International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

A Framework for Predicting Drug Target Interaction Pairs Through Heterogeneous Information Fusion

Ansa Baiju, Juliet Johny, Linda Sara Mathew

Abstract: ***Drugs, also known as medicines cure diseases by interacting with some specific targets such as proteins and nucleic acid. Prediction of such drug-target interaction pairs plays a major role in drug discovery. It helps to identify the side effects caused by various drugs and provide a way to analyze the chances of usage of one drug for various diseases apart from the one disease that is predefined for that drug. However, existing Drug Target Interaction prediction methods are very expensive and time consuming. In this work, we present a new method to predict such interactions with the help of bipartite graph, which represents the known drug target interaction pairs. Information about drug and target are collected from various sources and they are integrated using Kronecker Regularized Least Square approach and Multiple Kernel Learning method, to generate drug and target similarity matrices. By integrating the two similarity matrices and known DTIs a heterogeneous network is constructed and new DTI predictions are done by performing Bi Random walk in it.***

Keywords : Drug Target Interaction, Heterogeneous network, Multiple Kernel Learning, KronRLS, Bipartite graph.

I. INTRODUCTION

Since the numbers of diseases are increasing day by day, the relevance of medicines/drugs are very high and it became an unavoidable factor in human life. Drugs are chemicals that can be used to cure and prevent various diseases. They do so by interacting with some targets in our body such as ion channels, enzymes, receptors, amino acids etc. This is named as Drug Target Interactions (DTI) [1]. Targets are some types of molecules that are present in living organisms which are continuously modified by drugs. Each of the drugs have some well-known, specified targets that they are interacting with. For example, albuterol is a medicine which is used by asthma patients for respiratory related issues. It widens or open up the

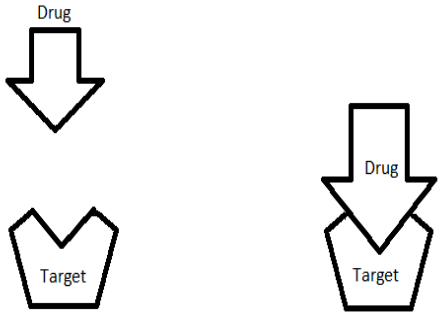

Fig. 1. Drug Target Interaction

Airways in lungs by attaching to some specific receptors, which is one of the well-known biological targets. Fig. 1 shows how drugs interact with targets. If the drugs need to be effective, it must attach with some targets. The chemical compound that is present in the drug will attach with the target and it will create some effects and later leaves the target.

Apart from the well-established and already available DTI, there can be hundreds and thousands of unknown drug target pairs. Identifying those drug target pairs are very important. Usually a drug D interacts with a target T1 in order to cure a disease S1. If this drug interacts with another target T2, then rather than curing the disease, there can be chances of some side effects. Side effects are the additional unwanted effects caused by the consumption of drugs. So by predicting a new DTI, the side effects that could be caused by the drugs can be identified [2]. Also if the drug D interacts with another target T2, then there can be chances that we could use the same drug for the treatment of another disease S2. In older days, the main idea of drug discovery was, one drug interact with one target to cure one disease. Nowadays, this concept is replaced with the idea that, one drug interacts with multiple targets to cure multiple diseases [3]. This has become a greater idea in drug discovery and it is termed as drug repositioning [4]. An example of drug repositioning is the case of raloxifene, a drug which was earlier used for the treatment of breast cancer. Later it was repositioned against osteoporosis, a bone disease that occurs in elder women which leads to easily breakage of bones. In short, main aims of DTI prediction includes:

- Identify side effects of drugs
- Identify chances of usage of a drug for various diseases

In order to identify new DTI pairs, generally two methods are there.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Ansa Baiju*, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. E-mail:ansabaiju96@gmail.com
Juliet Johny, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. E-mail:julietjohny5@gmail.com
Linda Sara Mathew, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. E-mail:lindasaramathew@gmail.com

Ligand based approach [5] and target based approach [6]. These are the widely used methods for discovering new drugs. Ligand are some specific molecules that attach to the biological targets. We have knowledge about the complete features of some already available ligands that are interacting with the ligands of targets and such information are used for the drug discovery. This is best when structural information of targets are unavailable. On the other hand, target based approach is used when the three dimensional structure of the target is available.

The rest of the paper is as follows. Section 2 describes the related works in this area. Section 3 describes the proposed work and the last section concludes the paper.

