Semantic Knowledge Graph Network Features for Drug Repurposing

Tareq B. Malas[1], Roman Kudrin[1,2], Sergei Starikov[1,2], Peter A.C.´'t Hoen[1], Dorien J.M. Peters[1], Marco Roos[1], Kristina M. Hettne[1]

[1] Department of Human Genetics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands
[2] Faculty of Bioengineering and Bioinformatics, Moscow State University, 119234 Moscow, Russia

Abstract. Given the significant time and financial costs of developing a commercial drug, it remains important to constantly reform the drug discovery pipeline with novel technologies that can narrow the candidates down to the most promising lead compounds for clinical testing. Computational approaches are used to expedite the drug discovery processes. Semantic knowledge graphs can assist these computational approaches, because they connect different biological databases and reflect the relationships between genes, pathways and diseases. Here, we took advantage of the Euretos Knowledge Platform (EKP), a commercial database that integrates more than 170 different biological resources including DrugBank, and evaluated the usefulness of the underlying semantic knowledge graphs to predict novel drug-disease associations. We extracted network-based features from the semantic knowledge graph and tested their ability to separate between the positive and negative data sets of drug-disease pairs. Our results showed that the extracted features such as the total number of intermediate concepts (count), the number of different semantic categories (diversity), and the predicates connecting a drug-disease pair were successful in separating the positive from the negative sets. These features provide a proof of concept for using semantic knowledge graphs for drug repurposing efforts. Our work reveals the added value of integrating different biological databases for solving complex biological questions.

Keywords: Drug repurposing, drug discovery, Semantic graphs, network mining, machine learning.

1 Introduction

In silico methodologies are becoming more important in the modern-day drug discovery pipeline. Computational drug discovery techniques accelerated the identification of drug targets and significantly contributed to the different stages of drug development [1]. Most efforts are concentrated into developing methods for the prediction of drug-target interactions that mitigate the expensive costs of experimental drug development

and optimization [2]. Moreover, these methods are allowing for drug repurposing efforts that identify new therapeutic applications for existing drugs and reduce research cost and time due to the existing extensive clinical studies [2, 3].

Given that the majority of diseases cannot be explained by single-gene defects but by the coordinated functions of their complex gene networks, drug development needs to shift its attention towards understanding network-based perspectives of disease mechanisms. Network-based approaches are providing important insights into the relationship between drugs and diseases. An investigation into the interaction between drug targets and disease genes revealed that they are not closely related [4]. Additionally, network-based approaches are showing promise in predicting novel targets and new uses for existing drugs [5]. Current network-based approaches rely on drug target profile similarities. These similarities are defined by either the number of targets two drugs share or the shortest paths between their interactomes. However, these studies focus only on using a limited number of databases related to protein drug targets, leaving a large amount of rich data untapped.

Semantic and text-mining approaches that screen hundreds of thousands of published literature articles have demonstrated the possibility of extracting concepts of biological meaning of various types. Semantic knowledge graphs are constructed to connect concepts of various ty

pes based utilizing a number of resources such as literature knowledge and biological databases. Such knowledge graphs can then be used to infer novel connections based on network mining methods [6, 7]. In addition to semantic connections, large efforts were made to integrate biological databases across gene, protein, pathway, disease and drug domains. The Euretos Knowledge Platform (EKP, http://www.euretos.com/) is a commercial database that integrates more than 170 different biological resources including semantic data [http://www.euretos.com/files/EKPSources2017.pdf]. These data sources are used by EKP to build a large network of connected biological concepts. Disease and drug concepts in EKP are directly or indirectly connected based on prior knowledge found in publications and/or other databases. We expect that leveraging a large set of databases will enhance our drug discovery ability and avoid relying on a single source of information to associate drugs to diseases. Each semantic type provides us with an additional layer of information that can be exploited to identify novel drug disease associations.

In this work, we have taken advantage of the EKP to evaluate the usefulness of the underlying semantic knowledge graphs to predict novel drug-disease associations. With the current exponential growth in biological data, semantic knowledge graphs have a great potential for drug discovery.

