Predicting Drug Mode of Action Using Drug-Drug Similarity Measures

Warunika Ranaweera, Rupika Wijesinghe, Ruvan Weerasinghe
University of Colombo School of Computing, Colombo 00700, Sri Lanka.
warunika@ieee.org, crw@ucsc.cmb.ac.lk, arw@ucsc.cmb.ac.lk

***Abstract*—Computational techniques developed to predict drug mode of action have been given prominence recently owing to their contribution in addressing the increasing drug rejection rates. Most of these recent models rely on the hypothesis that *similar drugs share similar effects*, where the similarity between drug pairs plays a major role in the process. Although several models have been proposed with diverse measures of similarity, each metric bearing its own advantage, all the models suffer from *missing data* when restricted by a specific similarity measure. In our study, we address the problem of missing data with the use of multiple similarity measures, while exploiting their individual advantage. For this purpose, we propose the *Friends of Friends approach for similarity approximation* by electing a mediator drug to indirectly find the similarity between a given drug pair. As a pilot study, we used the pair-wise chemical similarity to approximate the missing similarity distances between drug-induced gene expression profiles. Taking a step further, we built a drug-drug network from approximated similarity distances to obtain communities, in which, drugs that share therapeutic actions are grouped together using a clustering algorithm. Although the pair-wise similarity approximations produced an accuracy of 37%, we observed a precision rate of 0.95 for the drug communities we have obtained by replacing the top 20% of the pair-wise similarity scores, which proves that our approximations work best for drug pairs that display *highly similar* therapeutic actions.**

***Index Terms*—Drug discovery, similarity networks, community detection, gene expression profiling, cheminformatics.**

I. Introduction

ACCORDING to Kola and Landis [**?**], in the year 2000, lack of efficacy and unidentified side effects were together causing approximately 60% of the clinical trials to fail. As a result, in 2007, the U.S. Food and Drug Administration (FDA) agency recorded the fewest number of medicines approved since 1983, tallying to just 19 drugs [**?**]. This number rose only slightly in the following year to 21. At present, the introduction of a new drug incurs approximately $1.8 billion [**?**], which causes an unbearable loss in the event of a clinical failure of any drug candidate. Understanding the system-wide effects of a drug is important to prevent clinical failures and increase drug acceptance rates.

Nevertheless, the traditional one drug-one target paradigm hinders the understanding of the system-wide effects of a drug, which is usually the cause for unexpected drug effects [**?**]. Knowledge of the Mode of Action (MoA) of a drug is helpful to dissect what follows the drug/target interaction, thereby identifying the overall performance of the drug in terms of both efficacy and side effects.

One of the earliest attempts to computational drug target prediction is Virtual High-throughput Screening (VHTS), where hundreds and thousands of *in silico* compounds are docked against one target protein to find an energy-minimized binding affinity. Although this method accurately identifies drug targets, it produces a high computational complexity, such that, when VHTS is used to identify drugs with multiple targets, the computational cost it incurs is overwhelming [**?**]. Another widely used approach is the reconstruction of gene regulatory networks (GRNs) to predict pathways affected by a drug [**?**]. GRNs are complex and noisy, hence limiting the practical applications of that approach. Text mining [**?**], another commonly applied method to identify drug action, is limited to the current knowledge available in the literature; thus it cannot produce new experimental findings [**?**]. To avoid the complexities of these traditional methods, various novel approaches have been brought forth.

Recent advances in computational prediction of drug MoA have introduced drug-drug and target-target similarities to predict individual drug targets, side effects and pathways of novel/existing drug compounds. The common hypothesis underlying these techniques is that *similar drugs share similar effects*. Such methods identify drugs that share therapeutic effects by utilizing the similarity between drug pairs and target pairs; hence predicting the unknown therapeutic effects of a drug, while avoiding the complexities incurred in conventional methods [**?**], [**?**].

Although such novel prediction methods are proven to be effective, they are limited by the *specific measure of similarity* used in the model, such as drug side effects [**?**], drug-response gene expression profiles [**?**], or protein targets [**?**]. This dependency limits the applicability of the existing models only to the drugs which are equipped with the information required by the model. On the other hand, similarity in chemical structure [**?**], although highly available, does not necessarily represent therapeutic similarity in drugs [**?**]. These limitations in the existing network based prediction methods, which can be solved by utilizing the advantage of more than one similarity measure, forms the motivation for our work presented in this paper.

II. Network Based Mode of Action Prediction

In this section we survey the current methods available for network based drug action prediction.