II. RELATED WORKS

Kevin Bleakley proposed a new approach [7] to predict unknown DTI from chemical information and genomic sequence. The method is considered as a bipartite graph interference problem. All the problems which predict new drug compounds, new target compounds which are associate to target protein. The known bipartite local model predict the protein target for a given drug and also predict the drug target for the given protein. By training the local models the bipartite graph interference problem is solved and thus it predict a new edge linking drug nodes with the target nodes. The prediction is done by using support vector machine. In 2014, Ding H [8] used an machine learning approach and assume a similarity based method, ie, similar drug shares similar target. This approaches integrate the two types of similarities and identify the potential DTI. Methods like SVM, KNN and random forest classifier are used to predict the DTI. In the approach, it only uses the positive samples or known DTI in experiments. A supervised interference method [9] is used for predicting the new DTI and drug repositioning. The method is derived from recommendation algorithm. To predict the interaction, different inference methods are developed: Drug-Based Similarity Interference (DBSI), Target Based Similarity Interference (TBSI). Firstly, a bipartite network can create with known DTI data. Then NBI method is used to predict the new DTI in the drug target bipartite network.

Yan X [10] implement an method of network based label propagation with mutual interaction information derived from network and predict some novel DTI. It performs label propagation on drug and target similarity network. The label propagation on each network is assign an cluster structure. The label information from the other network is seen as mutual interactions. The drug and target network construct an heterogeneous network. A Kronecker Regularized Least Square – Multiple Kernel Learning (KronRLS-MKL) method [11] is developed for the DTI problems. It extend the the KronRLS method and utilized the concept of MKL. It integrates the drug and target kernels which indicates drug and target similarity matrices respectively. By integrating drug and target space, a drug-target space is created which can be further used for prediction. In [12] a DTI method is proposed to predict potential DTIs which can further used for drug repurposing. From the heterogeneous data source, we get the existing drugs by integrating diverse information of

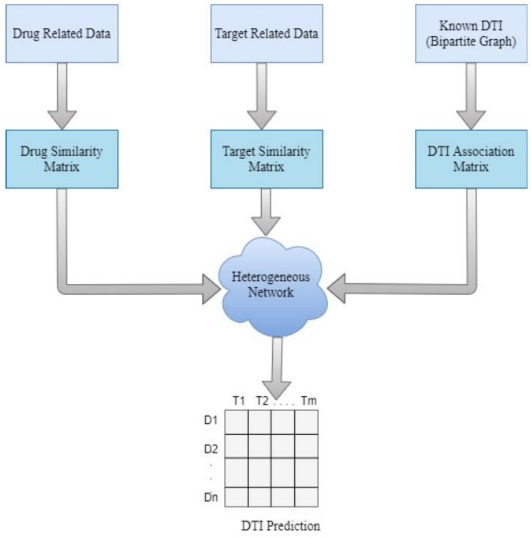

Fig. 2.Proposed Model

Drug and target. It also represents the drug and target by learning the low dimensional feature vector and also finding an best projection from drug space onto target space. In these approach it does not considered the cluster of network nodes and also does not recognize the effect of low similarity value.

III. PROPOSED WORK

The proposed model is a DTI prediction system that aims at providing the results through MKL. It mainly includes four steps. In-formation about drug and targets are collected from various sources such as Toxicogenomics database, Drug-Bank database, SIDER database, and Gene Ontology dataset. From the data collected first of all, drug and target similarity matrices are created. Various drug similarity matrices are integrated to form a single drug similarity matrix and target similarity matrices are integrated to form a single target similarity matrix using KronRLS method and MKL. With the help of the drug similarity matrix, target similarity matrix and available DTI pairs, a heterogeneous network is constructed and new drug target pairs are predicted by performing Bi-random walk on the network. The major idea is that, similar drugs have higher chances of interacting with the same target [13]. Fig. 2 shows the overall proposed model.

A. Construction of Similarity Matrices

In order to measure the similarity between two drugs D1 and D2, and two target T1 and T2, based on their side effects, disease relationships and gene ontology relationships, Jaccard
similarity measure is used. Thus drug-disease similarity matrix, drug-side effect similarity matrix, target-disease similarity matrix, target-sequence similarity matrix and target-GO similarity matrices are created.

$$S(D_i, D_j) = \left| \frac{A_i \cap A_j}{A_i \cup A_j} \right| \quad (1)$$

B. Information Fusion

KronRLS is the best approach for learning from graph based data. It consider the DTI prediction as a link prediction problem of the bipartite graph. Bipartite graph is a type of graph whose vertices can be represented as two disjoint sets and every edge connects between them. In the proposed model vertices indicates drugs and targets and edge indicates known interaction between a drug and target. Fig.3 shows an example of bipartite graph and its corresponding association matrix.

