where , c is the total number of complexes. We didn't connect complexes between themselves and we normalize the adjacency matrix between complexes and proteins from the PPI network based on its degree of the nodes.

The last and the bottom layer we connect the complexes to the pathways using weighted and unweighted edges to form a matrix $(CPa_{ij})_{cxp}$ where , p is the total number of proteins. Also we left the pathways unconnected in the study. . If we put all the matrices together we get a large transition matrix $\boldsymbol{W}$ of 19435 x 19435 elements given below as ,

$$\boldsymbol{W} = \begin{pmatrix} DD & DP & DC & DPa \\ DP^T & PP & PC & PPa \\ DC^T & PC^T & CC & CPa \\ DPa^T & PPa^T & CPa^T & PaPa \end{pmatrix}$$

This matrix can be written as ,

$$\boldsymbol{W} = \begin{pmatrix} DD & DP & 0 & 0 \\ DP^T & PP & PC & 0 \\ 0 & PC^T & 0 & CPa \\ 0 & 0 & CPa^T & 0 \end{pmatrix}$$

In the transition matrix W, 0 stands for a zero matrix indicating no transition between the nodes and superscript T stands for the transposition of the matrix.

We used two types of normalization for the full network,

- One is the column normalized where the edges are normalized by column sums,
- Laplacian based where the edge weights are normalized by source and target degrees,

We used two types of datasets,

- 60 Tissue based PPI networks,
- global PPI network where we use the average of the confidence scores of 60 different tissues and create a global network to predict the associations.

In total I used three type of datasets in order to check performance of the method and parameters,

- Laplace normalized Tissue non-specific network
- Laplace normalized Tissue based PPI network
- Column normalized Tissue based Network.

3.2.2 RANDOM WALK PROCESS:

The user gives the source disease and target pathway as the seed nodes. We initialize a query vector of seed nodes $(P_0)_{l+m+c+p}$ which represents the prior probabilities when a random walker starts its journey. In this vector all the seed nodes are initialized to 1 and the remaining ones to 0, we then normalize the query vector.

$$p_0 = \begin{bmatrix} (1-\gamma)u_0 \\ 0 \\ 0 \\ \gamma v_0 \end{bmatrix}$$

u_0 and v_0 be the initial probability vectors for disease network and pathway network and we initialize the protein and complex vectors as 0. Parameter γ controls the importance of two kinds of seed nodes, i.e. disease node and pathways node. We tested the importance parameter γ for different values ranging from 0.5, 0.7 and 0.9. We use $(P_t)_{l+m+c+p}$ to represent the probabilities that the random walker stays on the nodes at time steps t. After a number of iteration steps, the P_t converges to a steady-state probability vector $P\infty$, where we represent the probability $P\infty$, the probability of finding the random walker in the steady state, which can be determined by change between P_t and P_{t+1} ($L1$ norm) is less than 10^{-7} a random walker chooses the query of interest and at each time step of the walking process the walker may start a new journey with probability c or may move to its neighbor's with probability $1 - c$ according the transition value matrix W . The random walk is implemented on the heterogeneous network using the equation given below,

$$P_{t+1} = (1 - c) W^T P_t + cP_0 \qquad (4.3)$$

After the steady state is achieved we normalize the scores and further rank the candidates in a decreasing order of the probability scores. As a baseline we also made randomized networks for each of the type of parameters and tested the performance as well as the significance.

3.3 RESULTS AND DISCUSSION

3.3.1 EVALUATING PREDICTION PERFORMANCE USING LINK PERTURBATION:

The network based method aims to predict new pathways for a given disease. To evaluate our method we created a test set consisting of associated diseases, tissues, proteins and pathways and a set of control objects as those that neither link to the disease and protein in the training data not in the test data. Then we perturbed the network were we remove all the links between a disease and its associated proteins and calculate discriminant scores for both the test and the control objects, and we rank each test object against all control objects in non-ascending order according to their proximity scores. Repeating the above ranking procedure for all test cases, we obtain a set of ranking lists and further calculate some accuracy measure like auc, auc top 10%, bedroc [53] and enrichment factor.

The area under the receiver operating characteristic curve (AUC) is widely used to evaluate the performance of the ranking method. The advantage is, the value ranges from 0 to 1 with 0.5 corresponding to randomness. AUC has been criticized as an inappropriate method and is not sensitive to early recognition [52,88,89].

A key criterion for measuring the success of ranking prediction is the enrichment of annotated associations among top ranking associations. The higher the percentage of annotated associations among the top ranking associations, the better the performance of the prediction. The enrichment criterion is evaluated by a numerical factor (EF) defined as,

$$EFset = (Ha/Ht)/(A/D) \quad (3.4)$$

where , Ht total number of links retrieved, Ha is the total number of true links retrieved in the links list. A represents the total number of true links in the database and D stands for total number of interactions both positive and negative links. The enrichment factor reflects the capability of a screening application to detect true links (true positives) compared to random selection. Thus, its value should always be greater than 1 and the higher it is, the better the enrichment performance of the virtual screening. The EF overcomes this problem but it is dependent on the ratio of true links to non-links and the choice of X (ratio of the top ranked links). Similarly when we are predicting links it should rank true links in the top-ranking list. Metric likes ROC not sensitive to early recognition for example considering cases like where,

- true links are retrieved at beginning of a rank ordered list,
- where true links are randomly distributed

- where true links, which are retrieved in the middle of the rank, ordered list

In all of the above cases ROC is 0.5 but in terms of early recognition we see that case (1) is better than (2) and (3). To overcome these limitations methods such as RIE [88] and BEDROC [52] have been proposed. By changing the tuning parameter,α, one can test whether the method is able to rank true links early or not. In order to check performance we used two metric known as the relative rank and precision. For calculation of relative rank, we calculate the relative rank of all the predicted links for a particular disease by taking the ranks and dividing it by N (total number of true links) and then we calculate the average of the ranks for all the predicted pathway links for a queried disease. For precision criterion, we calculated the precision within the cutoff of top rank 100. In our test set a disease is associated with more than one pathway and proteins, for a particular disease we randomly select one of the pathway as the seed node. For a queried disease we remove all the disease – proteins links from the DP matrix, recalculate the transition value matrix in order to predict other associated pathways for the queried disease. As we remove all of the disease proteins links from the queried disease now the random walker will depend on nearest associated diseases and its associated proteins to make the walk and predict the pathways. We tested the parameters γ with different KNN and thresholds on the disease similarity network. We performed leave-one-out cross-validation experiment using this network.

3.3.2 COMPARISON OF DIFFERENT STRATEGY FOR CONSTRUCTING THE NETWORK

We considered two different strategies for creating the disease similarity network as the top layer of the disease-protein-complex-pathway network: one K nearest neighbor (KNN) and other is δ-threshold strategy. Also we considered two strategy to construct the protein- protein interaction network: one used edge reweighted network which consists of network from sixty different tissues and the other one we took the average of all the networks from all tissue related network and considered a tissue non-specific for prediction. We also checked the performance of the methods with different parameters and with different normalizations with a random network as a baseline. Table 1 and Table 2 represent the results of the KNN strategy and δ-threshold strategy respectively. We didn't present the full set of results.

γ	Relative Rank	Precision	AUC	AUCTOP (10%)	BEDROC	EF	Data Normalization	KNN
0.5	153.074	0.495	0.738	0.205	0.228	3.262	LaplaceNormalized tissue non-specific	20
0.5	157.564	0.499	0.733	0.202	0.223	3.267	Laplace Normalized Tissue Based	25
0.5	213.128	0.364	0.669	0.167	0.183	2.581	Normalized Tissue based	20
0.7	152.027	0.513	0.742	0.21	0.23	3.333	Laplace Normalized tissue non-specific	25
0.7	158.268	0.478	0.732	0.2	0.222	3.227	Laplace Normalized Tissue Based	30
0.7	158.268	0.38	0.679	0.158	0.162	2.591	Normalized Tissue based	15
0.9	156.937	0.498	0.737	0.202	0.224	3.328	Laplace Normalized tissue non-specific	20
0.9	155.982	0.496	0.735	0.202	0.222	3.265	Laplace Normalized Tissue Based	30
0.9	210.23	0.36	0.675	0.158	0.177	2.535	Normalized Tissue based	15

Table 3.2: Results of the K-nearest neighbor (KNN) based disease network using the parameters on table 3.2 and table 3.3 however, we shown the

KNN results and δ-threshold using different γ parameter values with best KNN value from different tissue based PPI networks and two types of normalizations.

We observe that our method is quite robust to the number of neighboring diseases in the KNN-strategy and well as the γ-threshold strategy for both the Laplace normalized tissue specific network and as well as the Laplace normalized tissue non-specific network. For KNN-strategy we have achieved a mean relative rank, AUC, AUCTOP, BEDROC, EF of 152.027, 0.742, 0.210, 0.230, 3.33 with Laplace normalization, using a tissue non- specific PPI network, KNN of 25 and γ of 0.7. We also found close scores for Laplace normalized tissue based PPI network with Relative Rank, AUC, AUCTOP, BEDROC, EF of 157.564, 0.733, 0.202, 0.223, 3.267 with KNN of 25 and γ of 0.5. We observe that for a column normalized tissue based PPI the Relative Rank, AUC, AUCTOP, BEDROC, EF is well below the laplacian Normalized ones which illustrate that Laplacian normalized increases the prediction performance. We noticed that for δ-threshold strategy we get a better performing model than the KNN-strategy. For δ-threshold of 0.3, we have achieved a mean relative rank, Precision, AUC, AUCTOP, BEDROC, EF of 149.012, 0.521, 0.747, 0.220, 0.240, 3.494 using Laplacian normalized tissue based PPI network with γ of 0.9. The results obtained are almost same as γ of 0.9 and δ-threshold of 0.3. Around these parameters the δ-threshold strategy holds the higher performance than KNN-strategy. Therefore we recommend to use δ-threshold strategy in disease similarity network along with laplacian normalization of matrices and for PPI we found

Tissue based PPI networks performance is higher than tissue non-specific one.