A network can be viewed as a visual abstraction of the relations that exist between a set of objects. In a biological context, vast arrays of datasets are available in public repositories which represent relationships between objects, such as

drugs, target proteins, diseases and genes. Thus, a network abstraction of the objects and their relationships included in these datasets will be immensely helpful in extracting useful information from them.

In the basic sense, a network is a *collection of objects* (drugs, targets, diseases or genes) represented by *nodes*, in which some node pairs are *linked* to each other by *edges* [**?**]. Based on the context, many different forms of relationships can define the links between objects; such as the similarity, interaction, co-operation and transition. When used in a pharmacological context, the nodes are usually drugs and targets, and the links are mostly formed by the similarity or the interaction between a pair of drugs. Hence, the measurement of similarity between drug pairs plays an important role in the construction of drug networks.

A. Measuring Drug-Drug Similarity

A given pair of drugs can be similar in many aspects, such as from the drug family, function, chemical structure or side effects. Various metrics have been introduced to calculate the similarity between a drug pair, so as to obtain a numerical value for the measurement of similarity between two drugs. The five notable similarity measures widely used in the literature can be listed as follows.

1) Similarity in ATC classification code
2) Similarity in compound chemical structures
3) Similarity in gene expression profiles (GEP)
4) Similarity in side-effects
5) Similarity in target proteins

B. Network Models: A Taxonomy

Although all the network models can be considered as graphs that link nodes through edges, a variety of models of networks have been considered in the recent literature; many of them differing from the types of vertices and edges, while some of them also differ by the type of the graph. This section provides a taxonomy of such different network models provided in the literature with various intentions: to predict unknown targets of drugs, their potential off-target effects, as well as the pathways affected by a drug under consideration.

1) Drug-Drug Networks: Drug-drug networks (widely referred to as drug networks) can be considered as the simplest form used in the network analysis approaches in pharmacology. Drug networks that exist in the literature are unipartite, which implies that the network is built considering only the homogeneous nodes and homogeneous links [**?**]. For a list of drugs, the construction of the drug network involves calculating the pair-wise distance between drug pairs using a suitable similarity measure, and interconnecting two drugs based on a given distance threshold.

Iorio et al. [**?**] uses the distance between drug-induced GEP as the similarity score in the drug network used in their study. Once the drug network is constructed, the authors detect communities of drugs (i.e. a tightly connected set of drugs), which they hypothesize as having the same MoA. Although efficient and accurate, relying on a single type of similarity measure may limit the applicability of this network model when the required data type is not available for a specific drug of interest.

2) Drug-Disease Networks: Drug repositioning studies are essential to predict new disease indications for existing drugs. Recent studies make use of disease-disease and drug-drug similarity, together with known drug-disease associations, to predict new drug treatments for diseases [**?**], [**?**]. Such networks are also useful to identify the underlying, hidden mechanisms of diseases. The major drawback of the drug-disease network model is the high false negative rates encountered [**?**], which degrades the value of this simple and efficient approach.

3) Drug-Target Networks: Drug-target interaction networks can be considered as the most commonly used network model in the context of drug discovery. These network models are usually bipartite, with nodes consisting of both drugs and targets, and the links between nodes representing drug-drug similarity, known drug-target interaction or target-target similarity. The main focus of this network model is to identify potential targets for drugs by analyzing the pair-wise similarity between two drugs or two proteins [**?**], [**?**], [**?**], [**?**].

One of the earliest drug-target bipartite networks is the model proposed by Yildirim et al. [**?**], in which two proteins are connected to each other if they share at least one drug and two drugs are connected to each other if they share at least one target protein. Drugs and targets are connected with each other from known indications. The similarity in side-effects can be used as another measure to interconnect drugs in a drug-target bipartite network. Campillos, et al. [**?**] experimentally confirms that common protein targets for unrelated drugs can be inferred using side-effect similarity. One of the major drawbacks in this approach is that phenotypic side-effects cannot be obtained for novel drugs prior to clinical trials.

C. Discussion

Although a large number of similarity measures exists to connect drugs, targets and diseases, each bearing their own advantages, most models limit their use to a specific measure of similarity. To overcome the unreliability of using a single similarity measure, there are methods that combine several measures together and select the best possible measure from the feature set [**?**], [**?**]. Such methods, however, still do not address the problem of *unavailability of required data*. Thus, there is a need for novel methods that utilize multiple similarity measures interchangeably upon their availability.

III. Materials and Methods

In this section we describe the research methodology, data and the experimental setup used when proposing a solution to the problem of missing data in network based methods for drug action prediction.

