Journal of Medical Systems (2019) 43: 92
https://doi.org/10.1007/s10916-019-1182-z

SYSTEMS-LEVEL QUALITY IMPROVEMENT

KGDDS: A System for Drug-Drug Similarity Measure in Therapeutic Substitution based on Knowledge Graph Curation

Ying Shen[1] · Kaiqi Yuan[1] · Jingchao Dai[1] · Buzhou Tang[2] · Min Yang[3] · Kai Lei[1]

Received: 30 October 2018 / Accepted: 22 January 2019 / Published online: 5 March 2019

Abstract
Measuring drug-drug similarity is important but challenging. Significant progresses have been made in drugs whose labeled training data is sufficient and available. However, handling data skewness and incompleteness with domain-specific knowledge graph, is still a relatively new territory and an under-explored prospect. In this paper, we present a system KGDDS for node-link-based bio-medical Knowledge Graph curation and visualization, aiding Drug-Drug Similarity measure. Specifically, we reuse existing knowledge bases to alleviate the difficulties in building a high-quality knowledge graph, ranging in size up to 7 million edges. Then we design a prediction model to explore the pharmacology features and knowledge graph features. Finally, we propose a user interaction model to allow the user to better understand the drug properties from a drug similarity perspective and gain insights that are not easily observable in individual drugs. Visual result demonstration and experimental results indicate that KGDDS can bridge the user/caregiver gap by facilitating antibiotics prescription knowledge, and has remarkable applicability, outperforming existing state-of-the-art drug similarity measures.

Keywords Drug-drug similarity · Knowledge graph · Therapeutic substitution · Medical knowledge curation · Visualization

Introduction

Drug-drug similarity metrics (DDS) in medicine has attracted substantial attention in recent years due to its broad applications in medical information retrieval and knowledge reasoning [1]. The most promising application scenario is therapeutic substitution, also known as therapeutic interchange and drug substitution [2]. In literature, drug-drug similarity measures have been extensively studied in the last decade to enable a proper interpretation of drug information [3].

Despite the effectiveness of previous studies, DDS measure remains a challenge in real-world applications for two reasons: (i) The issues of complex and diverse terminology, relations, hierarchies and attributes in the medical field remain yet to be resolved. Knowledge graph (KG) is a collection of relational facts that are represented in the form of triplets [4]. The existing computation measures of KG-based semantic similarity can be classified into path/depth-based similarity measures and corpus-based methods. However, path-based and depth-based similarity measures cannot adequately handle the computation

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Kai Lei
leik@pkusz.edu.cn

Ying Shen
shenying.sz@pku.edu.cn

Kaiqi Yuan
kqyuan@pku.edu.cn

Jingchao Dai
1801213351@pku.edu.cn

Buzhou Tang
tangbuzhou@hit.edu.cn

Min Yang
min.yang@siat.ac.cn

[1] School of Electronics and Computer Engineering, Peking University (Shenzhen), 518055 Shenzhen, People's Republic of China

[2] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), 518055 Shenzhen, People's Republic of China

[3] SIAT, Chinese Academy of Sciences, 518055 Shenzhen, People's Republic of China

between two concepts with the same path but different semantic similarity in the KG taxonomy [5], while Corpus-based methods are substantially dependent on the training corpus and susceptible to data sparseness and data noise [6]. (ii) Pharmacology information, which plays a crucial role in semantic comprehension and similarity metric, is yet to be well-researched. Although many studies struggle to measure the semantic similarity aimed at bio-medical category, they leverage knowledge from KG to conduct knowledge representation learning rather than capture the various medical features such as common specificity and local granularity [7, 8].

To alleviate these limitations, this paper presents a system KGDDS for node-link-based bio-medical Knowledge Graph curation and visualization, aiding Drug-Drug Similarity measure in therapeutic substitution of antibiotics. The KGDDS architecture is designed to support four key principles: 1) **Good Graphic Display**. KGDDS can display knowledge based on the following criteria: uniform edge length, uniform vertex distribution and showing symmetry; 2) **Easy Extensibility**. KGDDS is easily adapted and extended to fulfill additional aesthetic criteria; 3) **Good Interaction**. The flexible RESTful API can be accessed to obtain local data as requested by users easily. KGDDS presents information in both an intuitive vision and textual form formats; 4) **Domain specific**. KGDDS takes the viewpoint of a user, and bridge the user/caregiver gap by exhibiting clinical knowledge.