Figure 3.2 shows the precision plots for both the threshold and KNN strategy compared against each other with different types of datasets along with a randomized dataset. For KNN at different γ parameter values, we found that at KNN = 25 and γ = 0.7 it achieves a precision of 0.517 for top 100 rank pathways. In threshold strategy at different γ parameter values, we found that at threshold = 0.3 and γ = 0.9 it achieves a precision of 0.555 for top 100 rank pathways which confirms us that threshold strategy is suitable for our dataset.

Also as a baseline we also make randomized networks and made predictions and checked the performance for KNN strategy and δ-threshold at γ parameter values. The ROC plots are shown for both KNN and δ-threshold on **figure 3.3** illustrate better predictive performance of laplacian normalized tissue based network against random network.

The restart probability *c* determines the possibility of jumping from any node in the network back to the starting point of the query disease. With a large value of *c*, a random walker cannot go far away from the starting point and thus will mainly explore neighboring nodes of this point, while with a small value of *c*, the random walker is able to explore areas far away from the starting query disease. We computed Relative Rank, AUC, AUCTOP, BEDROC and Enrichment Factor for laplacian normalized tissue based PPI network with different Restarts from 0.1 – 0.9 as shown on table 3.4 we observe that our method shows small variations, however at restart of *c*=0.9, we have best AUC of 0.747, AUC TOP (10%) of 0.22, BEDROC

(α=20) of 0.24 and enrichment factor (EF) of 3.493. From these observations we conclude that the selection of parameter c = 0.9 results in improved performance of our approach as shown in table 3.3 .

Restart (c)	Relative Rank	AUC	AUCTOP	BEDROC	Enrichment
0.1	152.9424	0.721	0.184	0.193	3.173998
0.2	155.0584	0.718	0.185	0.191	3.143087
0.3	165.5083	0.7	0.159	0.167	2.88128
0.4	166.8327	0.697	0.164	0.171	2.729648
0.5	171.7294	0.689	0.162	0.167	2.674881
0.6	159.7915	0.712	0.177	0.191	3.025033
0.7	149.3544	0.74	0.205	0.226	3.244651
0.8	149.0128	0.745	0.213	0.236	3.407035
0.9	148.4792	0.747	0.22	0.24	3.494314

Table 3.3: Results of Laplacian Normalized Tissue based PPI with different restarts values

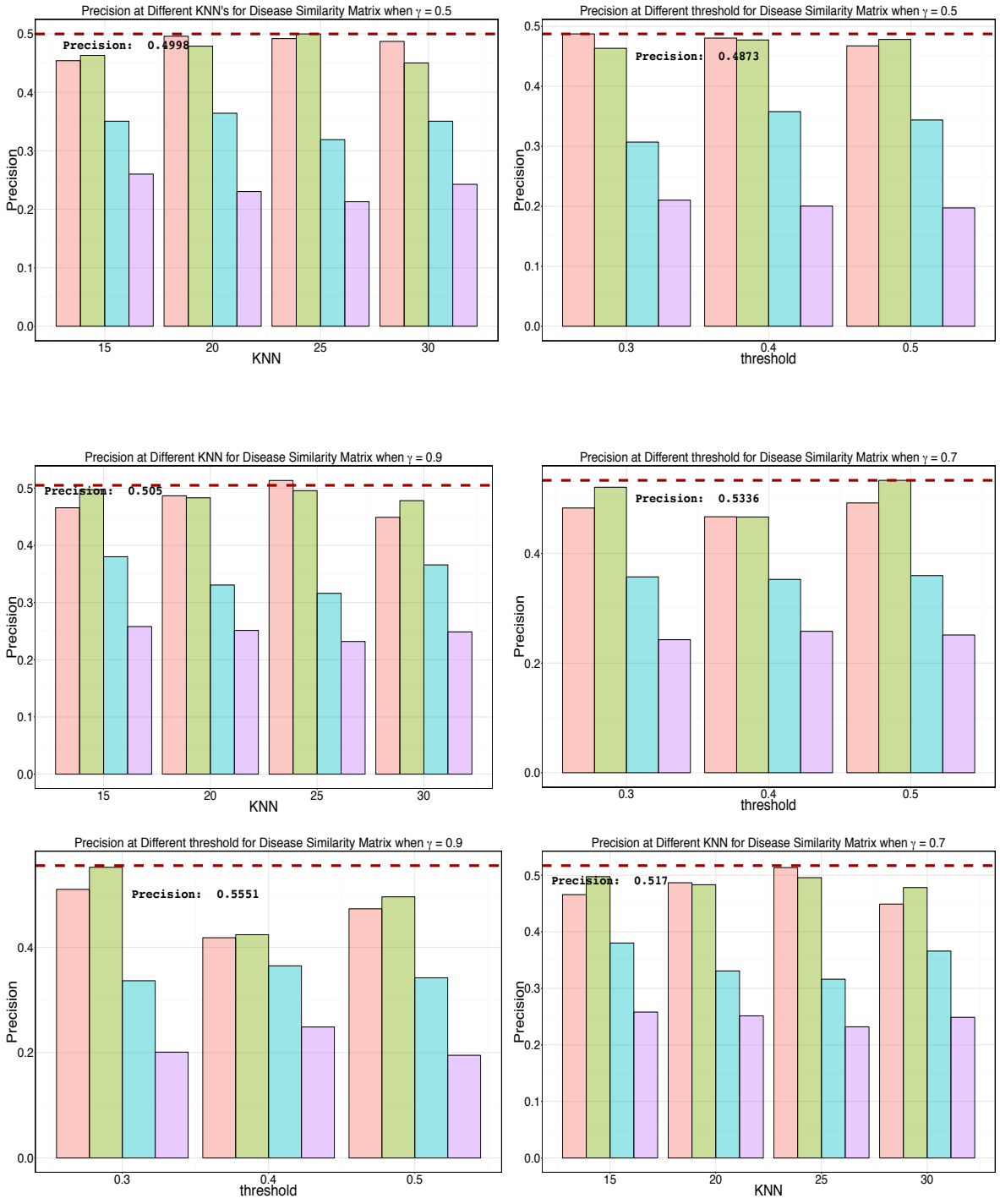

Figure 3.2: Showing precision plots for different γ's with KNN strategy and threshold

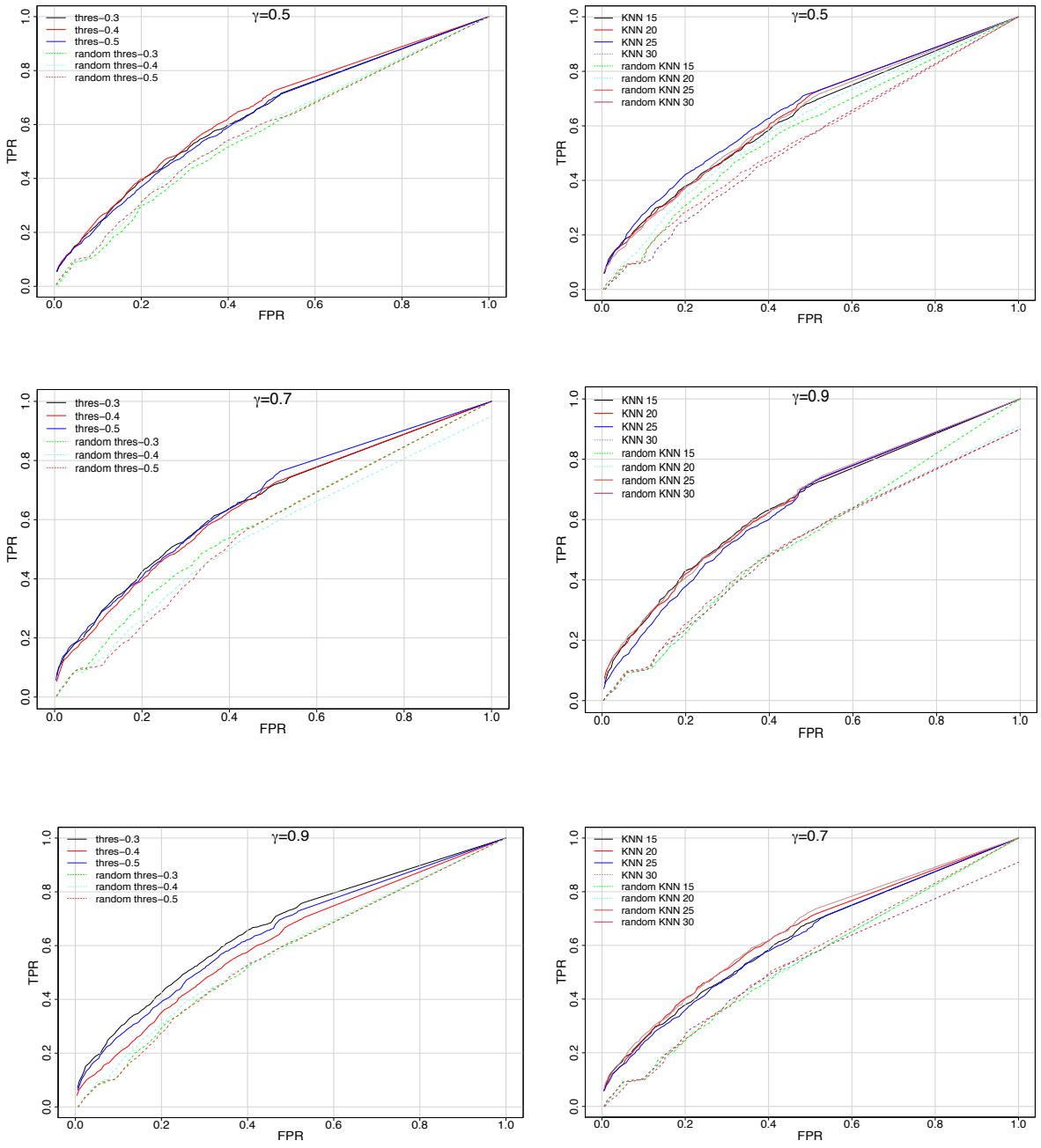

Figure 3.3: Showing ROC plots for different γ's with KNN strategy and threshold strategy.