A. Methodology

The initial step in identifying drugs that share a common MoA is the construction of a drug-drug network, in which the nodes represent drugs and the edges represent pair-wise distances calculated using a similarity measure. As discussed

in the previous section, the construction of drug networks from drug-drug similarity measures primarily require, *1)* selecting a suitable similarity measure; *2)* constructing distances between drug nodes; and *3)* interconnecting the nodes based on a distance threshold. To predict drugs with similar MoA, the constructed drug network should be clustered to obtain drug communities. Drugs falling in to the same community have a tendency to share the same MoA [**?**].

The step 2 of this process requires every drug under consideration to have the information necessary to calculate the similarity distances. For example, to construct a drug network using side effect similarity, only the drugs listed in side effect databases, such as SIDER [**?**], can be used. The major limitation introduced by these methods is the difficulty to obtain the information required by the model due to the fact that, some information, such as lists of side effects, is hard to obtain for new drugs that have not yet been tested at clinical trials. Our objective is to address this limitation of *missing data*.

The primary focus of our study lies on exploiting heterogeneous measures of similarity to approximate the *missing data* when constructing the drug network. Hence, we extend the primary steps in MoA prediction using drug network construction, which can be listed as follows.

1) Selecting suitable similarity measures.
2) Constructing pairwise similarity distances between drug nodes.
3) Approximating missing data.
4) Constructing the drug network using the novel, approximated, distances.

The final drug network is used to identify communities of drugs with similar MoA. Figure 1 provides a walk-through of the research design, which is comprehensively explained in this chapter.

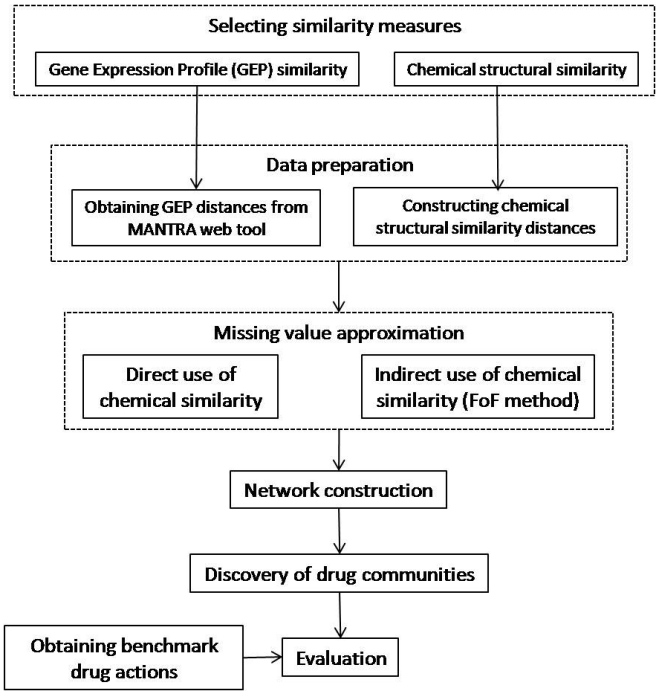

Fig. 1. A walk-through of the detailed research design.

B. Data and Experimental Setup

Since our primary focus lies on testing the suitability of heterogeneous measures, and not on introducing new similarity measures, methods for calculating similarity distances are directly obtained from the literature. Two similarity measures based on GEP and chemical structures, are initially chosen to test our approach for the approximation of missing values.

1) Similarity in GEPs: Drug-induced GEP (i.e. DNA microarrays) provide ample information to determine the genome-wide reactions that follow a drug-target binding [**?**]. Thus, the studies based on the assumption that drugs displaying similarities in transcriptional profiles may display similar MoA, have met with greater success [**?**], [**?**], [**?**]. Thus, transcriptional profiling is chosen as the primary similarity measure in our study, owing to the fact that the measure can be used independently to find drug actions.

MANTRA is an online drug network visualization and MoA detection tool constructed in a recent study by Iorio, et al. [**?**], which uses the similarity score between two gene signatures to build a drug-drug network. In this drug network, two drugs are connected with an edge, if their pair-wise similarity score is above a certain threshold. In our study, we directly obtain the pair-wise GEP distance values from MANTRA (http://mantra.tigem.it/), which consists of *31643* unique pair-wise distances corresponding to *1219* drugs.

2) Chemical Structural Similarity: When a pair of drugs have *highly similar* chemical structures, it is justifiable to hypothesize that they share similar drug targets [**?**]. Hence, our study utilizes chemical structures of drugs obtained from the publicly available drug repositories as the second pair-wise drug similarity measure.

