Drug-Drug Interaction Analysis Using Heterogeneous Biological Information Network

Kyubum Lee, Sunwon Lee, Minji Jeon,
Jaehoon Choi, and Jaewoo Kang*
Korea University
Seoul, Korea
{*kyubumlee, sunwonl, mjjeon, jaehoon, kangj*}
@korea.ac.kr
**Corresponding author*

***Abstract*—As the number of drugs increases, more prescription choices are available for physicians, and consequently the number of drugs administered together has increased. Researchers are working on finding multi-drug prescriptions that are effective and safe. An efficient method for finding DDIs plays a crucial role in this research.**

In order to address the problem, we construct a heterogeneous biological information network by combining multiple different databases and interaction information. Our network includes the information about genes, proteins, pathways, drugs, side effects, targets and their interactions. We propose a metric to measure the relation strength between two nodes in the network, which is based on the weighted sum of the numbers of paths containing different interaction types. We use the metric to score DDI candidates.

We found that the drugs sharing a disease are more likely to have a DDI than the drugs sharing a biomolecular target, and the metric using the weighted sum of the path numbers is effective to rank the potential DDIs. We validated the result with the PharmGKB DDI dataset and the Drugs.com drug interaction checker.

***Keywords*-DDI; drug interaction; biological network; systems pharmacology**

I. Introduction

Drug-drug interaction (DDI) is the reaction between two different drugs where one drug inhibits (antagonistic) or prohibits (synergistic) the effects of another drug directly or indirectly. An antagonistic DDI can reduce the effectiveness of drugs, whereas a synergistic DDI can cause overdose, over-reaction of biological systems, or serious side-effects.

As a result of more drugs being developed, there are more prescription choices for physicians, thus the number of drugs administered with other drugs has increased. Treatments for diseases like depression or diabetes and antiretroviral drugs for human immunodeficiency virus (HIV) are examples of multi-drug treatments that doctors prescribe the most [1]. Researchers are working on finding multi-drug prescriptions which are effective and safe without causing side-effects. An efficient method for finding DDIs plays a crucial role in this research.

Unfortunately, however, because of the overwhelming number of combination of drugs, it is almost impossible to consider all the possible pairs of drugs that are currently produced and prescribed. The number of US Food and Drug Administration (FDA)-approved drugs already amount to 1,787 that are chemically different, and their possible pairs surpass 1.5 million [2]. Since there are many more drug candidates, it is infeasible to do *in vitro* or *in vivo* experiments for all the possible pairs.

Another limitation of DDI research is the high costs. Performing all the experiments mentioned above for all the possible pairs of drugs is prohibitively expensive.

Because of these difficulties, many researchers have focused on *in silico* analyses for prescreening the DDI candidates in order to reduce the search space. Analyses of cohort studies were conducted but the number of findings was very limited [3]. There were efforts to find chemical similarity of drugs, or binding properties of drug pairs [1]. However, these efforts are not effective to find an indirect interaction that is caused by a remote chain reaction.

For the investigation of the indirect interactions, biological network information can be utilized. The advancement of medical science, molecular biology, and pharmacology made it possible to cumulate a sufficient amount of information on biomolecules and their interactions.

In order to address the above mentioned problems, in this work, we construct a heterogeneous biological information network using proteins, genes, drugs, pathways, side-effects, diseases, and the interactions between these biological objects. We use this network as an *in silico* test bench to find DDI candidates, and report important features for discovering DDIs. We also suggest a metric to measure the indirect interaction distance between two different drugs, which is an essential tool for facilitating the systematic prediction of a potential DDI.

We identify our contributions as follows:

1) We construct one of the largest heterogeneous biological information networks which contains various types of interactions between large numbers of heterogeneous

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

objects. We made the network available for research use at " http://infos.korea.ac.kr/ddianalysis ".

2) We study the properties of the network and investigate the characteristics of the network topology between two drugs known to have a DDI. Through this investigation, we identify the important network features for DDI prediction.

3) Finally, using the identified network features, we formulate a metric to measure the interaction distance between two DDI candidate drugs. We demonstrate the effectiveness of the proposed metric through an empirical study using the heterogeneous biological information network.

II. Related work

Biological network and its properties are researched in many fields. Some researchers have tried to combine different biological networks to make a dense, meaningful graph. Chen et al proposed a heterogeneous network combining a chemical information network, a PPI network, a drug-target network, and pathway information [4]. They tried to find drug targets by clustering the heterogeneous graph. Bio4j[1] is a project to combine different biological networks and to build a graph in a Neo4j graph database. Also, DrugBank [5] and PharmGKB [6] which are well known drug information databases are also heterogeneous biological networks in some ways.