3.4 CONCLUSION

In this chapter we have proposed a method for identification Biological pathways related to a query disease via random walking on a heterogeneous network that is composed of four different layers the disease layer, protein layer, protein complex layer and the protein pathway layer. We have shown a good performance of our method via large-scale leave out cross validation approach and optimized the parameters c, γ for better results. We tested two different types of disease similarity network types like KNN and threshold based and showed that threshold of 0.3 gave us better performance with restart c of 0.9. We would like to predict the pathways, given a specific disease symptoms like cough, fever, headache etc. Human phenotype ontology (HPO) [90] provides symptoms data for different diseases integrating the data in our system will be a direction worth exploring and user can input symptoms and then it can predict the biological pathways, the proteins which are involved and it can easily identify what type of drugs is suitable for queried symptoms.

CHAPTER 4

NETPREDICTOR R PACKAGE

4.1 INTRODUCTION

Social and biological systems can be represented by graphs where nodes represent individuals, biological experiments (protein, genes, etc.) web users and so on. Networks allow methods of graph theory to be applied to the task of predicting links. Link prediction predicts missing links in networks or links in future networks, it is also important for mining and analyzing the evolution of networks. Link prediction problem is a long-standing challenge in modern information science, and a lot of algorithms based on Markov chains and computer science community has proposed statistical models. The link prediction problem is usually defined in unipartite graphs. The netpredictor package is developed to solve the problem of bipartite link prediction using Random walk with restart (RWR) and network based inference methods (NBI). We plan to integrate variety of other algorithms in near future. All of the code is developed in R, which also provides parallel execution modes.

To compute the R package for prediction of missing links in a bipartite network/graph. The package provides utilities to compute missing links in a bipartite and well as unipartite network based on HeatS, Random walk with Restart (RWR), Network based inference (NBI) and combination of RWR and NBI . The package also allows one to compute Bipartite

network properties and well visualize communities and make a permutations test based prediction. With the advent of R open source statistical programming language [91] and gaining popularity of very useful "Shiny" package [92,93] that lets the programmers to create applications online, a new opportunity has been shown itself for creating netpredictor Shiny web app.

4.2 LINK PREDICTION IN NETWORKS

Link prediction is a new field of research in networks science and was first demonstrated by early 90's by Nowell-Kleinberg [16, 17]. They tried to evaluate set of different similarity measures between vertices of a graph in order to predict unknown edges (links). They are classified into two categories:

- Neighborhood based metrics and
- Path based metrics

In the netpredictor package we developed the methods below for finding missing links in a network.

4.2.1 LINK PREDICTION BY NEIGHBOURHOOD-BASED METRICS

The following are the methods for neighbourhood-based metrics. Let $\Gamma(x)$ be the set of neighbors of node x, and let $|\Gamma(x)|$ be the number of neighbors of node x.

• **Common Neighbors** (CN): CN is defined as the total number of nodes that two nodes x and y have common interaction with. More the number of links more significant the relation. It is defined using,

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)| \qquad (4.1)$$

• **Jaccard Coefficient** (JC) [18]: This is the extension of the common neighbors where it shows a proportion of nodes that are common between nodes x and y among all the nodes between x and y. The value is usually normalized between 0-1. It is defined using,

$$JC(x.y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \qquad (4.2)$$

• **Cosine Similairity** (CS): Cosine metric for two nodes x and y is defined as,

$$CS(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)|.|\Gamma(y)|}} \qquad (4.3)$$

• **Hub promoted Index** (HP) [19] : It defines the topological overlap between nodesx and y. The links adjacent to hubs are assigned high scores since the denominator is of lower degree of nodes. It is defined using the following equation,

$$P(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min(|\Gamma(x)|,|\Gamma(y)|)} \qquad (4.4)$$

• **Hub Depressed Index** (HD): This is similar to HP index but in this case the denominator is determined by the higher degree of the nodes.

$$HD(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max(|\Gamma(x)|,|\Gamma(y)|)} \qquad (4.5)$$

• **Adamic Adar Index** (AA) [20]: This index is similar to counting of common neigh- bors by assigning more weights to lower connected neighbors. It is defined as ,

$$AA(x,y) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log |\Gamma(z)|} \qquad (4.6)$$

• **Preferential Attachment** (PA) [21]: The PA metric indicates the new links will be more likely to connect higher degree nodes than lower ones. Since it does not need to know the neighborhood of each node, so it has minimal computational complexity. It is given using the following equation as,

$$PA(x,y) = |\Gamma(x)| \,.\, |\Gamma(y)| \qquad (4.7)$$

• **Resource Allocation** (RA) [22]: The Resource allocation index is similar to the adamic adar index which assigns more weights to lower connected neighbors but RA performs better with nodes having high average degrees. It defined by,

$$RA(x,y) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{|\Gamma(z)|} \qquad (4.8)$$

• **Leicht-Holme-Nerman Index** (LHN) [23]: This index assigns high similarity to node pairs that have many common neighbours between them where $|\Gamma(x)|.|\Gamma(y)|$ is the expected number of common neighbors. It is defined by ,

$$JC(x.y) = \frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \qquad (4.9)$$

4.2.2 LINK PREDICTION BY PATH-BASED METRICS

Using path-based metrics one can compute paths between two nodes as similarity between node pairs.

• **Local Path** (LP) [24]: The local path based metric uses the path of length 2 and length 3. The metric uses the information of the nearest

neighbours and it also uses the information from the nodes within length of 3 distances from the current node. The paths of length 2 is more relevant that paths of length 3 so a parameter β is applied the value of which is close to 0. One can compute it using the Adjacency matrix using the equation given below,

$$LP = A^2 + \beta A^3 \qquad (4.10)$$

- **Katz metric** [25]: Similarity measure based on all paths in a graph. This function counts all the paths between given pair of nodes with shorter paths counting more heavily. Parameters are exponential.

$|path_{x,y}^{l}|$ is the set of all the paths between x and y with length l and $\beta >$ 0.

$$Katz\,(x,y) = \sum_{l=1}^{\infty} \beta^{l}.\left|path_{x,y}^{l}\right| = \beta A + \beta^2 A + \beta^3 A^3 + \cdots \qquad (4.11)$$

- **Geodesic similarity metric**: This function calculates similarity score for vertices based on the shortest paths between x and y. Its given using the equation below,

$$Geodesic(x,y) = \sum_{l=1}^{\infty} |shoretst\ paths|_{x,y}^{l} \qquad (4.12)$$

- **Hitting time and Commute time** [26]: Hitting time is calculated based on a random walk starts at a node x and iteratively moves to a neighbor of x chosen uniformly at random. The hitting time Hx,y from x to y is the expected number of steps required for a random walk starting at x to reach y. Since this metric is not symmetric, for undirected graphs the commute time, $Cx,y = Hx,y + Hy,x$ can be used.

• **Random Walk with Restart** [26–28]: Random walk is a useful mathematical framework that provides a systematic way to measure importance of nodes in a network. The most widely known is the PageRank algorithm [29]. PageRank, developed for ranking web pages, measures page clicks of hypothetical web surfers who randomly click hyperlinks in the network of webpages. Since it is possible for the surfer to be trapped in a dead-end webpage that does not have any outgoing link, at each time step the surfer may jump to a random webpage with a probability c. Interestingly, this formulation also provides a simple way to define a random walk-based "distance" from a node a (or a set of nodes) to every other node, namely by allowing the random walkers to jump only to the source node a (or the source set of nodes) and restart from there. As a result, it is more likely to find the random walker at the vicinity of the source node than at a distant part of the network, and thus we are able to estimate the relevance (closeness) of each node with respect to the source node. The prediction method applies this idea to identify drugs and targets that are relevant to a set given set of drugs and targets.

Consider an undirected, unweighted network G = (V, E),where V is the set of nodes and E is the set of links. For each pair of nodes a,b∈V we can assign a proximity score by executing the following procedure:

– we start a random walker from a.

– At each time step, with the probability 1 - c, the walker walks to one of the neighbors, b, according to the transition value matrix $W_{ab} = \frac{S_{ab}}{K_a}$ where S_{ab} is the adjacency matrix of the network and (S_{ab} equals 1 if

node a and b are connected, 0 otherwise) K_a denotes the degree of a.

– With probability c , the walker goes back to a.

– After many time steps the probability of finding the random walker at node x converges to the steady-state probability which is our proximity score $S_{a \to x}$.

The random walk with restart, whose updating equation is shown as follows:

$$p_{t+1} = (1\text{-}c)W^T p_t + cp_0 \qquad (4.13)$$

Keep updating p until convergence; the stationary distribution vector p can meet,

$$pt = (1 - c)(I - cW^T)^{-1} p0 \qquad (4.14)$$

4.3 INSTALLATION

A stable tested version of from github using the devtools package. Installing the package from github is given below,

```
install.packages("devtools")

library(devtools)

install_github("abhik1368/netpredicter")
```

4.4 USING NETPREDICTOR STANDALONE R PACKAGE

One can look at the properties, which can be calculated on unipartite graphs.

```
require(igraph)
require(netpredictor)

g1   <-   upgrade_graph(erdos.renyi.game(100,
1/100)) V(g1)$name <- seq(1,100,1)
score_mat        <-        unetSim(g1,"aa")
head(which(score_mat!=0, arr.ind = T))

##  Common  neighbors  vertex  similarity
score_mat      <-      unetSim(g1,"cn")
head(which(score_mat!=0, arr.ind = T))

## Jaccard Index similarity
score_mat               <-
unetSim(g1,"jc")

## Dice similarity
score_mat <- unetSim(g1,"dice")

##  Katz  Index  similarity
score_mat               <-
unetSim(g1,"katz")

##  Geodesic  distance  vertex  similarity
score_mat <- unetSim(g1,"dist")

##  Cosine  vertex  similarity/  Salton  index
score_mat <- unetSim(g1,"cosine")

## Preferential attachment vertex similarity
score_mat <- unetSim(g1,"pa")

## Local Paths Index
## This function counts the number of two-paths and
##    three-paths    between
nodes.    score_lpsim    <-
unetSim(g1,"lp")

## Hub promoted Index
## This measures assigns higher scores to links adjacent to hubs
## (high degree nodes). It counts common neighbors of two vertices
## and weights the result.
score_hpsim             <-
unetSim(g1,"hpi")

## Similarity measure based on resource allocation process
## (number of common neighbours weighted by the inverse of degrees)
score_hpsim <- unetSim(g1,"ra")
```