SMILES is a *de facto standard* to represent the chemical structure of a molecule using ASCII strings [**?**]. It is widely used as a one-dimensional drawing of the chemical structure when the two- or three-dimensional representations are not appropriate. We acquired the SMILES representations for all the 1219 drugs from public data repositories, DrugBank [**?**], PubChem [**?**] and ChemBank [**?**], as well as from the related literature [**?**].

Once we obtained the chemical structures in SMILES format and converted them to fingerprints, we measured the similarity between two compounds using the *Tanimoto Coefficient*; one of the most commonly used metric in Cheminformatics to measure the similarity in molecules [**?**], [**?**]. Equation 1 was used for the Tanimoto calculation [**?**], where N_A denotes the number of bits (features) "on" in the bit string of structure A, N_B denotes the number of bits "on" in structure B and $N_{A \cap B}$ is the number of "on" bits common to both structures A and B.

$$T(A,B) = \frac{N_{A \cap B}}{N_A + N_B - N_{A \cap B}} \tag{1}$$

The data sets we acquired using the aforementioned methods are given below.

- L: List of all the drugs under consideration

- CS: An *N*N* matrix, where N is the number of drugs in list L. $CS_{a,b}$ contains the *chemical similarity* between drug A and drug B
- GS: An *N*N* matrix, where N is the number of drugs in list L (with some entries marked as null). $GS_{a,b}$ contains the *GEP similarity* between drug A and drug B

C. Missing Value Approximation

Although the GEPs are available for all the 1219 drugs present in the dataset, when introducing a *new drug* to the network, we cannot proceed further in the presence of missing gene expression profiles. In such a situation, the missing pair-wise distances of the new drug with all the other drugs should be either *approximated* or predicted by utilizing other heterogeneous information available for the drug under consideration. We address this problem by *indirectly* approximating the missing values using a **mediator drug** (described in the forthcoming section) which is selected based on the chemical structural similarity.

IV. Friends of Friends Approach for Similarity Approximation

Drug networks can be constructed using a variety of similarity measures; however, none of the measures complement the other. Although chemical similarity is highly available, it has been proven in the literature that the similarity in chemical structure cannot directly be used as the similarity in therapeutic action [**?**] or the similarity in drug targets [**?**]. This direct approximation fails due to the fact that, although high chemical similarity (tanimoto similarity score >0.8) can convey similarity in drug actions, the vice-versa does not necessarily hold true [**?**]: chemically dissimilar drugs can also display similar effects. Therefore, we devise the "Friends of Friends" method for similarity approximation by considering only the drug pairs with *high chemical similarities*.

A. Preliminary Hypothesis

Highly chemically similar drug pairs produce similar drug-induced gene expression profiles on the same cell line.

Figure 2 illustrates how our "**Freinds of Friends (FoF) approximation**" method infers the link between two drugs by assuming that the preliminary hypothesis is true.

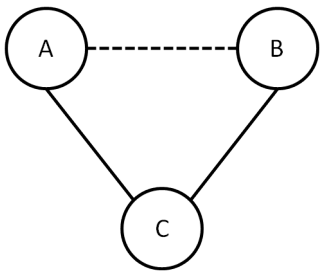

Fig. 2. Friends of friends approximation: If the link between drugs A and B is unknown, by taking C as the mediator which is chemically similar to B, we approximate the similarity between A-B as the similarity between A-C.

B. FoF Approximation

When the GEP of a new drug (B) is not known, the following algorithm is applied on the dataset to approximate the pairwise similarity of the new drug with all the other drugs in the network. For each drug pair (A-B), the first step is to approximate a mediator drug (C) and subsequently acquire the GEP similarity between A and C.

Algorithm IV.1: FoF($DrugListL, DrugB$)

$ListAL \leftarrow FoF(L, B)$
input: $GS[A]$, $CS[B]$, $list\ of\ drugs$
output: $Approximated\ List\ AL$

while $L.End$
do
- $A \leftarrow L.ReadLine()$
- $Min \leftarrow \infty$
- **for** $all\ C \leftarrow GS[A]$
  - **do if** $CS[B][C] > 0.6$
  - **then if** $GS[A][C] < Min$
  - **then** $Min = GS[A][C]$
- $GS[B][A] = Min$

return $(GS[B])$

C. From Similarity to Networks

We arranged the edges of the original GEP network in ascending order of their edge-weights, such that highly similar edges are ranked at the top, and applied a cut-off threshold of 0.59. Subsequently, we replaced the first k edges in an incremental manner (k = 1 to 105) using our approximated similarity scores, to test whether the network topology and the communities obtained will be affected by the slight variations in approximations.