2 Materials and Methods

2.1 Data Acquisition and Mapping in EKP

Drug disease pairs were acquired from Guney et al [8]. We specifically acquired the drug disease associations based on their analysis. We had 403 pairs of 239 drugs and 78 diseases that formed our positive "gold-standard" (GD) data. By randomly shuffling the 403 drug disease pairs of the positive dataset, we created 20 unique negative datasets that included 403 random drug disease pairs not seen in the positive dataset. We averaged the results of the negative datasets in the downstream analysis.

In EKP we first mapped the DrugBank IDs of the drugs in our datasets to drug concepts in EKP. We used full disease names to map the diseases in our dataset to disease concepts in EKP. Triples of drug disease pairs were identified in EKP if they were directly connected by at least one of the resources used in EKP (Figure-1). Predicates of drug disease triples were classified as "relevant" if they belonged to one of the following categories: "treats", "affects", "prevents", "disrupts". The LUMC has a local installation of this knowledge graph for research purposes.

2.2 Network Features

Network features were calculated for the intermediate concepts connecting drug disease pair. To evaluate if we could use the indirect associations to predict novel associations between drugs and diseases, we used the positive and negative datasets as follows. For each indirect association, we calculated a number of features and tested if these features could separate the two datasets. These features were calculated for each semantic subcategory (SubSemantic) available in EKP.

I. Count_normalized referred to as count in the following text:

$$(\text{“SubSemantic\_typeY”}) = X \div (y \times z) \tag{1}$$

X = total number of SubSemantic_typeY connecting the drug (y number of unique drugs making one drug concept) with disease (z number of unique diseases making one disease concept). The number of intermediate concepts between the drug and disease concepts was normalized by the multiplication of y and z.

II. Diversity = The total number of unique SubSemantic categories connecting the drug and disease concepts per semantic type.

III. Predicates from the drug concept to the intermediate concept and from the intermediate concept to the disease concept were combined and referred to as "predicate path". We used the Chi-square test to identify, within each semantic subcategory, the most enriched paths in the GD vs the negative dataset (cutoff p-value < 0.05). We filtered out paths that made up less than 1% of the total amount of paths within each semantic subcategory.

For I and II we used the Kolmogorov–Smirnov to test the similarity of the distribution of scores between the positive and the negative datasets (cutoff p-value < 0.05)

3 Results and Discussion

3.1 Concept Mapping and Direct Associations

We acquired the dataset of curated drug disease relationships (drugs used in the treatment of certain diseases) from Guney et al [8]. The GD dataset included 239 drugs, 78 diseases and 403 drug-disease pairs. For the negative dataset, we reshuffled the GD into 20 random datasets. The results of the negative datasets were averaged and compared to the GD.

We used DrugBank IDs available in the GD dataset to map drugs from the GD and negative datasets into EKP concepts and we used the full disease name to map diseases, since no unique identifier was supplied in the GD dataset. Out of 239 drugs, 235 were mapped successfully. All diseases were mapped successfully into EKP. Whendisease or drug term mapped to more than one concept in the EKP, this was corrected for (Figure-1).

Using the EKP we retrieved the triples for drug-disease pairs found in the GD and negative datasets. Each semantic triple consists of a subject-predicate-object, where the subject and the object refer to the drug and the disease respectively, and the predicate refers to the relationship connecting them. From the pairs found in the GD, 83% mapped to a triple in the EKP, whereas in the negative datasets 22% of the pairs mapped to a triple in the EKP. Moreover, from the mapped triples in the GD, 90% had a predicate type that we consider positive for a drug-disease association i.e. 'treats', compared to 75% in the negative datasets. These results demonstrate that the drug disease pairs in the GD and the negative datasets are different in two main aspects. 1). Most of the GD drug disease pairs could be represented in direct triples owing to prior knowledge of the pair's relationship. 2). The type of the predicates is different when comparing the triples of the GD and negative datasets, where the GD contains a higher proportion of the "relevant" predicates. The observed 22% of random drug disease pairs that mapped to triples in EKP could be explained by the smaller proportion of "relevant" predicates

compared to GD. These triples would contain negative drug disease indications or a drug that treats a side symptom of the disease.