Next we look at the properties, which can be calculated on bipartite graphs.

```
library(igraph)
library(netpredictor)
data(Enzyme)

## Get the Enzyme and compound adjacency matrix
A <- t(enzyme_ADJ)

## degree Centrality of the Bipartite Graph
get.biDegreeCentrality(A,SM=FALSE)

## Compute Graph density of Bipartite Graph
get.biDensity(A)

## Compute betweeness centrality of Bipartite Graph
get.biBetweenessCentrality(A)


## Projects Bipartite Networks into monopartite networks default method
## is shared neighbours.
get.biWeightedProjection(A,weight = TRUE)
```

Next, we will use the different methods to predict links. Here we have shown examples based on drug target prediction. With the growing understanding of complex diseases, the focus of drug discovery has shifted away from “one target, one drug” model, to a new “multi-target, multi-drug” model. Predicting potential drug-target interactions from heterogeneous biological data is critical not only for better understanding of the various interactions and biological processes, but also for the development of novel drugs and the improvement of human medicines. To predict polypharmacology people use bayesian methods, SVM and Random Forest models, but in all of those algorithms the methods depends on labelled data to predict unknown links. Network based approaches does not rely on labelled data . Two of the algorihtms implemented in this package Random walk based Restart (RWR) and Network based Inference (NBI) to do

it. For performing RWR we used Drug target network, which is a bipartite graph in which every links connects drugs to proteins.

```
data(Enzyme)
## load the adjacency matrix
A <- enzyme_ADJ
## load the chemical similarity matrix calculated from other
## packages or softwares
S2 = enzyme_Csim
## load the protein similarity matrix
S1 = enzyme_Gsim
## Convert the adjacency matrix to igraph object
## because biNetwalk function used igraph object
g1 = graph.incidence(A)
## Run the RWR in bipartite network.
pScore <- biNetwalk(g1,s1=S1,s2=S2,normalise="laplace", dataSeed=NULL,
                restart=0.8, parallel=FALSE, multicores=NULL, verbose=T)
```

In this example we attempt to use the dataseed file, which contains the pairs relations between targets and drugs. This can be useful when one is trying to investigate relations for a specific set of relations. The Drug names and proteins names should be included in the adjacency matrix when one uses the file option to provide dataseed. In the dataseed file the first column contains the proteins names and the second column the drug names. Ouput is a matrix of unique drugs against the number of targets in the adjacency matrix.

```
library(netpredictor)
data(Enzyme)
A <- t(enzyme_ADJ)
g1 <- upgrade_graph(graph.incidence(A,mode = 'all'))
S1 = enzyme_Csim
S2 = enzyme_Gsim
## Read the dataseed file from the user
dataF<-  read.csv("seedFile.csv",header=FALSE)
Mat <- biNetwalk(g1,s1=S1,s2=S2,normalise="laplace", dataSeed =dataF,
                 restart=0.8,parallel=FALSE, multicores=NULL, verbose=T)
```

In this next example we will see how we can plot the significant communities of drugs from the final RWR computed matrix. For community detection we used the walktrap algorithm [13], which places nodes into communities based on neighborhood similarity from short random walks. We also input a list of drugs as vector and retrieve top 10 interactions for each of those drugs. In this package after getting the results one can easily write the results in GML format for visualization in Gephi or cytoscape. It also support export to GEXF format (Gephi specific file format). Below shows the example of exporting to GML format.

Drug	Pnames	score	Type
D00014	hsa2936	0.396969	True Interaction
D00014	hsa2950	0.3973772	True Interaction
D00014	hsa1719	0.0011479	Predicted Interaction
D00014	hsa55312	0.0010483	Predicted Interaction
D00014	hsa7172	0.0010385	Predicted Interaction
D00014	hsa4128	0.0009746	Predicted Interaction

Table 4.1: Table shows the example of drug target interactions

```
A <- enzyme_ADJ
S1 = enzyme_Gsim
S2 = enzyme_Csim
g1 = graph.incidence(A)
Q  =  biNetwalk(g1,s1=S1,s2=S2,normalise="laplace",dataSeed=NULL,restart=0.8,
              parallel=FALSE,multicores=NULL,  verbose=T)
head(getTopresults(A,Q,top=10,druglist=NULL))
##Saving the top results in GML format for visualization in Gephi.
g<-graph.data.frame(result[,1:2],directed=FALSE)
## Set the edge values
g <- set.edge.attribute(g, "weight", value=result[,3])
saveGML(g,"netresult.gml","netresult")
```

Select a druglist to get the results for each of the drugs.

```
drugs = c("D00014","D00018", "D00029", "D00036","D00045","D00049")

result <- getTopresults(A,Q,top=10,druglist=drugs)
```

The net.perf function samples and removes links from the adjacency matrix and predicts them and calculates area under accumulation curve, AUC, BEDROC (bdr), and Enrich- ment factor (EF). The area under the receiver operating characteristic (ROC) curve (AUC) is widely used metric for evaluation of predictive models. The advantage of using AUC is that it is bounded, between 0 to 1 with 0.5 corresponding to random prediction. But AUC method has been critized in cheminformatics based virtual screening methods because it is not sensitive to early recognition compounds. The EF tries to solve early recognition problem but it is dependent on the ratio of actives to inactives and the choice of subset X (fraction of active and inactive set). To try and overcome these limitations numerous other evaluation methods, such as robust initial enhancement (RIE) [10] and Boltzmann-enhanced discrimination of ROC (BEDROC) have been proposed [11]. Sheridan et al. developed an exponential weighted scoring scheme RIE which gives heavier weight in “early recognized” hits.

The BEDROC is constructed on top of RIE by, in essence, forcing the RIE to be bounded by 0 and 1, avoiding the dependence on the active/inactive ratio. In the example below we remove 50 links and repredict those links. While re-predicting them we calculate performance metrics like AUC, bedroc and enrichment factor. As the number of links (re- links) increases the performance of prediction drops. 'Calgo' option uses different algorithms like NBI,RWR and netcombo.

```
 data(Enzyme)

A = enzyme_ADJ
S1 = enzyme_Gsim
S2= enzyme_Csim

## We want to remove the links from the links which has
## two or more interactions.
m = net.perf(A,S1,S2,alpha=0.5,lamda=0.5,relinks = 50,numT=2,Calgo="nbi")
```

In 2010 Zhou et al. [], proposed a recommendation method based on the bipartite network projection technique implementing the concept of resources transfer within the network. The method developed here is based on Alaimo etal.[32]. The example given below one can use both the methods, using similarity matrices or simply use heatS equation with the adjacency matrix. The nbiNet function is developed to perform the prediction.

Type	AUAC	AUC	AUCTOP	BECROC	EFC
RWR	0.932	0.976	0.606	0.520	9.206
NBI	0.923	0.976	0.676	0.500	5.468
NetCombo	0.936	0.976	0.602	0.520	9.087

Table 4.2: Results network performance using all the algorithms.

```
data(Enzyme)
A <- t(enzyme_ADJ)
S1 = as.matrix(enzyme_Csim)
S2 = as.matrix(enzyme_Gsim)
g1 = graph.incidence(A)
## Using the similarity matrices
P1 <- nbiNet(A,alpha=0.5, lamda=0.5,  s1=S1, s2=S2,format = "matrix")
```

I can also compute performance metrics using different algorithms like NBI, RWR and netcombo (fusion of results of NBI and RWR) to get auac auc, auctop, bdr and ef so that one can compare the performance using different algorithms.

```
library(netpredictor)
library(igraph)
data(Enzyme)
A <- t(enzyme_ADJ)
S1 = as.matrix(enzyme_Csim)
S2 = as.matrix(enzyme_Gsim)
## Use all the algorithms NBI, RWR and netcombo
m = net.perf(A,S1,S2,relinks = 50,numT=2,Calgo="all")
tab  <-  rbind(data.frame(m[[1]]),data.frame(m[[2]]),data.frame(m[[3]]))
```

To calculate the significance of an interaction, I first compute the association score between drug a target and we want to found out whether the predicted association score is significant or not. We make 1000 permutations of the association matrix and similarity matrix and compute NBI scores for 1000 random matrices and then we used a

normal distribution to calculate p-value. Then I convert the original compute score to an associated Z-score. Once the Z-score is found the probability that the value could be less the Z-score is found using the pnorm command. Also for a two sided test we need to multiply the result by two. Box below gives an idea how we can achieve this. We can create a significant network based on these significant associations found. The example give below shows the computation using network-based inference for 1000 permutations.

```
data(Enzyme)
A <- t(enzyme_ADJ)
S1 = as.matrix(enzyme_Csim)
S2 = as.matrix(enzyme_Gsim)

## Compute NBI
P1 <- nbiNet(A,alpha=0.5, lamda=0.5,   s1=S1, s2=S2,format = "matrix")

## Create a list where to store the matrices
perm = list()

## Set a random seed
set.seed(12345)

## Compute scores for 1000 permutations where you sample the matrix
## everytime
for ( i in 1:10){
    A <- t(enzyme_ADJ)
    A  <-  A[sample(nrow(A)),sample(ncol(A))]
    S1 = as.matrix(enzyme_Csim)
    S2 = as.matrix(enzyme_Gsim)
    S1  <-  S1[sample(nrow(S1)),sample(ncol(S1))]
    S2  <-  S2[sample(nrow(S2)),sample(ncol(S2))]
    R1 <- nbiNet(A,alpha=0.5, lamda=0.5,   s1=S1, s2=S2,format = "matrix")
    perm[[i]] <- R1
}
extractUC  <-  function(x.1,x.2,...){
    x.1p <- do.call("paste", x.1)
    x.2p <- do.call("paste", x.2)
    x.1[! x.1p %in% x.2p, ]
}
## Get the mean and standard deviation of the matrix

mean_mat <- apply(simplify2array(perm), 1:2, mean)
sd_mat   <-   apply(simplify2array(perm), 1:2,  sd)

## Comput the Z-score of matrix
Z <- (P1 - mean_mat)/sd_mat
Z[is.nan(Z)] = 0
## Compute the significance score
sigNetwork  <-  2*(pnorm(-abs(Z)))

## Get the significant interactions where, P < 0.05
sigNetwork[sigNetwork  <   0.05] <- 1
sigNetwork[sigNetwork != 1] <- 0

sum(A) ## Total number of interactions we had earlier
[1] 1515

sum(sigNetwork) ## Total number of interactions after computation.
[1] 2368
```