D. Drug Community Identification

Although the number of edges starts to decrease when the approximation count exceeds 29, it does not cause a major effect on the network topology due to the fact that only a handful of links are discarded from the network. To prove this statement, we applied the Girvan-Newman clustering algorithm on the new network at each k approximations and obtained new drug communities. The 22 drug communities we acquired from the network at 50 approximations are illustrated in Figure **??**, where each community has an average of 7 drugs per community. Each new community is compared with the original set of drug communities identified by Iorio et al. [**?**] The validation of the drug communities obtained is discussed in the forthcoming section.

V. Results and Discussion

This section presents the results of the evaluation we have carried out on both approximated similarity scores and the communities we have obtained.

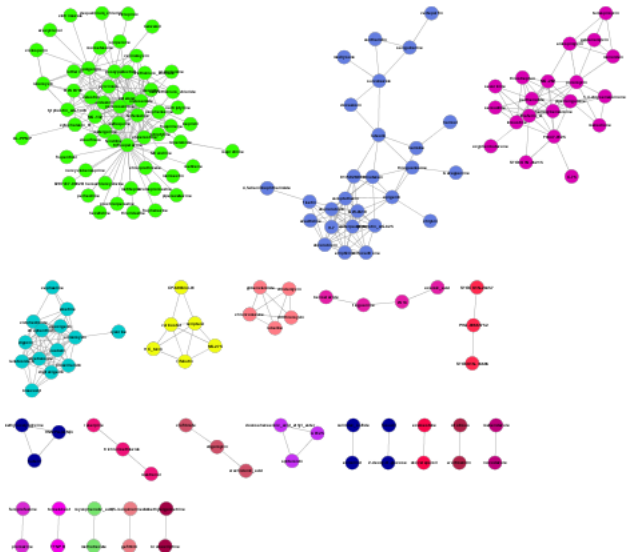

Fig. 3. Drug communities constructed by applying the Girvan-Newman clustering algorithm on the drug network obtained by approximating the top 50 edges. Number of clusters = 22.

A. Evaluation of the FoF Approach

We approximated GEP similarity scores for 3 randomly selected drugs from the drug list L: Helveticoside, Digoxin and Ouabain. When the approximations for all three drugs are compared against actual similarity distances using *simple linear regression* (Figure **??**), all three datasets displayed a high level of correlation (Pearson's correlation-coefficient: r) (Table **??**), with the equation **??** of the straight line.

$$y = \beta_1 x + \beta_2; \beta_1 = 0.996, \beta_2 = -0.0015 \tag{2}$$

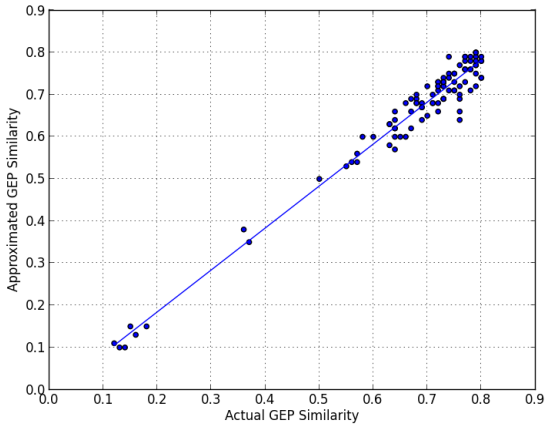

Fig. 4. Comparison of actual similarity scores with approximated distances using a simple linear regression model. Drug under consideration: Helveticoside, sample size = 102.

TABLE I
STATISTICS DERIVED BY APPLYING SIMPLE LINEAR REGRESSION ON THE APPROXIMATED VALUES OF THREE DRUGS.

Drug	**Size**	R^2	**Slope**	**Constant**	**r**
Helveticoside	103	0.967	0.996	-0.0015	0.984
Digoxin	62	0.963	0.966	0.0031	0.981
Ouabain	99	0.959	0.977	0.0008	0.979

B. Z-test

The difference between actual values and approximated values (d) should be minimized to have successful approximations. For a single drug B, 1218 pairwise similarities can be acquired from the chemical similarity matrix (CS). Hence, for each drug k, d is calculated for all the 1218 approximations, and the mean difference ($\overline{x_k}$) is calculated using equation **??**.

$$\overline{x_k} = \frac{\sum_{i=1}^{N} d_i}{N}; N = 1218, k = 1, .., 1219 \tag{3}$$

To test whether the approximated values represent actual values, we hypothesize that, for all the approximations, the difference between actual and estimated is equal to 0. This leads to the hypothesis, **the mean difference of the populations (μ_d) is equal to 0**. Thus, the *null hypothesis* and the *alternative hypothesis* are formed as, $\boldsymbol{H_0 : \mu_d = 0}$, $\boldsymbol{H_1 : \mu \neq 0}$. We chose z-statistics to test the null hypothesis with a 95% confidence interval. Standard error (S.E.) is calculated using equation **??**.