A lot of research was done to find drug targets and drug pairs sharing targets. An increasing amount of researches are emphasizing the importance of *in silico* biology to find new drug targets and drug candidates using the biological network [7], [8].

One approach in finding the drug target was the use of side-effect similarity. Campillos et al found the side-effect information of drugs, and tried to use that information to find the drug target [9]. They assumed if drugs share more side-effects, the drugs tend to have the same drug targets. Their research motivated our work. In our study, we found that drugs sharing a target disease are more likely to have a DDI than drugs sharing a biomolecular target. It may sound somewhat counter-intuitive; however, it can also be interpreted as a pair of drugs designed for the same disease are likely to have been investigated more vigorously for DDIs than a pair of drugs designed for different diseases but sharing the same biomolecular target. We will discuss more on this in Section IV.

III. Method

A. Our Assumption for Biological Network

Our method is based on the following assumption: If two drugs share more properties (i.e., more neighbors in the network), they tend to have a DDI. This assumption can be explained in two aspects: drug similarity and target sharing. The more properties they share such as side-effects or drug-gene interactions, the more likely they are to share targets [9].

Although drug pairs do not share the properties directly, an indirect chain reaction can occur. For example, let us assume that drug A influences target T, and drug B interacts with protein P. In this situation, if protein P is known to inhibit target T, despite drug A and B not sharing the target directly, they have a very high chance of having a DDI. Not only bio-macromolecule targets, but also genes, proteins, and pathways related with the drugs directly/indirectly can be hidden/unintended targets of the drugs. Previous research which focuses on finding direct target-sharing of drugs fails to notice indirect, hidden, and unintended drug targets. Moreover, the number of known drug targets is very limited and many drugs are not fully researched about their influence on targets [10].

As such, we made an assumption that all the biological objects can be drug targets, or at least carrier of the influence to the targets. The more drugs share biological objects, the possibility of the DDI between drugs will be higher.

B. Construction of Heterogeneous Biological Information Network

We built a heterogeneous biological information network to find interactions between two different drugs. As mentioned above, a DDI can occur indirectly. Most of the biological reactions are related with complex interactions between biological objects. A DDI study requires understanding complex cross interactions of biomolecules and events that occur in the system, because drugs react to various target molecules, and there are very diverse types of drugs prescribed today. For these reasons, a heterogeneous biological information network that captures diverse biomolecules and their relations emerges as a valuable resource for DDI research.

To build a heterogeneous biological information network, we combined more than 8 databases from different

Table I
Biological Network Databases Included in Our Network

Database Name	Domain	Relation Type
PharmGKB	Drug Information	Drug-Disease Drug-Drug Drug-Gene Gene-Gene Disease-Disease
DrugBank	Drug Information	Drug-Gene Drug-Pathway Drug-Symptoms
Meddra	Drug Information	Drug-Symptom
HPRD	Protein-Protein Interaction	GeneOntology-Protein Protein-Protein
String	Protein-Protein Interaction	Protein-Protein
BioGRID	Genetic and Protein Interaction	Gene-Gene
Ensembl	Genome Browser	Gene-Protein
KEGG	Pathway Information	Pathway-Pathway Pathway-Gene Gene-Disease
TTD	Target Information	Target-Disease Target-Pathway Target-Drug
MeSH	Medical Terms	Synonym

[1] http://www.bio4j.com/

sources(Table I). The databases have their own systematic ways to name the objects. We used MeSH, a biological thesaurus that provides lists of synonyms for bio-medical terms [11]. We used the MeSH concept preferred terms to disambiguate the object names from different sources. For the object names that could not be resolved using MeSH, we matched the terms manually with help from domain experts.

Table II
PROPERTIES OF OUR HETEROGENEOUS BIOLOGICAL INFORMATION NETWORK

Categories	Num. of Nodes	Average Degree
Drug	20098	12.12
Genes+Proteins	47734	30.39
Diseases	2116	17.86
Pathways	2156	49.61
Side-effects	449	170.11
Targets	2057	35.61
gene-ontology	238	162.81
Total	74,848	27.09

C. Measuring the Relation Strength of Two Nodes

Since one of the main tasks of this study is predicting a potential relation between two nodes in the network, we need to develop a metric function that can measure relation strength between two arbitrary nodes in the network. The simple approach to this problem is to count all the possible paths between two nodes, and use that number as the relation strength between the two. The possible path between two drug nodes can be considered as a candidate drug-drug interaction path. Counting them all means counting all the possible candidates of DDI paths.