We deploy the system on a server, which is equipped with 32GB of RAM and 8 Intel Xeon E5–2630 CPUs of 2.4GHz; 4 cores each. The machines run a 64-bit CentOS Linux. The KGDDS is developed by Java 1.8 in IntelliJ IDEA 2017. Spring Boot greatly accelerated the initial construction and development of Spring applications. Tomcat, which is embedded in Spring Boot, enables a one-click deployment of WAR file, simplifying the application deployment to the server configuration process. After packing codes into a war package, one can successfully deploy the system by entering the command line "java -jar packageName.war". Then one can access the system in a web browser. MyBatis automatically completes the interaction between the system and the database. The queried data is cached, greatly accelerating the follow-up of the same data query. Once the templates of the website have been decided, FreeMarker displays backend data from the backend and relieves the developer from web designing. D3.js converts the json data sent from the backend into SVG, and displays it on the system page. D3.js speeds up the generation and loading of front-end, large-scale data visualizations. KGDDS takes less than 100 ms to display the knowledge graph after receiving the user's query.

The main contributions of our approach are three-folded:

(1) We propose a novel neural network model, which leverages pharmacological and pharmaceutical knowledge from external knowledge bases, so as to broader medical knowledge for similarity measures.
(2) We construct a large-scale antibiotic relevant knowledge graph to aid the drug-drug similarity computation, addressing the data skewness and knowledge incompleteness. The KG visualization can be accessed via www.iasokg.com.
(3) Based on experimental results on Drugbank, the method proposed herein achieves better performance than existing methods in drug-drug similarity measures.

Related work

Knowledge graph construction

In the field of medicine, clinicians cannot obtain the necessary knowledge regarding clinical processes from inadequate medical databases [8, 9]. Some platforms in the general field, e.g. CKAN[1] and CN-Dbpedia,[2] cannot provide high-quality domain knowledge to meet the practical needs of the medical industry [10]. To address this problem, many studies have investigated constructing medical knowledge graphs [11], including Drugbank,[3] Disease Ontology (DO),[4] Infectious Disease Ontology (IDO)[5] and Dengue ontology (IDODEN).[6]

However, most medical knowledge graphs are published without sophisticated visualization tools that are equipped with advanced knowledge harvesting and analytics support [12]. Most current visualization techniques are applicable to graphs with specific structural properties [13, 14]. The visual representation of large datasets enables faster analysis by end users since human beings are better at "recognition" tasks than at "memorization" tasks [15]. As a result, there is a growing need for effective medical knowledge graph visualization for design and browsing [16].

Drug-drug similarity

Recently, there is a growing interest in computationally predicting potential drug-drug interactions (DDI). These approaches are broadly classified as either similarity [17] or feature-based [18] DDI predication methods. The core idea of the similarity-based approach [19] is to predict the presence of interactions between a pairwise of drugs by comparing it to known interacting drug pairs. However, existing similarity-based approaches are difficult to distinguish between a low

[1] https://ckan.org/
[2] http://kw.fudan.edu.cn/apis/cndbpedia/
[3] https://www.drugbank.ca/
[4] disease-ontology.org/
[5] infectiousdiseaseontology.org/
[6] https://bioportal.bioontology.org/ontologies/IDODEN

similarity value between two drugs without proper features [20–22]. In addition, these existing techniques for drug similarity measures rely on a limited number of data sources (e.g., DrugBank) that can provide only partial information about a subset of drugs of interest, leading to varying levels of incompleteness [23].

Different from the aforementioned studies, we develop new DDS measures within neural network architecture by simultaneously considering the drug pharmacology and the domain-specific knowledge included in the KG. We also define drug features by learning low-dimensional embeddings of drugs from textual and graph-based datasets.