Let D={D1,D2,. . . Dn} be a set of drugs and T={T1,T2,. . . .Tm} be a set of targets. Known DTI pairs are extracted from the bipartite graph and those set of training inputs are indicated by Xi (Known DTI pair). Let Y be their labels.

Y={1, if there is an edge from Di to Ti , 0, else}

Regularized Least Square (RLS) [14] method which is used to solve the least square regression problem. Least square regression problem is to find the best line or curve (regression line or curve) to fit some sets of variables so that they can be classified to the predefined labels. In our case, we have two labels. One is a DTI pair and the other one is not a DTI pair. Vertical distance from the data points to the line or curve should be as low as possible. A line or curve which satisfies this is the best one which reduces the overfitting. Overfitting, which is also termed as variance is the sum of the squares of the errors. It is an error that occurs when our model is too adaptive to the training data. Regularization is a method used to avoid overfitting. This is done by adding a regularization parameter (λ > 0) to the loss function. RLS minimizes the following, since loss function, which is the difference between predicted and actual value should be minimum.

$$J(f) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_K^2 \quad (2)$$

KronRLS is applied when training data is paired input such as known DTI. Feature representations of the inputs (drug and target) are given as two kernel matrices, KD and KT. A pairwise kernel is created by the kronecker product of these two base kernels. K = KD ⊗KT . This is done to combine the drug space and target space into drug-target space for efficient
prediction.

In our scenario, since the data is collected from various sources, rather than a single kernel represent the drug space, we have multiple kernels (similarity matrices). The same applies for target also. Here comes the importance of MKL [15], which is a flexible learning model that uses some kernels for the learning process. It is best when we need to combine data from various sources. It allows us to create kernels for each of the source. Kernels usually represent a notation of similarity of the data. We can combine the base kernels in a linear optimization manner and can generate a single kernel. Grouping kernels is one way to combine different data. Since each of the drug and target similarity matrices are considered as kernels, they can be joined together as follows. βi indicates

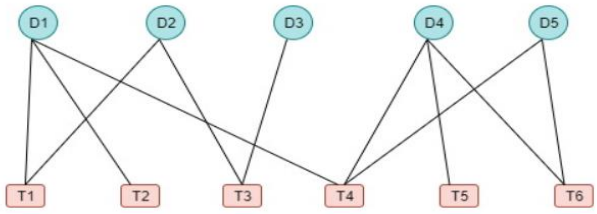

Fig. 3.Bipartite graph

Table- I: Associatin Matrix

D	T1	T2	T3	T4	T5	T6
D1	1	1	0	1	0	0
D2	1	0	1	0	0	0
D3	0	0	1	0	0	0
D4	0	0	0	1	1	1
D5	0	0	0	1	0	1

The weight of the kernel Ki and n is the number of kernels used.

$$K=\sum_{i=1}^{n}\beta_i K_i \quad (3)$$

With the help of the above equation, optimal drug kernel and target kernel can be created by combining the drug and target similarity matrices. Let K_D^* be the optimal drug kernel and K_T^* be the optimal target kernel.

$$K_D^* = \beta_1 K_1 + \beta_2 K_{2+\cdots+}\beta_n K_n \quad (4)$$

$$K_T^* = \beta_1 K_1 + \beta_2 K_{2+\cdots+}\beta_m K_m \quad (5)$$

N and m are the total number of drug and target kernels respectively. Since kernels usually represent a notation of similarity of the data, the above equations can be re-written in such a way that, they represent the ultimate drug and target similarity matrices after the information fusion from heterogeneous sources. S_D is the drug similarity matrix and S_T is the target similarity matrix.

$$S_D = \beta_1 S_1 + \beta_2 S_{2+\cdots+}\beta_n S_n \quad (6)$$

$$S_T = \beta_1 S_1 + \beta_2 S_{2+\cdots+}\beta_m S_m \quad (7)$$

Our assumption is that drugs having higher similarity have higher chances of interacting with the same targets. Similarly, targets having higher similarity have higher chances of interacting with the same drug. From this concept, drugs having very lower similarity and targets having very lower similarity are less significant. So they can be neglected and replaced with 0.