Another approach is to consider the different types of nodes in the paths. We divide them into 5 different categories: paths including a target, paths including a disease node, paths including a pathway, paths including a side-effect node, and paths purely consisted with genes and proteins. The number of paths in each category is used to predict the DDI. More specifically, we compute the weighted sum of the numbers of the paths in different categories and use it as a score for DDI prediction.

IV. EMPIRICAL VALIDATION OF THE PROPOSED NETWORK

A. DDI dataset

As mentioned earlier, the PharmGKB drug-drug interaction dataset is used to evaluate our methods. The number of DDI information we collected from this database totaled 264 interaction drug pairs. Among these drug pairs, 191 drugs appear on the list at least once. We generated all the possible pairs from the 191 different drugs and obtained $18,145 (= 191 \times 190/2)$ drug pairs. Finally, we divided the drug pairs into two groups, the DDI pair group (264 drug pairs) and the unknown group (17881 drug pairs). The amount of DDI information found in this database was very limited, and there are many DDI drug pairs which have not yet been discovered or have not entered into the database. Hence, we refer to the 17,881 drug pairs as the unknown DDI group.

The reason why we generated the unknown DDI group using only the 191 drugs (those that has at least one known DDI) is that we wanted to reduce the potential bias on the amount of available information between well-known drugs and new drugs as much as possible. We assumed that if a drug has been on the market for a long time, there is a higher chance of a DDI being discovered for the drug. Given that, the 191 drugs appearing in the DDI list can be assumed to have been under a fair amount of research; thus the chances of the two drugs in the list to have an undiscovered DDI is much lower than the chances of an undiscovered DDI existing between two drugs outside the list (for which no DDI has been discovered yet).

B. DDI Path Collection Using Graph Traversal Algorithms

We found paths between drugs by using graph traversal algorithms. Two different traversal algorithms are used for this test.

1) 2-hop Traversal: 2-hop traversal algorithm finds all the possible paths in a 2-hop distance. It also means that there is only one node in a path between two nodes. This neighbor-sharing between two drugs is the basis of former DDI research [9]. We performed the same experiments and set it as a base line for our research. 2-hop traversal algorithm is only considering the directly connected nodes from both drug nodes.

2) 2,3-hop Traversal: The last traversal algorithm that we used is a 2,3-hop traversal algorithm. In this method, we find all the possible paths between two nodes in a 3-hop distance. We did not explore 4-hop distance and up because paths with distance greater than 4 tend to start including a good deal of random connections. It is not surprising given the fact that biological network is a scale free network [12].

C. Comparison of Different Traversal Algorithms

Table III shows average number of paths returned by both of the traversal algorithms and their t-test results. There are on average 39.02 paths between two drugs known to have a DDI, returned by the 2,3-hop traversal algorithm while on average 24.74 paths are returned for drug pairs

Table III
COMPARISON OF THE NUMBER OF PATHS RETURNED BY DIFFERENT TRAVERSAL ALGORITHMS

	2-hop traverse			2,3-hop traverse		
	DDI	Unknown	p-value	DDI	Unknown	p-value
#all paths	27.75	19.36	7.82^{-12}	39.02	24.74	1.13^{-13}
#target paths	0.04	0.06	5.27^{-02}	0.66	0.22	1.63^{-11}
#disease paths	2.04	0.50	2.67^{-13}	2.71	0.72	1.49^{-16}
#gene/protein paths	6.132	2.98	3.28^{-15}	16.57	8.07	4.96^{-9}
#pathway paths	0.54	1.44	9.40^{-07}	0.54	1.44	9.40^{-07}
#side-effect paths	19.00	14.42	3.73^{-06}	19.00	14.42	3.73^{-06}