4.5 USING NETPREDICTOR R SHINY WEB APPLICATION

The interface is consisted of two parts; web interface and web server. Both of these components are controlled by the code that is written within the framework of Shiny application in R. The building block of Shiny package is based on reactive programming. Since the ma- jor task of the web-based application is to get the inputs and produce outputs, the whole programming language is designed in a reactive programming approach so that a change in any input instantly change the end result. The Shiny application automatically updates the data tables and graphs in real-time. This is an advantage for the web applications that rely on user inputs. The shiny procedure can provide different outputs without the need to refresh the web page. Within the shiny package, ordinary controllers or widgets are provided for ease of use of application programmers. Many of the procedures like uploading files and refreshing the page for drawing new plots and tables are provided automatically. Websockets are exclusively important in situations where there is constant back and forth dialogue or data exchange between the clients and servers. The communication between the client and server is done over the normal TCP connection. The bulk of live data traffic that is needed for many of web applications (i.e. online games) between the browser and the server is facilitated over the websockets protocol. This protocol operates separately and only handshake between the client and server is done over the HTTP protocol. The duplex connection is open all the time and therefore the authentication is not needed when exchange is done. Figure

2.1 shows the web framework architecture for R shiny Web app.

In order for a Shiny app to execute, we have to create a Shiny server installed in linux or CentOS. Shiny follows a pre-defined way to write R scripts. It consists of server.R and ui.R, which need to be in same directory location. If a developer want to customize the user interface shiny can also integrate additional CSS and JavaScript within the web application. The netpredictor shiny app is available at https://github.com/abhik1368/Shiny_NetPredictor .

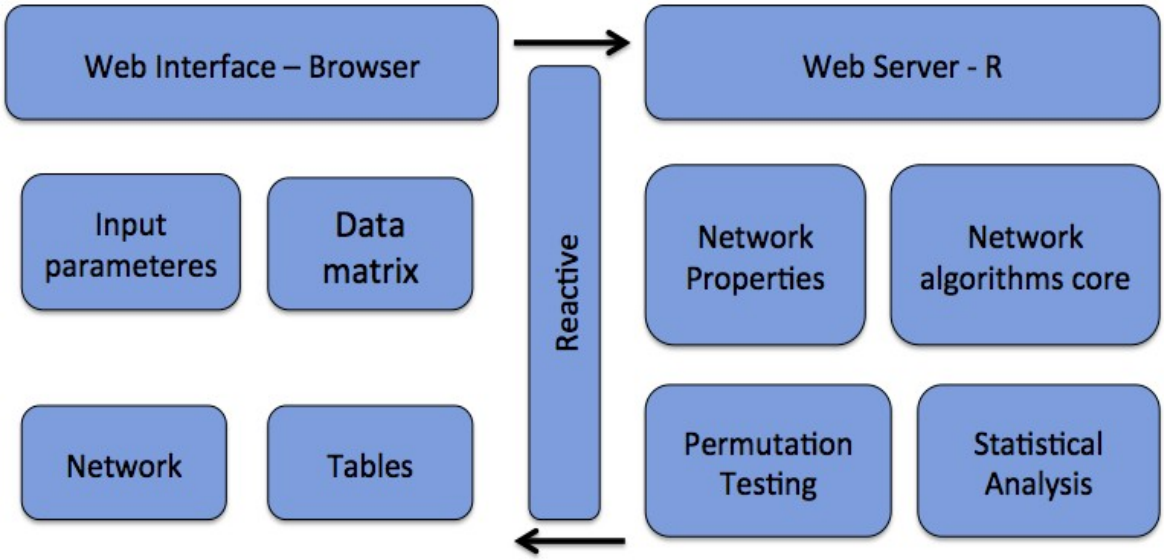

Figure 4.1: Diagram showing web architecture of NetPredictor Shiny app.

4.6 DESCRIPTION

4.5.1 LOADING DATA

One can load their own data or can use the given sample datasets given in the software. For the custom dataset option one needs to upload bipartite adjacency matrix along with the drug similarity matrix and protein sequence matrix. From the given datasets - enzyme, GPCR, Ion Channel and Nuclear Receptor in the application one can load the data and set the parameters for the given algorithms and start computations. Figure 4.2 shows the start page.

4.5.2 RESULTS

Once the data is loaded in the workspace and prediction button is pressed it instantly shows up network properties of the bipartite network in network properties tab and the predicted results of the given algorithm used. The results are easily downloadable as a csv file. It also shows up the interactive network plot it comes with true and predicted interactions in the network. The network can be downloaded as GML file.

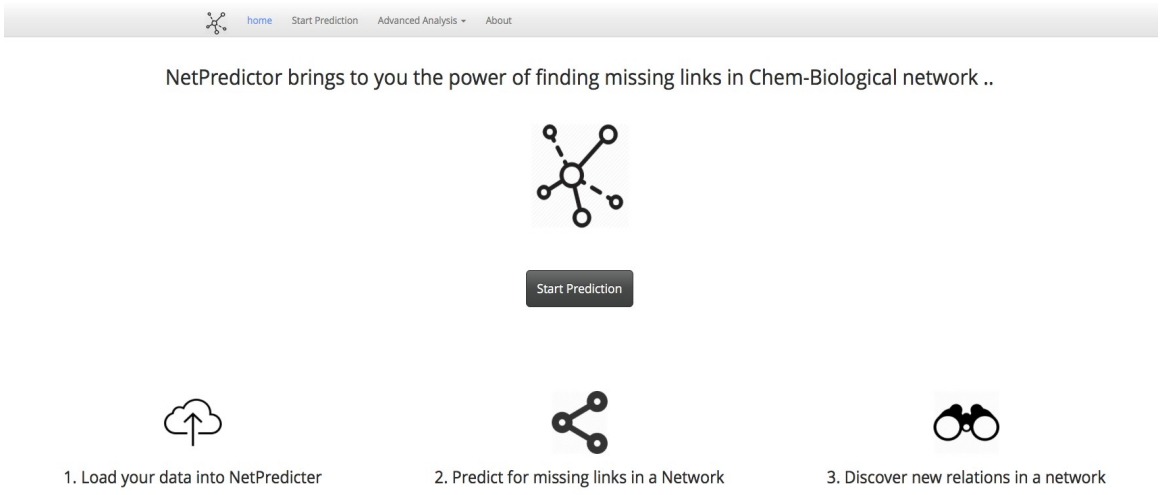

Figure 4.2: Diagram starting page of NetPredictor Shiny app.

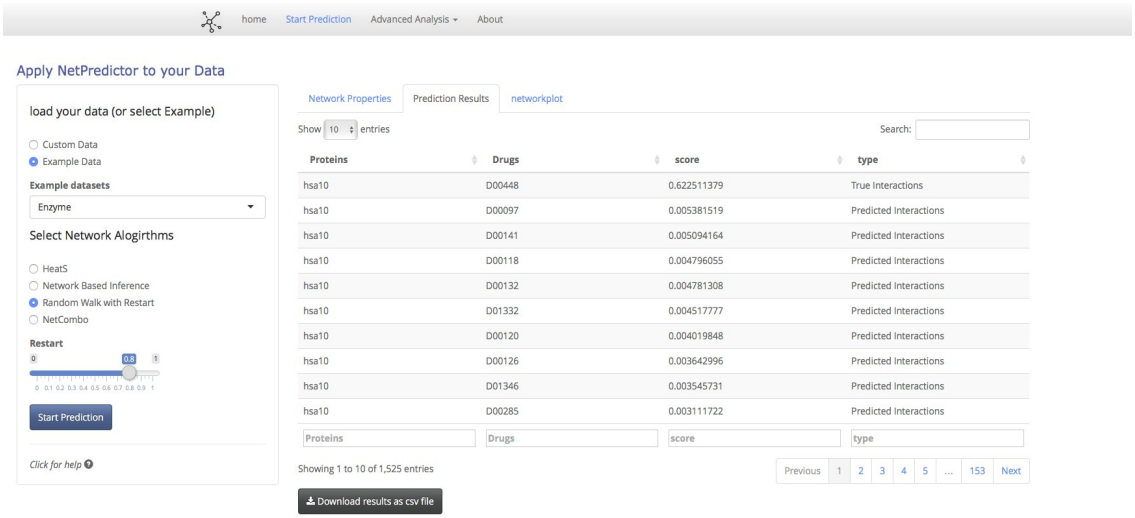

Figure 4.3: Diagram showing results page of NetPredictor Shiny app.