3.2 Evaluating the Indirect Drug-Disease Associations

As we are interested in drug repurposing, we were looking for novel associations between drugs and diseases. We utilized the indirect drug disease associations as a basis for our method, where we aim to mine the full EKP graph of indirect drug disease associations for strong candidates using network based features. To identify which features are useful, we used the GD and the negative datasets and evaluated several network features on the indirect associations retrieved from them. In the EKP, 14 semantic types are defined based on the semantic groups as defined by the Unified Medical Language System [9], with a number of semantic subcategories under each semantic type. Our analysis of indirect associations, i.e. drugs and diseases that connected via a third concept, was done per subsemantic category.

All 403 drug-disease associations in the GD and negative dataset were connected by at least one intermediate concept from the semantic types available in EKP. Out of the 14 possible semantic categories, 12 were found to connect a drug and a disease. We next evaluated which semantic and semantic subcategories were the most informative. Using the count diversity feature, defined as the total number of a certain intermediate concept connecting a drug disease pair, the semantic type 'Chemicals & Drugs' was the most informative intermediate semantic type and distinguished the positive and negative sets best (Kolmogorov-Smirnov p-value: $7.4 \cdot 10^{-23}$). Density plots of the count values per semantic and semantic subcategory in both the GD and the negative data reveal visually that the GD contained a higher number of indirect concepts in most semantic categories compared to the negative dataset, such as "Chemicals & Drugs", "Anatomy", "Disorders" and "Procedures" semantic categories (Figure-2A, Table-1).

Another feature we investigated was the diversity of the different semantic types connecting a drug disease pair. In this analysis we compared the total number of unique semantic categories and semantic subcategories in the drug disease pairs of the GD and negative datasets. As observed for the count feature, the GD drug disease pairs displayed a higher semantic diversity in their intermediate concepts (Figure-2B).

We also investigated the predicate types that connect the indirect concept with the drug disease pairs. In this analysis we used two predicates, the one connecting the drug with the intermediate concept and the one connecting the intermediate concept with the disease concept. The combination of these two predicates in this order is referred to as the predicate path. Using the chi-squares test we investigated if there were predicate paths that are enriched in the GD and negative datasets. We found the most enriched paths in the "Amino Acid, Peptide or Protein" and "Pharmacologic Substance" seman-

tic subcategories (Figure-2C). For example, the path "drug →*is compared with* → Pharmacologic Substance → treats →Disease" that belongs to the "Pharmacologic Substance" semantic subcategory is strongly enriched in the GD that can be interpreted as drugs that are known to be similar in function or chemical properties can be repurposed for the same disease.

These results indicate that the type of, count and the predicates relating to the intermediate concepts connecting a drug and a disease pair were informative in differentiating positive and negative datasets. The added values of using a diverse set of semantic categories was demonstrated. In the count feature, we found almost all semantic categories shifted towards higher values in the GD when compared to the negative data. Additionally, the diversity feature revealed that the GD tends to have a higher number of semantic categories and subcategories as intermediate concepts connecting drugs and diseases. Having the 'Chemicals & Drugs' as the most differentiating semantic category also demonstrates the importance of looking at drug properties and not completely relying on the drug targets.

In contrast to other tools, our methodology is different in a number of ways. The quantity and diversity of databases that we included is larger and the content much richer than other comparable tools. In terms of quantity we have taken advantage of EKP that integrates more than 170 resources. Other network-based tools such as SLAP [6] and ProphNet [10] include 17 and 3 databases respectively. In terms of diversity, EKP includes databases that span drug, disease, phenotype, protein, gene and molecular pathways. Additionally, EKP takes advantage of mining the PubMed published literature. To our knowledge this is the most resource inclusive effort in network-based drug disease associations. Our methodology utilizes drug disease connections beyond the commonly used drug-targets-disease framework to expand the possibilities to include other semantic categories, such as drug-drug and disease-disease similarities, phenotypes, pathways, proteins and biological function annotations. Our method utilizes semantic knowledge graphs properties and can be extended to other semantic knowledge graphs that contain drug and disease concepts.