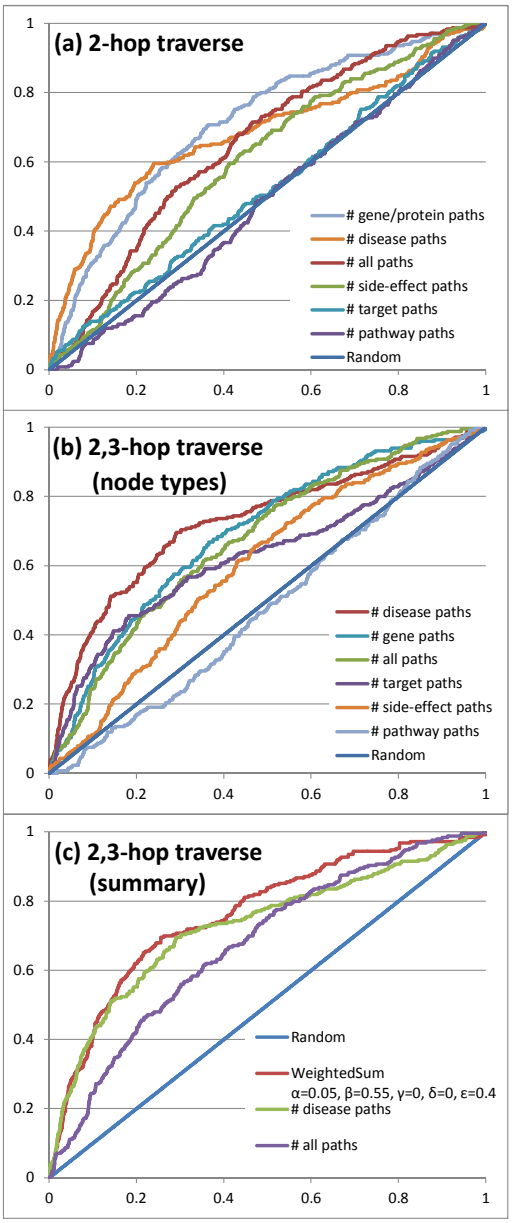

Figure 1. ROC Curves of DDI Pair Ranks

in the unknown group. Similarly, the second row shows the average number of paths containing a biomolecular target. The first column (0.66) represents the average number of "target paths" between drug pairs in the DDI group while the second column (0.22) represents the same between drug pairs in the unknown group. The third column shows *p*-value representing the statistical significance of the difference between the two groups. As we can see, 2,3-hop traverse shows better results than the 2-hop traversal algorithms.

D. Different Path Types and Their Relevance to DDI

In Figure 1-(a), and Figure 1-(b), we can compare the results to see which path types are most relevant to DDIs under different traversal schemes. In both tests, the DDI pairs that were ranked by the number of disease node-containing paths showed the best result.

Contrastingly, the results of the DDI pairs ranked by the number of target node-containing paths and the number of pathway node-containing paths were particularly bad. Many DDI research is based on the assumption that if drugs share targets, then they tend to have a DDI. However, in our experiments, we found a total of 281 target-sharing drug pairs, but only 10 pairs were known to have DDIs.

On the other hand, disease noticeably showed the highest relevance with a DDI in our network. It can be interpreted as a pair of drugs designed for the same disease are likely to have been investigated more vigorously for DDIs (thus more chance to discover a DDI) than a pair of drugs designed for different diseases but happened to share the same biomolecular target. Nonetheless, further investigation is needed to draw a concrete conclusion on this problem. We leave this problem for future work.

E. Finding Optimal Combination of Different Types of Paths

As we can see in Table III, different types of paths have different discrimination power. The difference in the number of disease paths appears to have the greatest statistical significance ($p < 1.49^{-16}$). It means that if we were to use only one type of path in our analysis, we should use the disease paths. However, if we can find a proper combination of the different types we may improve the prediction results. The following formula represents the weighted sum of the different types of paths between two drugs d_1 and d_2.

$$S(d_1, d_2) = \alpha N_t + \beta N_d + \gamma N_p + \delta N_s + \epsilon N_g \qquad (1)$$

where, α, β, γ, δ, and ϵ are user specified parameters such that $\alpha+\beta+\gamma+\delta+\epsilon = 1$; N_t,N_d,N_p,N_s, and N_g represent the number of paths containing a target node, a disease node, a pathway node, a side-effect node, and a gene/protein node, respectively.

We evaluated the above score for all pairs of drugs with varying parameter values to find the optimal combination. This test was performed using the 2,3-hop traversal method as it is the better performer between the two traversal algorithms. The AUC (Area Under the Curve) of the ROC curve was calculated for each trial and we found the best result occurs when α is 0.05, β is 0.55, γ is 0, δ is 0, and

Table IV
AREA UNDER THE CURVE(AUC) OF DIFFERENT TYPES OF PATHS

	2-hop traverse	2,3-hop traverse
All Paths	0.650759	0.672082
Target Paths	0.520563	0.627106
Disease Paths	0.679518	0.724022
Gene/Protein Paths	0.709098	0.674937
Pathway Paths	0.486221	0.480901
Side-effect Paths	0.604431	0.601766
Optimal Weight	-	0.759348

ϵ is 0.4(Figure 1-(c)). As expected, the disease path has the biggest impact on the DDI prediction. The AUC comparison result is given in Table IV.