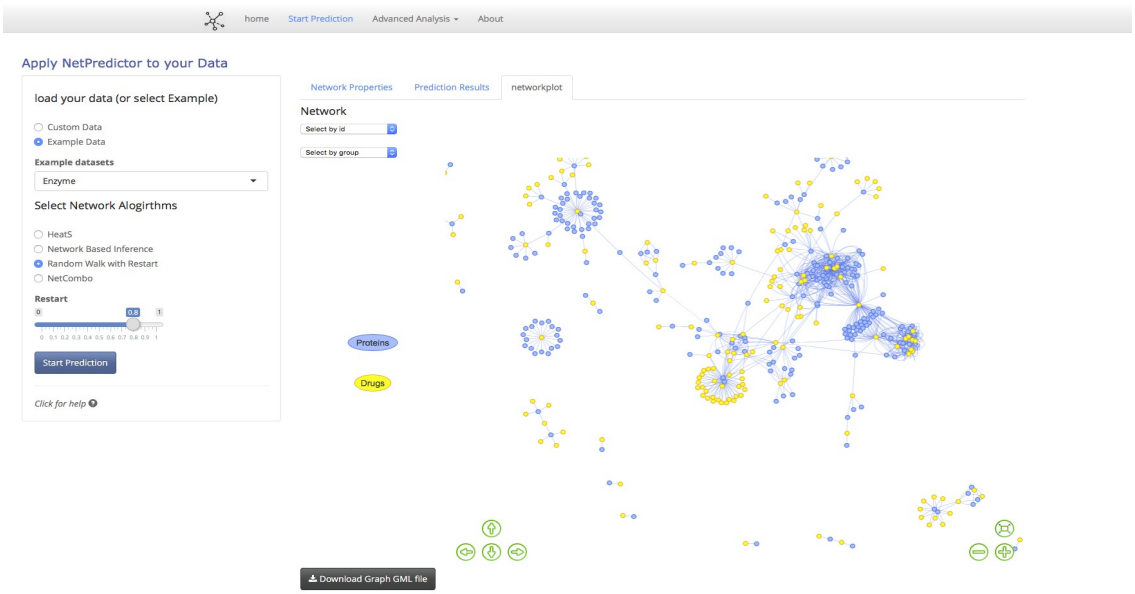

Figure 4.4: Diagram shows network plot page NetPredictor Shiny app.

4.5.3 ADVANCED ANALYSIS

In the advanced analysis tab one can compute the different statistical metrics of your given data for three different algorithms NBI,RWR and NetCombo with given set of parameters. This can easily identify the performance of the network algorithms on your data. Figure 4.5 shows the advanced analysis tab. A user can run number of times the algorithms with different sets of parameter settings. Two parameters are provided i.e removing the random links from a network with drugs having more than given frequency of targets. If the frequency of the targets is select as 0 then all the drug target relations are selected and if it is 2 only those drugs having more than 2 targets links will be removed from those drugs. The permutation testing can be also

performed using this tool. The method usually computes significant relations based on the number of random permutations with the original predicted matrix. One can select the significance level and get the results with significance level less than that given value. Figure 4.6 shows the results of prediction.

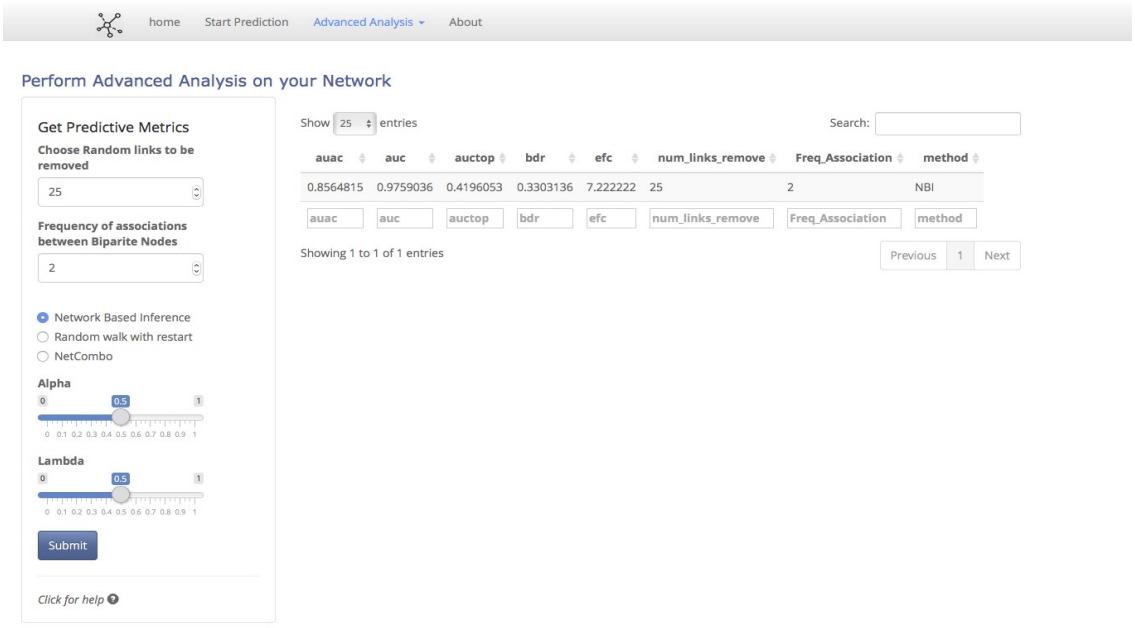

Figure 4.5: Diagram shows advanced analysis page NetPredictor Shiny app.

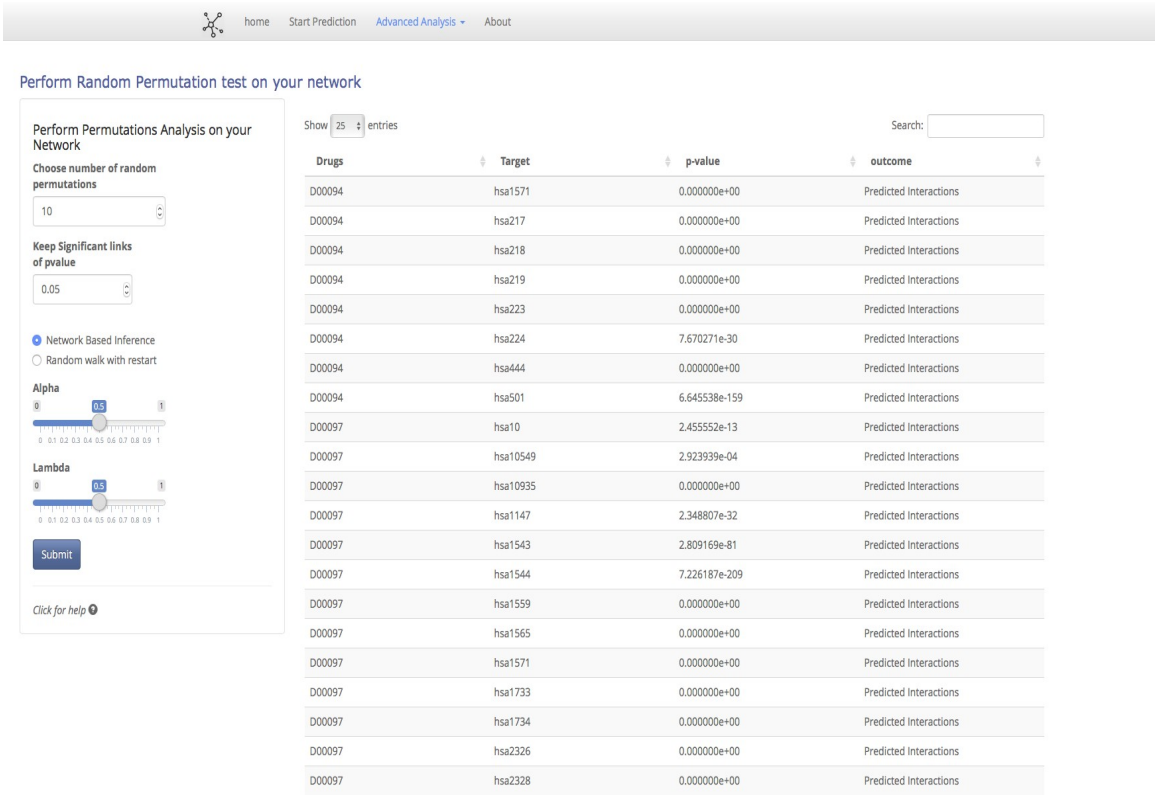

Figure 4.6: Diagram shows Random permutations results page NetPredictor Shiny app.

CONLCUSION

The netpredictor standalone package and shiny application helps in identification of missing links in bipartite and unipartite network. This application is not useful for biomedical domain but it can also be used in searching links in social informatics. The standalone is built using R and the web platform is built using R Shiny web, which integrates packages like shinythemes, shinysky, data.tables.js, vis.js, d3.js, gridster.js, igraph and reshape2.

CHAPTER 5

FUTURE WORK AND SUMMARY

5.1 Future Work

Understanding polypharmacology is a critical problem in drug discovery especially of-class target polypharmacology which causes adverse effects and side effects. Random walk with restart has lot of potential in understanding of target-mediated effects and also will be useful for drug-repurposing applications. Other major areas where it can be used is in prioritization of candidate genes for various diseases, metabolic pathways, prediction of disease associated microRNAs, clustering of proteins based on protein-protein interactions, image segmentation and etc.

5.2 Conclusion

In this study, I examined the method of link prediction. Although the link prediction is not a new problem in informatics but it seems that the traditional methods have not been up with the recent development network science. One of biggest challenges in link prediction is using multi-dimensional and multi-partite networks where each of the link associations could have a different meaning and consists of different classes of nodes in it respectively. For example, multiple relations can exist between a ligand and target, a drug binds to a protein target based on its activity, whether it is an inducer, activator or enhancer, it can show adverse events. In my opinion in-depth understanding, complex networks (example making use of the modular structure of the network and hierarchical organization) can help in the development of advanced link

prediction algorithms. In this work, I developed a novel technique of random walk with a restart to predict drug-target interactions and identification of metabolic pathways of a given disease condition. Studying drug target interaction using traditional techniques is time-consuming and needs some special software packages to understand potential interactions. However, we can reduce the time gap by using methods like random walk with restart to rank the interactions and then study significant interactions using special cheminformatics tools. Chapter 1 introduced some traditional ways of doing ligand based and target-based methods to study drug target interactions and introduced some link prediction methods in using neighborhood based and path-based metrics. Focusing on the path-based metrics, I introduced the concept of random walk with restart and how we can use to predict missing links in a network. Chapter 2 introduces random walk restart in heterogeneous drug target network where we integrated drug-drug chemical similarity network, drug-target network and target-target network based on protein sequence similarity. I thought using different chemical features would result in a different ranking of targets, but surprisingly I observed using four different chemical features and optimizing a parameter η I achieved similar kind of results, which indicated using commercial or open source chemical similarity fingerprints for drug network the results doesn't vary much. Next, Chapter 3 focuses on the using RWR to a four-layered (disease layer, protein layer, protein complexes layer and the protein-biological pathway layer) network for identification Biological pathways related to a query disease. The protein layer consists of a ppi network from a tissue based gene expression, which connects to disease layer via association of disease tissues associations. Based on the

query disease a given tissue is selected, and a particular protein-protein network is loaded and then we generate the full four-layered network and get the pathway predictions. We tested two different types of disease similarity network using K-Nearest Neighbor and similarity threshold based and showed that threshold of 0.3 gave us better performance with restart c of 0.9.