$$\boldsymbol{Standard\ Error = \frac{\sigma}{\sqrt{n}}} \tag{4}$$

The hypothesis was tested on 143 drugs, out of which, only 53 drugs failed to reject the null hypothesis. If the number of drugs tested is $\boldsymbol{N_A}$ and the number of drugs that failed to reject the null hypothesis is $\boldsymbol{N_{H0}}$, using equation **??** we arrive at an accuracy 37.6% for the 95% confidence interval test.

$$\boldsymbol{Accuracy = \frac{N_A}{N_{H0}}} \tag{5}$$

C. Reasons for low accuracy

We have identified four possible reasons that contribute to the low accuracy levels at the 95% confidence interval hypothesis test. Out of them, the most contributing factor, which suggests that the preliminary hypothesis only works for some specific drugs, is described in this section in detail.

1) Finding Possible Drug Classes: Out of the drugs that failed to reject the hypothesis, we further conducted experiments to find any overrepresented drug classes. The class of each drug is taken as its ATC code prefix [**?**]. As depicted in Figure **??**, out of the drugs that failed to reject the hypothesis, the highest percentage of drugs belong to the class C (Cardiovascular System), and the second highest is the class J (Antiinfectives for Systemic Use). However, it is worth noting that the drugs that rejected the hypothesis have similar distributions of the classes C and J due to the over-representation of drugs belonging to these two classes in our

complete dataset. Hence, with the use of ATC therapeutic categories, we could not find a drug class which specifically agrees or disagrees with our preliminary hypothesis.

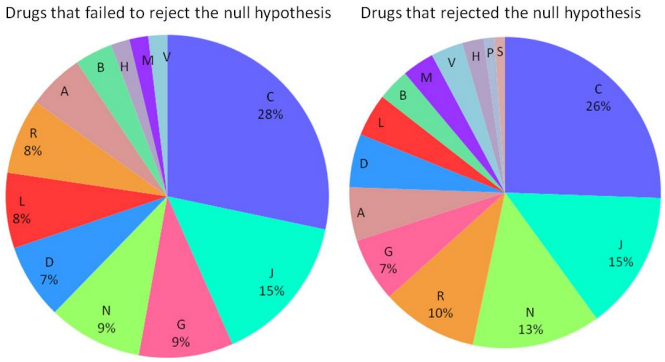

Fig. 5. Class distribution of drugs that failed to reject the null hypothesis (left) and drugs that rejected the null hypothesis (right).

2) Filtering Out Highly Similar Drugs: Owing to the fact that we could not find a specific therapeutic category which significantly complies with our preliminary hypothesis, we conducted the "0.4 threshold test" in order to distinguish the types of similarity links for which our Friends of Friends approach works the best. The steps we carried out for this experiment are:

1) Prepare $\boldsymbol{List_{AcS}}$: For actual pair-wise GEP similarity scores, obtain a list in ascending order such that the drug pair with the highest actual GEP similarity is ranked at the top. Filter out drug pairs with the similarity score (s) >0.4. The final list we obtained, consists of 63 drug pairs (out of 31643).
2) Prepare $\boldsymbol{List_{ApS}}$: Repeat step 1 for approximated pair-wise GEP similarity scores. Final list consists of 105 drug pairs.
3) Compare drug pairs in $\boldsymbol{List_{AcS}}$ with the drug pairs in $\boldsymbol{List_{ApS}}$.

The "0.4 threshold test" applied on $\boldsymbol{List_{ApS}}$ yielded a precision rate of 50.0% and a recall rate of 82.54% when compared with $\boldsymbol{List_{AcS}}$.

D. Validating drug links on therapeutic similarity

Initially, we validated the pair-wise drug-drug similarity scores by matching the ATC code of the two drugs in a pair. If the drug class (or the subclass) is similar in two drugs, we consider the pair, a true positive.