Medical knowledge graph

Knowledge graph construction

In this study, we reuse existing knowledge bases to alleviate the difficulties in building a high-quality domain knowledge graph. The KG hierarchical conceptual schema is arranged based on knowledge acquisition associated with infectious disease diagnosis and antibiotics prescriptions. The schema covers the following nine dimensions: Disease, Infection site, Bacteria, Animal, Symptom, Symptom Type, Situation, Complication, and Antibiotic. A MySQL database was created based on this schema.

Based on the existing knowledge bases, the KG completion is carried out by filling the existing knowledge bases into schema through knowledge acquisition and matching. The existing knowledge bases include the aforementioned DO and IDO ontology, as well as NCBI organismal classification ontology, Human Phenotype Ontology (HPO), and DrugBank.

These existing knowledge bases were matched in the MySQL database based on the components “class name” and “alias” (defined by the relation “hasExactSynonym” in OWL) through entity linking. Classes not found in the database are considered and labelled as new classes. For classes that have already existed in the database, we merge the information between the new input class and the existing class.

For the KG construction, the Owlready[7] package was used to convert the MySQL database to OWL (Web Ontology Language) format. To carry out the medical KG visualization, KG components are retained in MySQL database for easy querying.

Subgraph probability

Generally, KG is build-up with entities and the relations between entities and without probability distribution, the latter of which is crucial in various domain applications such as decision making support and probabilistic inference. For a large KG, determining the probability between relations requires tremendous amount of time. Moreover, the uneven entity distribution in the corpus may lead to deviation of probability, resulting in low precision of probability acquisition of the whole KG. In this study, we provide insights into probability distributions for the subgraph consisting in individual nodes and edges.

Subgraph model Let $G^{w}(V, E, W)$ be a probabilistic subgraph with $V = (p_0, p_1, p_2, \cdots, p_n)$ being a set of vertexes made up of a center node p_0, where p_i denotes the edge node of the subgraph. E is a set of edges. $W = \left(w_{p_i}\right)_{\{p_i\}\in V_e}$ is the co-occurrence frequency of words in websites and can be considered as a set of probability values of nodes in KG:

$$\mathrm{Freq}(p_i) = log(|V| + 1)/(DF(p_0, p_i) + 1) \tag{1}$$

where $|V|$ is the set of documents containing the center node p_0, $DF(p_0, p_i)$ is the number of documents containing center node p_0 and edge node p_i. We normalize the weight of each edge as:

$$\mathrm{w}_{p_i} = \frac{\mathrm{Freq}(p_i) - min\left\{\mathrm{Freq}\left(p_j\right)\middle|p_j \in V_e\right\}}{max\left\{\mathrm{Freq}\left(p_j\right)\middle|p_j \in V_e\right\} - min\left\{\mathrm{Freq}\left(p_j\right)\middle|p_j \in V_e\right\}} \tag{2}$$

In this context, we get the probabilistic edge as shown in the left panel of Fig. 1. Considering that all weights are mutually independent, the probabilistic path (see the 2nd panel in Fig. 1) from one node to another node through the centroid node is the product of the weights of the two paths.

All paths going through the same centroid are clustered to form an initial subgraph. The probability distributions of these paths are evaluated (see the 3rd panel in Fig. 1). Since not all paths are reasonable (e.g. some probabilities are too low to be considered), we follow the constraints of weights in W to filter useless paths:

$$\sum_{\{p_i\}\in V_e} w_{p_i} = 1 \tag{3}$$

Force-directed layout Using a force-directed layout, we compute the node positions. For each node, its positions as an approximation to its distribution in 2D space (see the 4th panel in Fig. 1). Each displacement of p_i can be calculated as follows:

$$displacement(p_i) = \frac{w(p_0)(d - \|p_i - p_0\|)v}{(w(p_i) + w(p_0))\cdot\|p_i - p_0\|} \tag{4}$$

where d is the maximum distance between the center node p_0 and the edge node p_i, $\|\cdot\|$ denotes the Euclidean norm, v is the unit-length direction vector of the line passing through p_0 and