As shown in Table III, Table IV and Figure 1, better results can be obtained when we consider the nodes that interacts directly and indirectly with neighbors rather than focusing on the nodes with direct interactions. The improved results of the 2,3-hop traverse group can be attributed to the longer path that can include more bio-macromolecular targets, target candidates (such as proteins, genes, and pathways), and related diseases, thus capturing more potential indirect relations.

Diseases are complex biological phenomenon which we are unable to completely understand the mechanisms occurring inside. A drug which acts on a disease will influence the middle section of the mechanism. If two different drugs act on the same disease, it may work on the same disease-causing mechanism. If two different drugs influence the same target, or the same disease-causing mechanism at the same time, the probability of a perturbation of biological systems occurrence is high.

F. Validation using Drugs.com drug interaction checker

We used the PharmGKB DDI dataset as a test set for our method; however, the number of DDI pairs was comparatively small to validate all the results of our experiments. Drugs.com provides much more information about DDIs, however, they do not provide raw data of their database or API to collect them. The only way to access the information is to directly query the system for each drug using the drug interaction checker. We validated 100 pairs from the best score and 100 pairs from the worst score produced by our optimal weight scoring method. In the drug interaction checker, the DDI is classified into 4 categories, such as major, moderate, minor, and not discovered. We excluded in the evaluation the drugs that does not exist in the Drugs.com drug interaction checker. Table V shows the result of the validation.

As we can see, most of the high score drug pairs have major DDI, but for the low score drug pairs, not discovered was dominant.

Table V
THE RESULT OF DRUGS.COM DRUG INTERACTION CHECKER

DDI Level	High 100	Low 100
Major	**38**	10
Moderate	36	9
Minor	2	4
Not discovered	22	**77**

V. CONCLUSION

We built a heterogeneous biological information network using proteins, genes, drugs, pathways, side-effects, diseases, and the interactions between these biological objects. We used this graph as an *in silico* test bench to find DDIs. We found that paths containing a disease node are more relevant to finding DDIs. We introduced a metric to measure the relation strength between two nodes, which is a linear combination of the numbers of paths with different types of nodes. Finally, we reported the optimal weight combination for predicting DDIs using our network.

REFERENCES

[1] B. Apsel and J. A. e. a. Blair, "Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases," *Nature Chemical Biology*, vol. 4, no. 11, pp. 691–9, 2008.

[2] F. D. A., *Approved Drug Products with Therapeutic Equivalence Evaluations*. US Food and Drug Administration, 2012, vol. 32nd Edition.

[3] S. I. Berger and R. Iyengar, "Network analyses in systems pharmacology," *Bioinformatics*, vol. 25, no. 19, pp. 2466–72, 2009.

[4] B. Chen, D. Wild, and R. Guha, "Pubchem as a source of polypharmacology," *J Chem Inf Model*, vol. 49, no. 9, pp. 2044–55, 2009.

[5] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "Drugbank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D1035–41, 2011.

[6] M. Hewett, D. Oliver, D. Rubin, K. Easton, J. Stuart, R. Altman, and T. Klein, "Pharmgkb: the pharmacogenetics knowledge base," *Nucleic acids research*, vol. 30, no. 1, pp. 163–165, 2002.

[7] S. Ekins, J. Mestres, and B. Testa, "In silico pharmacology for drug discovery: applications to targets and beyond," *Br J Pharmacol*, vol. 152, no. 1, pp. 21–37, 2007.

[8] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature Chemical Biology*, vol. 4, no. 11, pp. 682–690, 2008.

[9] 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, 324IS Times Cited:174 Cited References Count:27.

[10] N. M. Penrod, R. Cowper-Sal-lari, and J. H. Moore, "Systems genetics for drug target discovery," *Trends in Pharmacological Sciences*, vol. 32, no. 10, pp. 623–30, 2011.

[11] L. S. Knecht and S. J. Nelson, "Mapping in pubmed," *J Med Libr Assoc*, vol. 90, no. 4, p. 475, 2002.

[12] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality iprotein networks," *Nature*, vol. 411, no. 6833, pp. 41–2, 2001.