In chapter 4 we introduce the netpredictor package to compute properties and predict links in a bipartite and unipartite network. I also develop some functions to compute node centrality measures in two-mode networks using bipartite network projection. I have developed three algorithms to predict missing links in a bipartite network namely NBI, HeatS and RWR. Apart from using a standalone package a web-based software developed in R shiny. Shiny is a platform to develop web-based applications using R. The matrix computations are done using the Revolutions Analytics parallel package and later on upgraded to Revolutions R open which includes intel math kernel (MKL) which provides BLAS and LAPACK library functions. For Mac OSx users, it uses ATLAS blas library functions. MKL uses, as many parallel threads as there are number of cores. The shiny web app uses some javascript libraries like bootstrap.js for creating navbars and tabs, vis.js for developing the interactive network based visualization, data.tables.js for generating tables, gridster.js for moving the network properties grid around the page. It uses some packages like shinythemes, shinyBS, shinyjs, shinysky for look and feel of the web application.

BIBLIOGRAPHY

[1] Scannell J.W, Blanckley A., Boldon H., and Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov*, 11:191–200, 2012.

[2] Munos B. Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov*,8:59—968, 2010.

[3] B. Booth and R Zemmel. Prospects for productivity. *Nat Rev Drug Discov*, 3:451–6, 2004.

[4] C.R. Chong. and D.J. Sullivan Jr. New uses for old drugs. *Nature*, 448:645–646, 2007.

[5] Tartaglia LA. Complementary new approaches enable repositioning of failed drug candidates. *Expert Opin Investig Drug*, 15:1295–8, 2006.

[6] Nielsch U., Schafer S., and Wild H. One target-multiple indications: a call for anintegrated common mechanisms strategy. *Drug Discov Today*, 12:1025–31, 2007.

[7] Eric Alm and Adam P Arkin. Biological networks. *Current Opinion in Structural Biology*, 3(2):193 – 202, 2003.

[8] Yvonne C. Martin, James L. Kofron, and Linda M. Traphagen. Do structurallysimilar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358, 2002.

[9] M.A. Campillos. Drug target identification using side-effect similarity. *Science*, 321(5886): 263–266, 2008.

[10] Schierz AC. Virtual screening of bioassay data. *Jcheminf*, 1(21), 2009.

[11] Seal A., Passi A., Jaleel U.C.A., Wild D.J., and OSDD Consortium. In-silico predictive mutagenicity model generation using supervised learning approaches. *Jcheminf*, 4(10), 2012.

[12] Cumming J.G, Davis A.M., Muresan S., Haeberlein M., and Chen H. Chemical predicive modelling to improve compound quality. *Nature Reviews Drug Discovery*, 12:948– 962, 2013.

[13] Fredrik S, Anders K, and Christian S. Virtual screening data fusion using both structure and ligand-based methods. *J Chem Inf Model*, 52(1):225–232, 2012.

[14] Salam N.K., Nuti R., and Sherman W. Novel method for generating structure-based pharmacophores using energetic analysis. *J Chem Inf Model*, 49:2356–2368, 2009.

[15] Dixon S.L., Smondyrev A.M., Knoll E.H., Rao S.N., Shaw D.E., and Freisner R.A. Phase: a new engine for pharmacophore perception, 3d qsar model development, and 3d databases screening: 1 methodology and preliminary results.*J Comput Aided Mol Des*, 20:647–671, 2006.

[16] David Liben-Nowell and Jon M. Kleinberg. The link prediction problem for social networks. *J Comput Aided Mol Des*, CIKM:556–559, 2003.

[17] Mohammad Al Hasan and Mohammed J. Zaki. A survey of link prediction in social networks. *Social Network Data Analytics.*, pages 243–275, 2011.

[18] P. Jaccard. Etude comparative de la distribution florale dans une por-tion des alpes et de jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[19] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi.Hierarchical organization of modularity in metabolic networks. *Science*, 297:1553, 2002.

[20] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. social networks,.*Social Networks*, 25(3):211–230, 2002.

[21] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*,286:509–512, 1999.

[22] Zhou T., Lu L., and Zhang Y.C. Predicting missing links via local information. *Eur Phys JB*,71:623–630,2010

[23] Leicht E.A., Holme P., and Newman M.E.J. Vertex similarity in networks. *Phys RevE*, 73:026120, 2006.

[24] Lu L., Jin C.H., and Zhou T. Similarity index based on local paths for link prediction of complex networks. *Phys Rev E*, 80:046122, 2009

[25] Katz L. A new status index derived from sociometric analysis. *Psychometrika*, 18:39– 43, 1953.

[26] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans Knowl Data Eng*, 19:355–369, 2007.

[27] Abhik Seal, Yong-Yeol Ahn, and David J Wild. Optimizing drug–

target interaction prediction based on random walk on heterogeneous networks. *J Cheminform.*, 7:40, 2015.

[28] Sebastian Kohler, Sebastian Bauer, Denise Horn, and Peter N. Robinson1. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, 82:949–958.

[29] Langville A.N. and Meyer C.D. Google's pagerank and beyond: the science of search engine rankings. *Princeton University Press.*

[30] Magger O, Waldman YY, Ruppin E, and Sharan R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Computational Biology.*, 82(8):e1002690, 2012

[31] Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, and Tang Yun. Pre- diction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, 2012.

[32] Alaimo S, Pulvirenti A, Giugno R, and Ferro A. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, 29(16).

[33] Chen X, Liu MX, and Yan GY. Drug–target interaction prediction by random walk on the heterogeneous network. *Mol BioSyst*, 8:1970–1978, 2012.

[34] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharana. Predict: a method for inferring novel drug indications with application to personalized medicine. *Mol BioSyst*, 7:496, 2011.

[35] Meng Zhou, Xiaojun Wang, Jiawei Li, Dapeng Hao, Zhenzhen Wang, Hongbo Shi, Lu Han, Hui Zhou, and Jie Sun. Prioritizing candidate disease-related long non-coding rnas by walking on the heterogeneous lncrna and disease network. *Mol BioSyst*, 11:760, 2015.

[36] Campillos M, Kuhn M, Gavin A-C, Jensen L.J, and Bork P. Drug target identification using side-effect similarity. *Science*, 321:263–266, 2008

[37] Cheng F, Liu C, Jiang J, and et al. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8:e1002503, 2012.

[38] Yamanishi Y, Araki M, Gutteridge A, Wataru H., and Minoru K. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24:i232–40, 2008.

[39] Yamanishi Y, Kotera M, Kanehisa M, and Susumu G. Drug-target interaction pre- diction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26:i246–54, 2010.

[40] Wild D.J., Ding Y., Sheth A.P., Harland L., Gifford E.M., and Lajiness M.S. Systems chemical biology and the semantic web: what they mean for the future of drug discovery research. *Drug Discov Today*, 17:469–474, 2012.

[41] Bleakley K. and Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25:2397–2403, 2009.

[42] Chen B., Dong X., Jiao D., Wang H., Zhu Q., Ding Y., and Wild D.J. Chem2bio2rdf: a semantic framework for linking and mining chemogenomic and systems chemical biology data. *BMC Bioinform*, 11:255, 2010.

[43] Bin Chen, Ying Ding, and David J. Wild. Assessing drug target association using semantic linked data. *PLoS Comput Biol*, 8:e1002574, 07 2012.

[44] Knox C., Law V., Jewison T., Liu P., Ly S., and Frolkis A. etal. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, 34:D668–D672, 2011.

[45] Durant J.L., Henry D.H. Leland B.A., and Nourse J.G. Reoptimization of mdl keys for use in drug discovery. *J Chem Inf Comput Sci*, 42:1273–1280, 2002.

[46] Rogers D. and Hahn M. Extended-connectivity fingerprints. *J Chem Inf Comput Sci*, 50:742–754, 2010.

[47]https://community.accelrys.com/message/23572357., 2013. [Online; accessed 4 May 2013].

[48] ROCS,OpenEye Scientific Software. http://www. eyesopen.com/rocs.

[49] Morgan H.L. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem Doc.*, 5:107–112., 1965.

[50] OMEGA, OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com/rocs.

[51] Langville A.N. and Meyer C.D. Google's pagerank and beyond: the science of search engine rankings. *Princeton University Press.*

[52] Seal A., Yogeeswari P., Sriram D., Consortium OSDD, and Wild D.J. Enhanced ranking of pknb inhibitors using data fusion methods. *J Cheminform*, 5:2, 2013.

[53] Truchon J.F. and Bayly C.I. Evaluating vs methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model*, 47:488:508, 2007.

[54] Top pharmaceuticals http://cbc.arizona.edu /njardarson /group / top-pharmaceuticals-poster.

[55] Gaulton A., Bellis L.J., Bento A.P., Davies M. Chambers J., and Hersey A. etal.Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, 40(D1):D1100–D1107, 2012.

[56] Roth B.L., Lopez E., Patel S., and Kroeze WK. The multiplicity of serotonin recep- tors:uselessly diverse molecules or an embarrassment of riches. *Nucleic Acids Res*, 40(D1):D1100– D1107, 2012.

[57] Wang Y., Xiao J., Suzek T.O., Zhang J., Wang J., and Bryant S.H. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*, 37:W623–W633., 2009.

[58] Goh K, Cusick ME, Valle D, Childs B, Vidal M, and Barabási AL. The human disease network. *Proc Natl Acad Sci*, 2:8685–8690, 2007.

[59] Goh K.I and Choi IG. Exploring the human diseasome: the human disease network. *Brief Funct Genomics*, 11(6):533–542, 2012.

[60] Zhao S and Iyengar R. Systems pharmacology: Network analysis to identify multiscale mechanisms of drug action. *Annu Rev Pharmacol Toxicol*, 52:505–521, 2012.