[7] https://pypi.python.org/pypi/Owlready

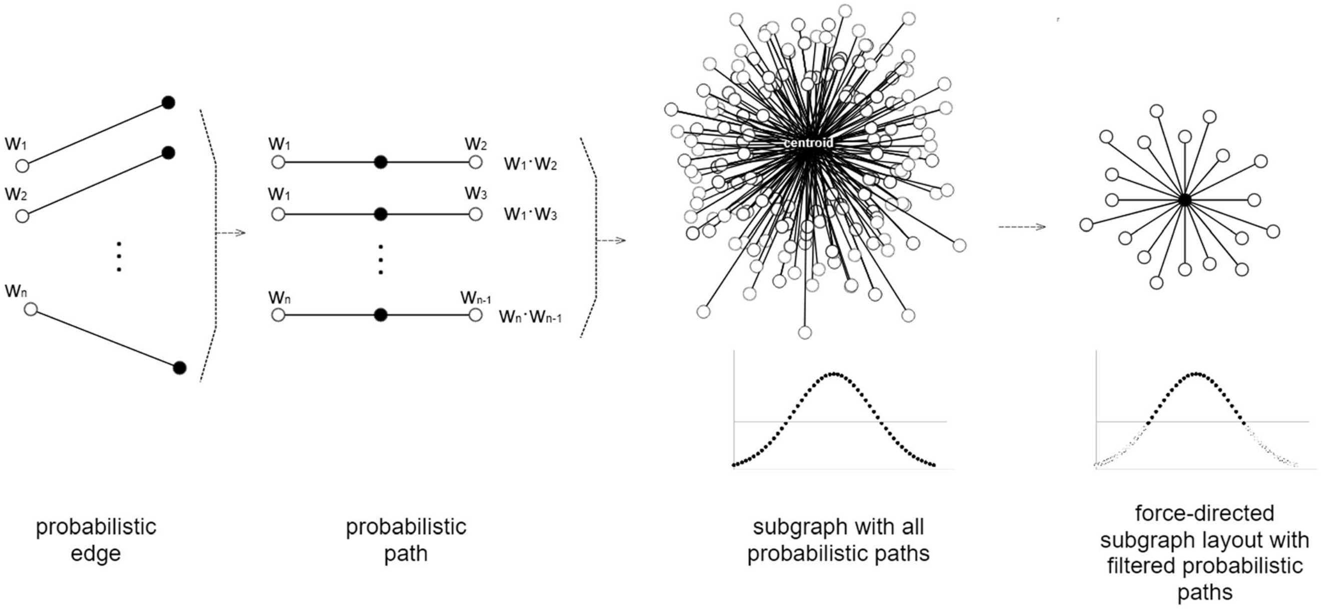

W1
W2
Wn
W1
W2
W1·W2
W1
W3
W1·W3
Wn
Wn-1
Wn·Wn-1
centroid
probabilistic edge
probabilistic path
subgraph with all probabilistic paths
force-directed subgraph layout with filtered probabilistic paths

Fig. 1 Overview of the probabilistic graph layout process

p_i. Eq. (4) can converge reliably after cycling constraints of the displacement of nodes relatively few times, and can well recognize the potential flexibility and scalability of a general framework for complex 2D graph layout built on Euclidean distance constraints.

Knowledge graph edge splatting

We believe that the KG structure and association of node and edges should be visible at the same time. We depict edges between node instances by adopting hierarchical edge bundling [24], even though this conceals the distribution of edges.

In terms of hierarchical edge bundling, we define a set of edges between two sets of sample nodes as one level of hierarchy. Formally, a bundled edge is defined as rational quartic Bézier curve $C(t)$ with nodes $n_i \in \mathbb{R}^2$ and corresponding weights $w_i > 0$:

$$C(t,n,w) = \sum_{i=0}^{4} \binom{4}{i} t^i (1-t)^{4-i} w_i n_i \tag{5}$$

where n_1 and n_3 are the centroids of the clusters. n_0 and n_4 represent the source and target positions of sample nodes, respectively. The remaining point n_3 is the midpoint of the segment (n_1, n_3).