4 Conclusions

Computational efforts in drug discovery are gaining popularity for their ability to reduce the costs involved in drug development. Network-based approaches are currently being used for drug repurposing efforts. We have taken advantage of the EKP that integrates more than 170 biological sources. Leveraging 12 semantic categories that are found in the EKP to connect drug and disease pairs, we identified three main network features that showed significant differences in the characteristics of the intermediate concepts connecting the drug disease pairs in the Gold Standard and negative datasets. These features can be readily used to build a classifier that will mine the full EKP graph

to propose novel drug disease associations. Additional network features that are tailored to specific semantic types can be further extracted to fine tune the performance of the classifier.

This work demonstrates that semantic knowledge graphs have a strong potential in mitigating drug discovery efforts. We expect semantic graphs to grow with the exponential growth in data generation in life sciences. Thus, rendering semantic knowledge graphs even more valuable for drug discovery.

Table 1. Top 5 most significant semantic subcategories based on count feature.

Semantic type(subcategory)	Semantic Subcategory	Kolmogorov-Smirnov p-value
Organic Chemical	Chemicals & Drugs	7.39E-23
Pharmacologic Substance	Chemicals & Drugs	1.09E-22
Indicator, Reagent, or Diagnostic Ai	Chemicals & Drugs	7.12E-19
Hazardous or Poisonous Substance	Chemicals & Drugs	1.2E-16
Chemical Viewed Structurally	Chemicals & Drugs	2.72E-16

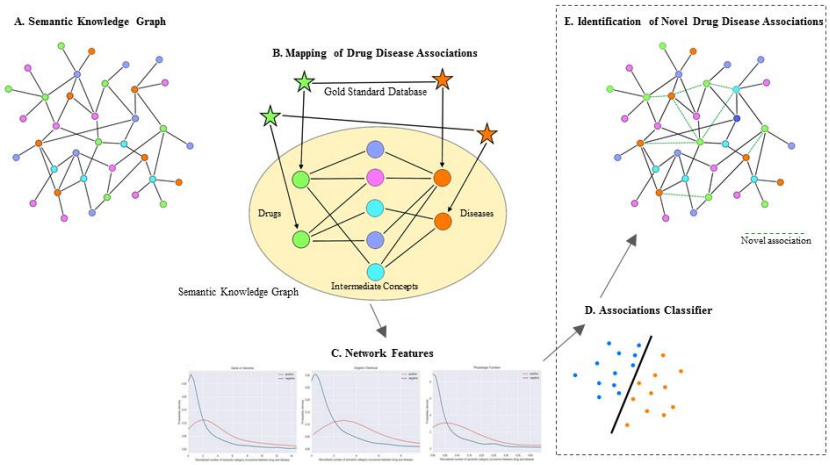

Fig. 1. We have used the Euretos Knowledge Platform (EKP) as the semantic knowledge graph in this analysis. Biological concepts (e.g. drugs, diseases, genes) are represented as circles, with different colors suggesting the variety of semantic types in the EKP (A). The drug disease pairs we acquired from an independent source were mapped to EKP concepts (B). Notably, mapped pairs were connected by intermediate concepts of 12 out of 14 different semantic types. We extracted network features from the intermediate concepts (C) to use them in building a classifier (future work) (D) to predict novel drug disease associations in the semantic network (E). Black dashed line reflects ongoing parts D and E of which their results are not included in this manuscript.