[61] Rahnenführer J., Domingues S.F, Maydt J, and Lengauer T. Calculating the statisti- cal significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–29, 2004.

[62] Moreau Y and Tranchevent L.C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature reviews Genetics*, 13:523–536, 2012.

[63] Yu S Van Vooren S Van Loo P et al. Tranchevent LC, Barriot R. Endeavour update: a web resource for gene prioritization in multiple species. *Nature reviews Genetics*, 13:523–536, 2008.

[64] Huttlin E. L., Jedrychowski M. P., Elias J. E., Goswami T., Rad R., Beausoleil S. A., Villén J.and Haas W., Sowa M. E., and Gygi S. P. A tissue-specific atlas of mouse protein phosphorylation and expression. *cell*, 143:1174–1189, 2010.

[65] K. Lage, N. T. Hansena, and Karlberg E.O et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci*, 105(52):20870–20875, 2008.

[66] K. Lage, Karlberg O.E., and Størling Z.M. et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316, 2007

[67] Vanunu O., Magger O., Ruppin E., Shlomi T., and Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1):e1000641, 2010.

[68] Jacquemin T. and Jiang R. Walking on a tissue-specific disease-protein complex het- erogeneous network for the discovery of disease-related protein complexes. *BioMed Research International*, 6:Article ID 732650, 2013.

[69] van Driel A.M, Bruggeman J., Vriend G., Brunner G.H, and Leunissen M.A.J. A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5):535–542, 2006.

[70] Zhong Q, Simonis N, Li Q.R., Charloteaux B, and Heuze F et al. Edgetic perturbation models of human inherited disorders. *Mol Syst Biol*, 5(321), 2009.

[71] Smedley D., B. S. Haider, and Ballester et al. Biomart—biological queries made easy.*BMC Genomics*, 10(1), 2009.

[72] Ruepp A., Brauner B., and Dunger-Kaltenbach I. et al. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(1):D646—D650, 2008.

[73] Kamburov A, Stelzl U, Lehrach H, and Herwig R. The consensuspathdb interaction database. *Nucleic Acids Research*, 41:D793—-D800, 2013.

[74] Kanehisa M., Goto S., Sato Y., Furumichi M., and Tanabe M. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40:D109–14, 2012.

[75] Hebert J Gong L Sangkuhl K Thorn C et al. Whirl-Carrillo M, McDonagh E. Pharma- cogenomics knowledge for personalized medicine. *Clinical Pharmacology Therapeutics.*, 92(4):414–417, 2012.

[76] Timothy Jewison, Yilu Su, and et al. Fatemeh Miri Disfany. Mpdb 2.0: Big improve- ments to the small molecule pathway database. *Nucleic Acids Res*, 42:D478–D484, 2014.

[77] Croft D., O'Kelly G., Wu G., Haw R., Gillespie M., Matthews L., and Caudy M. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39:D691–D697, 2011.

[78] Kelder T., van Iersel M.P., Hanspers K., Kutmon M., and Conklin B.R. etal. Wikipathways: building research communities on biological pathways.

Nucleic Acids Res, 40:D1301–D1307, 2012.

[79] Green ML Kaiser D Krummenacker M Karp PD. Romero P, Wagg J. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol*, 6(R2), 2005.

[80] Darryl Nishimura. Biocarta. *Biotech Software Internet Report*, 2(3):117–120, 2001.

[81] Kandasamy K., Mohan S.S., Raju R., Keerthikumar S., Kumar G.S., and Venu- gopal A.K. et al. Netpath: a public resource of curated signal transduction pathways. *Biotech Software Internet Report*, 11(R3), 2010.

[82] Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, and Goryanin I. The edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3(135), 2007.

[83] De Jong K.A. and Spears W. An analysis of the interacting roles of population size and crossover in genetic algorithms. *Proceedings of the First International Conference on Parallel Problem Solving from Nature*, 1990.

[84] Moreau Y and Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet*, 13(523–536), 2012.

[85] Piro RM and Di Cunto F. Computational approaches to disease-gene prediction:rationale, classification and successes. *FEBS J.*, 279(678–696), 2012.

[86] Zhao ZQ, Han GS, Yu ZG, and Li J. Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Comput Biol Chem*, 57(21–28), 2015.

[87] Li Y and Li J. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC Genomics*, 13(7:S27), 2012.

[88] Sheridan R.P., Singh S.B., Fluder E.M., and Kearsley S.K. Protocols for bridgingthe peptide to nonpeptide gap in topological similarity searches. *J Chem Inf Comput Sci*, 41:1395–1406, 2001.

[89] S. J. Swamidass, C. A. Azencott, K. Daily, and P Baldi. A croc stronger than roc: Measuring, visualizing and optimizing early retrieval. *Bioinformatics*, 26:1348– 1356, 2010.

[90] Robinson P.N, Köhler D, Bauer S., D. Seelow, Horn D, and Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83(5):610–615, 2008.

[91] R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. Accessed 2014,December 30.

[92] RStudio and Inc. shiny: Web Application Framework for R. http://shiny.rstudio.com/. 2013.

[93] Introducing Shiny: Easy web applications in R — RStudio Blog."[Online].Available: http://blog.rstudio.org/2012/11/08/introducing-shiny/. 2012.

[94] Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, Zhang YC. Solving the apparent diversity accuracy dilemma of recommender systems. Proc. Natl Acad. Sci. USA. 2010;107:4511

[95] Berger, S. I., & Iyengar, R. (2009). Network analyses in systems pharmacology. Bioinformatics, 25 2466–2472.

[96] Hert J, Willett P, Wilton DJ, Acklin P, Azzoui K, Jacoby E, Schuffenhauer A: Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. JChemInf Comput Sci 2004 44:1177-1185.

[97] Open-source platform to benchmark fingerprints for ligand-based virtual screening S Riniker, GA Landrum Journal of cheminformatics 5 (1), 1-17.

[98] Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P: Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real datsets. J Chem Inf Model 2012, 52:2884-2901.

Vita

EDUCATION

Indiana University Bloomington *January 2016*
PhD in Chemical Informatics
Minor in Network Science
Overall GPA: 3.9

West Bengal University of Technology *January 2010*
Dual Degree in Masters in Bioinformatics
& Technology
Overall GPA: 7.9/10

EXPERIENCE

AbbVie *May '15 – Nov '16*
Cheminformatics Intern

I worked in several projects at abbvie with a focus on large scale data visualization, building web apps in R shiny for drug adverse events and document search. I also worked on core chemistry related projects on Pipleline pilot example developing some new topological descriptors for compound design. One of the major projects with pipeline pilot was using the open patent chemistry data (SureChEMBL) to search and bioactivity of compounds and build fast structure and substructure search of 17 million compounds on amazon cloud.

European Molecular Biology lab, Cambridge, U K *May '14 – Aug'14*
Cheminformatics Intern

Worked with ChEMBL team on data virtualization converting UniChem Oracle database to Post- gres and perform benchmarking of similarity searching on 70 million records using SQL and NoSQL techologies.

Dow Agrosciences *May '13 - Aug '13*
Cheminformatics Intern

Worked on building pipline for 2D and 3D Virtual Screening and developed user interface and visual- ization results for 2D and 3D virtual screening using Pipeline Pilot web forms.

Indo-US Science Technology Fellowship *May '12 – Aug'12*
Cheminformatics Intern
Identication of Mycoabacterial PknB Inhibitors using datafusion methods.

Indiana University, Bloomington *2015*

Phd in Cheminformatics

Associate Instructor for I571 (Introducing Cheminformatics) and I590 (Data science in Drug discovery health and Translational Science.) courses.

SELECTED PUBLICATIONS

- Seal, A., Ahn Y.Y.,Wild, D.J. Optimizing Drug target predictions based on random walk with restart on heterogeneous networks.(Article submitted to Journal of Cheminformatics)
- Applications of the YarcData Urika in Drug Discovery and Healthcare R Henschel, A Seal, JJ Yang, DJ Wild, Y Ding, A Thota, S Michael, M Gianni, J Maltby - 2014

  https://cug.org/proceedings/cug2014_proceedings/includes/files/pap168.pdf
- Seal, A., Yogeeswari, P., Sriram, D., Wild, D.J. 3D virtual screening of PknB inhibitors using data fusion methods. Journal of Cheminformatics,5:2 2013.
- Seal, A.,Passi A.,Jaleel U.C A.,Consortium OSDD.,Wild, D.J.In-silico Predictive Mutagenicity Model Generation Using Supervised Learning Approaches Journal of Cheminformatics 4:10 2012(Highly Accessed)
- Seal, A., Gupta, A., Mahalaxmi,M., A,Riju., Singh T.R., and Arunachalam,V., (2012) Tools,resources and databases for SNPs and indels in sequences: A review, IJBRA, 2013.(under print)
- Docking study of HIV phytochemicals with Reverse transcriptase by Abhik Seal, Riju Aikkal,Mriganka ghosh. Bioinformation 5(10): 430-439 (2011).
- Singh T.R., Gupta., Aykkal R., Mahalaxmi M., Seal, A.. and Arunachalam V. 2011 Computa- tional identication and analysis nucleotide polymorphisms and insertions/delete ons in expressed sequence tag data of Eucalyptus. J.Genet. 90
- Seal, A. Higher Eciency In Prediction Of TIBO Activity By Evolutionary Neural Network. Avail- able from Nature Precedings (2011)

JOURNAL REVIEWER

- Journal Of Cheminformatics
- BMC Bioinformatics
- Journal of Molecular Graphics & Modelling.

AWARDS

- $2000 Special fellowship for development of data science course materials from IU School of Informatics and Computing 2014.
- American Chemical Society CINF award for excellence in science ,2013 ($1000)
- Indo US science Research Intern awarded travel and stipend to Birla Institute of Technology to do study on PknB inhibitors.
- Indiana University Teaching Assistantship.
- University topper in Msc(TECH) in Bioinformatics at West Bengal University of Technology.
- Travel fund to Second US-Indian Network Enabled Research Collaboration Workshop, March 22-23,2012Wasington DC .