C. Heterogeneous Network Construction

Based on the similarity matrices obtained in the previous stage, two networks are created which represents the drug and target similarity. This drug similarity network and target are integrated with the known DTI pairs from the bipartite graph to form a heterogeneous network.

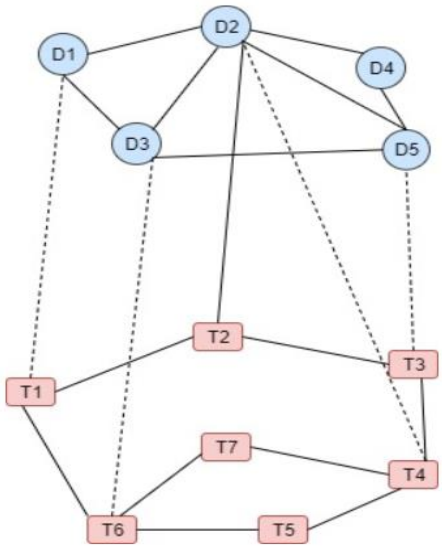

Fig. 4.Heterogeneous Network

Fig. 4 shows how a heterogeneous how a heterogeneous network will look like. Dark line between drug D2 and target T2 indicates that there exist a known interaction between these two and similarly, the dotted line indicates no interaction between them.

D. Bi-Random Walk

Bi-random walk algorithm [16] performs random walks on the drug similarity network and target similarity network simultaneously. The main aim is to increase the number of drug-target association or paired association between the two networks. Let D_{n*n} be the drug similarity matrix and T_{m*m} be the target similarity matrix where n and m are the total number of drugs and targets respectively. A_{m*n} is the association matrix of known DTIs obtained from the bipartite graph. q and r are the number of maximum iterations on drug and target network. A new association matrix, R_{m*n} is created where Rij indicates the degree of association between drug i and target j. R_D indicates random walk on drug network and R_T indicates random walk on target network.

$$R_D = \alpha * S_D * R_{D-1} + (1-\alpha)A \quad (8)$$

$$R_T = \alpha * S_T * R_{T-1} + (1-\alpha)A \quad (9)$$

α is a decay factor which can take values {0,1}. A new association is evaluated by its distance to the known association.

IV. RESULT

The accession numbers of the drugs such as DB00357, DB02721 etc. are used for our study. Similarly, UniProtIDs such as, P05108 are represented as targets. DB00773 is named as Etoposide which is an anti-cancer chemotherapy drug. P05108 is the UniProtID of a target named cholesterol side-chain cleavage enzyme. Fig.5 shows the known DTI pairs. For example drug DB00357 and target P05108 is already a known DTI pair. It indicates that drug DB00357 is already interacting with target P05108 in order to cure some diseases.

	A	B	C
1	Drug	Gene	
2	DB00357	P05108	
3	DB02721	P00325	
4	DB00773	P23219	
5	DB07138	Q16539	
6	DB08136	P24941	
7	DB01242	P23975	
8	DB01238	P08173	
9	DB00186	P48169	
10	DB00338	P10635	
11	DB01151	P08913	
12	DB01244	P05023	
13	DB01745	P07477	
14	DB01996	P08254	
15	DB04800	P18031	
16	DB08352	Q16539	
17	DB00133	P21549	
18	DB00163	P21266	
19	DB00197	P10632	
20	DB06777	P08684	
21	DB01151	P10635	
22	DB00356	P08684	
23	DB01589	P34903	
24	DB01272	P20645	
25	DB08846	Q14534	

Fig. 5.Known DTI Pairs

By utilizing the concepts of MKL and KronRLS approach, two similarity matrices are created representing drug and target. Fig. 6 shows the drug similarity matrix. Similarly, a target similarity matrix is also created. Fig. 7 shows the obtained DTI prediction output. It indicates the percentage of interaction between a specified drug, DB08846 and all other targets. For example, the interaction between drug DB08846 and target Q96199 is predicted as 91.44%. It shows that there exist high chances of interaction between these two and can be considered as a strong DTI pair in the future drug discovery.

Similarly, drug DB08846 and target P35247 have less chances of interaction between each other, which can be considered as a very weak DTI pair.