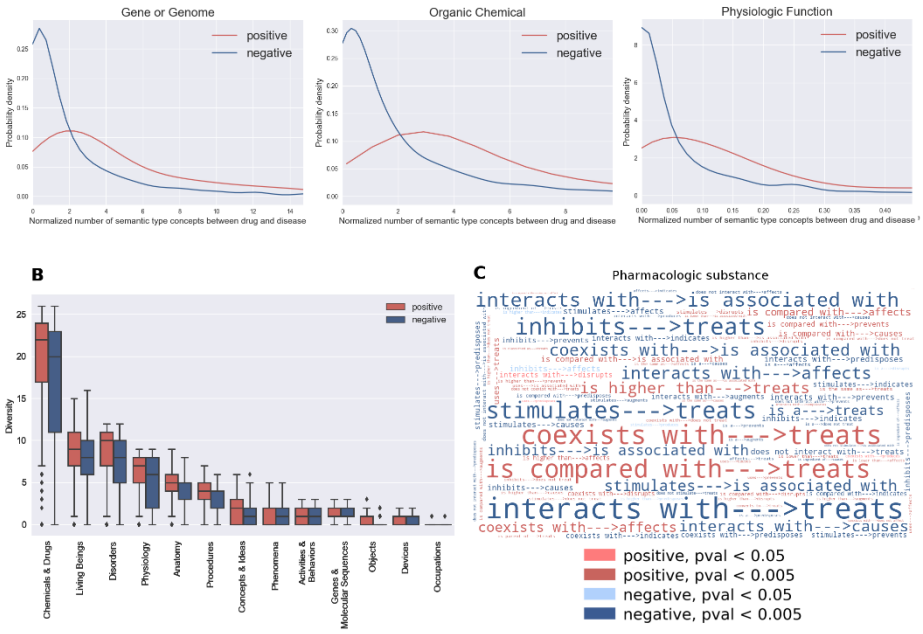

Fig. 2. A). Density plots of the count feature of three semantic subcategories. The higher the count value on the x-axis the higher the higher this semantic subcategory is found as an intermediate concept between drug and disease pairs.
B). Boxplots representing the diversity feature for concepts in each of the 12 semantic categories. For each semantic category, we have calculated the presence of each of the subcategories belonging to that semantic category.
C). Word Cloud representation of predicate paths of the "Pharmacological Substance" semantic category. P-values of the chi-square test residuals were used as an input to the cloud to calculate the enrichment of each path in either the positive and the negative datasets.

5 Acknowledgments

The research leading to these results has received funding from the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7/2077-2013 under REA grant agreement no. 317246. In addition, the European Commission (FP-7 project RD-Connect, grant agreement No. 305444).

6 Competing Interests

Kristina M. Hettne has performed paid consultancy since November 1, 2015, for Euretos b.v, a startup founded in 2012 that develops knowledge management and discovery

services for the life sciences, with the Euretos Knowledge Platform as a marketed product

7 References

1. Choi, S., Macalino, S.J.Y., Cui, M., Basith, S.: Expediting the Design, Discovery, and Development of Anticancer Drugs using Computational Approaches. Curr. Med. Chem. (2016).
2. Glick, M., Jacoby, E.: The role of computational methods in the identification of bioactive compounds. Curr. Opin. Chem. Biol. 15, 540–546 (2011).
3. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R.: Combining drug and gene similarity measures for drug-target elucidation. J. Comput. Biol. 18, 133–145 (2011).
4. Yildirim, M.A., Goh, K.-I., Cusick, M.E., Barabási, A.-L., Vidal, M.: Drug-target network. Nat. Biotechnol. 25, 1119–1126 (2007).
5. Wu, Z., Wang, Y., Chen, L.: Network-based drug repositioning. Mol. Biosyst. 9, 1268–1281 (2013).
6. Chen, B., Ding, Y., Wild, D.J.: Assessing drug target association using semantic linked data. PLoS Comput. Biol. 8, e1002574 (2012).
7. Hettne, K.M., Thompson, M., van Haagen, H.H.H.B.M., van der Horst, E., Kaliyaperumal, R., Mina, E., Tatum, Z., Laros, J.F.J., van Mulligen, E.M., Schuemie, M., Aten, E., Li, T.S., Bruskiewich, R., Good, B.M., Su, A.I., Kors, J.A., den Dunnen, J., van Ommen, G.-J.B., Roos, M., 't Hoen, P.A.C., Mons, B., Schultes, E.A.: The Implicitome: A Resource for Rationalizing Gene-Disease Associations. PLoS One. 11, e0149621 (2016).
8. Guney, E., Menche, J., Vidal, M., Barábasi, A.-L.: Network-based in silico drug efficacy screening. Nat. Commun. 7, 10331 (2016).
9. McCray, A.T., Burgun, A., Bodenreider, O.: Aggregating UMLS semantic types for reducing conceptual complexity. Stud. Health Technol. Inform. 84, 216–220 (2001).
10. Martínez, V., Cano, C., Blanco, A.: ProphNet: a generic prioritization method through propagation of information. BMC Bioinformatics. 15 Suppl 1, S5 (2014).