Positive predicted value (PPV) graph in Figure **??** is obtained by comparing the ATC code of drugs in a pair, and gradually increasing the number of drug pairs under consideration. Equation **??** is used to calculate the PPV at each step. For the 1-length prefix, the similarity between ATC codes of the drug pairs perfectly matched up to 20 top-ranked drug pairs.

$$\boldsymbol{PPV = \frac{TP}{(TP + FP)}} \tag{6}$$

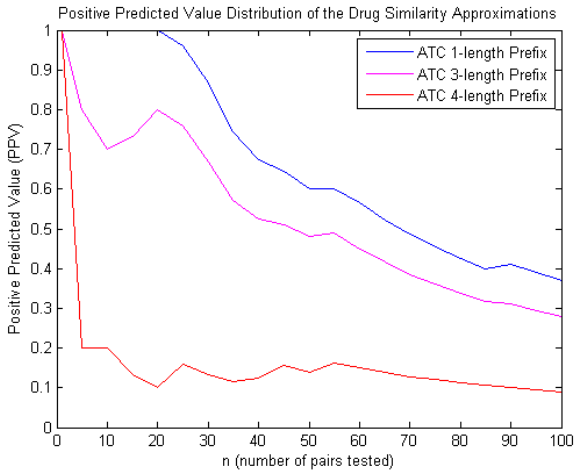

Fig. 6. Rate of decrement in the number of edges of the new drug network with respect to the number of similarity approximations.

E. Validating drug communities

In the previous section we described the construction of the new drug network by approximating the top k edge-weights in an incremental manner (k = 1, 15, 30, 50, 75, 105), and the identification of drug communities that possibly share a mode of action.

We compared each drug community with the original set of communities from the literature and assigned a scalar value (i.e. Jaccard Coefficient [**?**]) to represent the degree of the overlap between two drug communities. The scale of the Jaccard Coefficient is such that, 1.0 represents a one to one overlap and 0.1 represents a very poor overlap. During the comparison of two communities, we pronounce two communities as overlapping if the Jaccard Coefficient exceeds 0.7. If the number of overlapping communities are referred to as true positives, the precision and recall rates at each iteration k can be easily approximated. Table **??** provides the precision and recall rates, as well as the average Jaccard Coefficient values of the overlapping communities, which provides an indication of the overlap between the communities obtained from the new network and the communities obtained from the original network.

TABLE II
ANALYSIS OF THE DRUG COMMUNITIES.

Network	Edges	Cluster Count	Matching clusters	PPV	Recall	Avg Jaccard
K=15	511	22	22	1.0	1.0	1.0
K=30	510	22	22	1.0	1.0	1.0
K=50	508	22	22	1.0	1.0	1.0
K=75	501	22	22	1.0	1.0	0.99
K=105	485	20	19	0.95	0.86	0.97

F. Observations

GEP similarity distances approximated using the FoF method is only 37.6% accurate if we consider a one-to-one

mapping of actual scores and approximated scores. Nevertheless, the "0.4 threshold test", which filters out drug pairs with low similarity scores ($s > 0.4$), yielded high recall rates (82.54%). Thus, we can conclude that our FoF method is better at approximating unknown pair-wise GEP similarity of **highly similar drugs**. Furthermore, we also carried out the test by taking chemical similarity scores directly as the GEP similarity distances, which yielded low precision (0.74%) and recall rates (52.38%), confirming that our results are not achieved by random chance or the biasness of the dataset.

Furthermore, if we consider the validation of the drug communities, the precision rate of 1.0 for k = 15, 30, 50 and 75 indicates that, although the number of edges have decreased at each iteration, the community formation of the network was not affected by this change. This provides further evidence that the FoF approximation works best for highly similar drug pairs.

VI. Conclusion and Future Work

In this study, we attempted to gain an understanding of the unknown modes of action of drugs by relying on the assumption that *similar drugs share similar effects*. We built on existing models that interconnect drugs based on their pair-wise similarity, where we specifically addressed the issue of *missing data* in the presence of a single measure of similarity. The objective of our work was to exploit the advantage of heterogeneous measures of similarity to avoid the limitation of a single measure. Hence, we devised the *Friends of Friends approach for similarity approximation*, which exploits the advantages in multiple similarity metrics when deriving the missing data.

To demonstrate our suggested approach, we conducted a pilot study by obtaining drug-drug similarity distances based on *highly reliable* GEPs and approximating the missing scores based on *readily available* chemical structures. The preliminary hypothesis behind our approach is that "highly chemically similar drug pairs produce similar drug-induced gene expression profiles on the same cell line". Subsequently, by combining the actual and approximated similarity scores, we constructed a drug network, which was further analyzed to identify communities of drugs that share a specific MoA.

In this work, we have used the FoF method to approximate missing information only on a GEP similarity based drug-drug network. As future work, we propose to test the applicability of this method across several other similarity-based models, which we have discussed in section II, such that its applicability will not be limited by the model used. For instance, we can further test the suitability of using chemical structural similarity to approximate the side effect similarity of drug pairs [**?**]. If the cross applicability is ensured, the FoF approximation can be provided as a platform to build drug networks with multiple similarity measures.