Fig. 6.Drug-Similarity Matrix

Fig. 7.DTI Prediction

V. CONCLUSION

DTI is an interesting area in the drug discovery. Drug discovery is a method which identify new drugs and their new targets. A novel method is proposed to predict unknown drug-target interactions. Prediction of new DTI pairs is very important Pharmacogenomics and drug discovery. It can be used to identify the side effects of existing drugs and with the help of this we can identify the chances usage of a drug for various diseases. Several methods are already available for predicting new DTI pairs. But most of them are very expensive and time consuming. Also some of the exiting methods only use limited number of heterogeneous data sources in order to predict the DTIs. So an efficient and reliable method is required for the prediction of DTIs. Proposed method integrates the drug-related and target-related heterogeneous information from various sources and aggregate them using MKL. The main idea of the method is to form drug and target similarity matrices by integrating the drug related data and target related data using KronRLS. These similarity matrices are adjusted in such a way that drugs or targets having low similarity value is neglected. From the drug and target similarity matrices, drug and target similarity networks are constructed respectively. Then these similarity networks of drug and target are integrated with known DTI bipartite graph to form a drug target heterogeneous network. Finally, bi-random algorithm is implemented on the network and identify the new DTI. The model performs best in identifying new DTI pairs and it offer great accuracy.

REFERENCES

1. S Anusuya, M Kesherwani, K. V. Priya, A. Vimala, G. Shanmugam, D.Velmurugan, and M. M. Gromiha, "Drug-Target Interactions: Prediction Methods and Applications," Current protein peptide science, vol. 19, p. 537, 2018-01-01 2018.
2. Korotcov, A., Tkachenko, V., Russo, D. P. Ekins, S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery datasets. Mol. Pharm. 14, 4462–4475 (2018).
3. J. L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker, and R. A. Houghten, "Shifting from the single to the multi-target paradigm in drug discovery," Drug discovery today, vol. 18, no. 9-10, pp. 495–501, 2013.
4. Cheng, F., Liu, C., Jiang, J, 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput. Biol. 8 (5), 357–372.
5. Yamanishi,Y.,Araki,M.,Gutteridge,A.,2008.Prediction of drug target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24 (13), i232–i240.
6. Klabunde, T., Hessler, G., 2002. Drug design strategies for targeting G-Protein-Coupled receptors. Chembiochem 3 (10), 928–944.
7. Kevin Bleakley and Yoshihiro Yamanishi,2009 Supervised prediction of drug–target interactions using bipartite local models Bioinformatics Vol. 25 no. 18 2009, pages 2397–2403.
8. Ding,H.,Takigawa,I., Mamitsuka, H., 2014. Similarity-based machine learning methods for predicting drug–target interactions : a brief review. Brief. Bioinformatics 15(5),734–747.
9. Cheng,F.,Liu,C.,Jiang,J., 2012.Prediction of drug-target interaction and drug repositioning via network-based inference. PLoS Comput. Biol.8(5),357–372.
10. Yan,X.-Y.,Zhang,S.-W.,Zhang,S.-Y.,2016. Prediction of drug–target interaction by label propagation with mutual interaction information derived from heterogeneous network.Mol.Biosyst.12(2),520–531.
11. Nascimento,A.C.,Prudˆencio,R.B.,Costa,I.G.,2016. A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinformatics17(1),1.
12. Luo,Y.,Zhao,X.,Zhou.J.,2017.A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information.Nat.Commun.8(1).
13. Luo, H., Wang, J., Li, M., 2016. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics 32 (17), 2664.
14. J. Hainmueller, C. Hazlett, "Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach", Political Anal., vol. 22, no. 2, pp. 143-168, 2014.
15. Mehmet Gonen and Ethem Alpaydin. Multiple Kernel Learning Algorithms. Journal of Machine Learning Research (JMLR), 12:2211–2268, 2011.
16. Chen,X.,Liu,M.-X.,Yan,G.-Y.,2012.Drug–target interaction prediction by random walk on the heterogeneous network. Mol. Biosyst. 8 (7), 1970–1978.

AUTHORS PROFILE

Ansa Baiju received Bachelor of Technology in Computer Science and Engineering from Younus College of Engineering, Kottarakkara in 2018 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Machine Learning and Data Mining.

Juliet Johny received Bachelor of Technology in Computer Science and Engineering from Amal Jyothi College of Engineering, Kanjirappally in 2017 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Big Data Analytics and Data Mining.

Linda Sara Mathew is currently working as an assistant professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. She received her B-Tech Degree in Computer Science and Engineering in 2002 from Mahatma Gandhi university and M-Tech in Computer Science and Engineering from Anna University, in 2011. She has around 15 years of teaching experience. Her research interest include Data Mining, Neural Network, Image Processing and Soft Computing.