Acknowledgment

We thank Dr. Dilhari Attygalle for providing the statistical insight and Dr. Chandanie Wanigatunge for providing the pharmacological background.

References

[1] I. Kola and J. Landis, "Can the pharmaceutical industry reduce attrition rates?" *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 711–716, 2004.

[2] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nature reviews. Drug discovery*, vol. 9, no. 3, pp. 203–214, Mar. 2010.

[3] N. P. Tatonetti, T. Liu, R. B. Altman *et al.*, "Predicting drug side-effects by chemical systems biology," *Genome Biology*, vol. 10, no. 9, p. 238, 2009.

[4] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery." *Nature chemical biology*, vol. 4, no. 11, pp. 682–90, Nov. 2008.

[5] Y. Tamada, S. Imoto, K. Tashiro, S. Kuhara, S. Miyano *et al.*, "Identifying drug active pathways from gene networks estimated by gene expression data," *Genome Informatics Series*, vol. 16, no. 1, p. 182, 2005.

[6] P. Agarwal and D. B. Searls, "Literature mining in support of drug discovery," *Briefings in Bioinformatics*, vol. 9, no. 6, pp. 479–492, Nov. 2008.

[7] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drugtarget interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[8] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, "Combining drug and gene similarity measures for drug-target elucidation," *Journal of computational biology*, vol. 18, no. 2, pp. 133–145, 2011.

[9] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Globalized bipartite local model for drug-target interaction prediction," in *Proceedings of the 11th International Workshop on Data Mining in Bioinformatics*, ser. BIOKDD '12. New York, NY, USA: ACM, 2012, pp. 8–14.

[10] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, 2008.

[11] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi *et al.*, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proceedings of the National Academy of Sciences*, vol. 107, no. 33, pp. 14 621–14 626, 2010.

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

[13] B. Chen, Y. Ding, and D. J. Wild, "Assessing drug target association using semantic linked data," *PLoS Comput Biol*, vol. 8, no. 7, p. e1002574, Jul. 2012.

[14] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a highly connected world*. Cambridge Univ Press, Jun. 2010, vol. 8.

[15] F. Iorio, R. Tagliaferri, and D. di Bernardo, "Identifying network of drug mode of action by gene expression profiling," *Journal of computational biology:a journal of computational molecular cell biology*, vol. 16, no. 2, pp. 241–51, Feb. 2009.

[16] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine." *Molecular systems biology*, vol. 7, 2011.

[17] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PLoS ONE*, vol. 4, no. 8, p. e6536, Aug. 2009.

[18] M. Kissa, M. Scroeder, and G. Tsatsaronis, "Towards an integrated compound to compound relatedness measure," in *ECCB-ISMB 2013 BioLINK SIG (BioLINK 2013)*, Jul. 2013.

[19] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biol*, vol. 6, no. 1, Jan. 2010.

[20] A. B. Parsons, A. Lopez, I. E. Givoni, D. E. Williams, C. A. Gray, J. Porter, G. Chua, R. Sopko, R. L. Brost, C.-H. Ho *et al.*, "Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast," *Cell*, vol. 126, no. 3, pp. 611–625, Aug. 2006.

[21] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *Science Signaling*, vol. 313, no. 5795, p. 1929, 2006.

[22] D. Weininger, "Smiles, a chemical language and information system," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[23] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D668–D672, 2006.

[24] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "Pubchem: a public information system for analyzing bioactivities of small molecules," *Nucleic Acids Research*, vol. 37, no. suppl 2, pp. W623–W633, 2009.

[25] K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber, and P. A. Clemons, "Chembank: a small-molecule screening and cheminformatics resource database," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D351–D359, 2008.

[26] J. J. Babcock, F. Du, K. Xu, S. J. Wheelan, and M. Li, "Integrated analysis of drug-induced gene expression profiles predicts novel herg inhibitors," *PLoS ONE*, vol. 8, no. 7, p. e69513, Jul. 2013.

[27] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "Stitch: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D684–D688, 2008.

[28] P. Baldi and R. Nasr, "When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values," *Journal of chemical information and modeling*, vol. 50, no. 7, pp. 1205–22, Jul. 2010.

[29] P. Willett, "Similarity-based virtual screening using 2d fingerprints," *Drug Discovery Today*, vol. 11, no. 2324, pp. 1046–1053, 2006.

[30] A. Skrbo, B. Begović, and S. Skrbo, "[Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes]," *Medicinski arhiv*, vol. 58, no. 1 Suppl 2, pp. 138–41, 2004.

[31] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bulletin de la Société Vaudense des Sciences Naturelles*, vol. 44, pp. 223–270, 1908.