Drug-drug similarity measure

KGDDS (Fig. 2) is a similarity metric of medicine predicted by the **Drug Similarity Computation Module** with pharmacology features and KG features. The **User interaction model** allows users to input the name of an antibiotic and obtain the drug similarity results.

Input representation

Pharmacology embeddings

We explore the pharmacology features that can simplify the semantic representation of medicines.

Side effect Given a drug d, its side effect information can be obtained from SIDER database. We use the Inverse Document Frequency (IDF) to reduce the impact of high frequency words and pay more attention to low frequency words:

$$\mathrm{IDF}(s, Drugs) = \log\Big((|Drugs|+1)/\Big(DF(s, Drugs)+1 \tag{6}$$

where s indicates a side effect, $Drugs$ stands for the set of all drugs, and $DF(s, Drugs)$ is the number of drugs with the side effect s. The vector of the side effect of a drug is $sider(d)$, and its value is $IDF(s, Drugs)$. The side effect-based similarity of two drugs d_1, d_2 is the cosine distance between the vectors $sider(d_1)$ and $sider(d_2)$.

Drug mechanism and physiological effect Given a drug, its drug mechanisms and physiological effect are both collected from NDF-RT. Based on these two features separately, the similarity of two drugs d_1 and d_2 is calculated same as the drug side effect.

We use a feedforward layer to reduce the dimensionality of the pharmacology features from over 7000 dimensions to 500

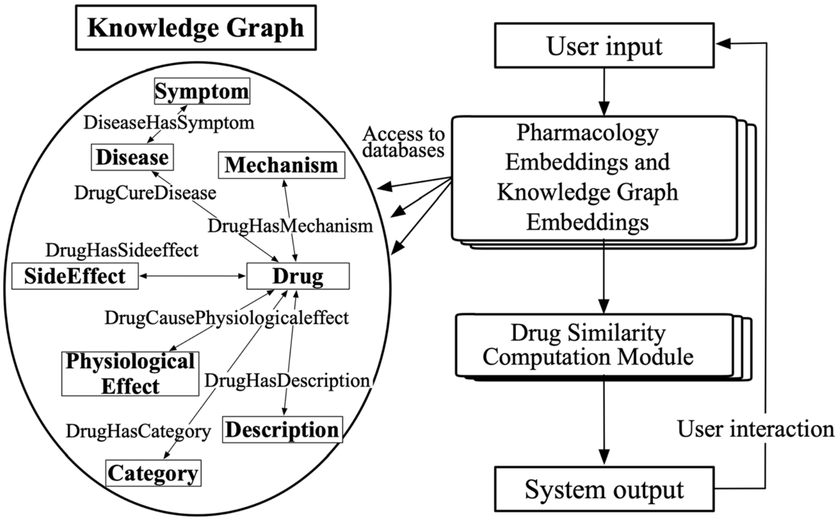


Fig. 2 Drug-Drug similarity measure framework

dimensions, so as to improve the the interpretability and calculation of feature embeddings.

Knowledge graph embeddings

Structural feature LINE [25] is adopted to learn the hierarchy embeddings from the taxonomy of DrugBank. Different from the structure-based methods that mainly concern the path connection and node degrees of KG, structural embeddings learn the KG properties by projecting the whole knowledge graph into a low dimension vector space. The similarity between the two drug vectors can be calculated by the cosine distance between them.

Textual feature The entity description and other textual information of concepts in KG usually carry conceptual semantic information. Based on the entity description, the BM25 algorithm [26] is adopted to calculate the textual similarity. Given a drug d_1 and its description D that containing keywords q_1, q_2, … , q_n, we can compute the BM25 score of drug d_2:

$$\mathrm{score}(d_1, d_2\ ) = \frac{\sum_{i=1}^{n} IDF(q_i)\cdot f(q_i, D)\cdot(k_1+1)}{(f(q_i, D) + k_1\cdot(1-b+b\cdot|D|/avgl))} \tag{7}$$

where $IDF(q_i)$ stands for the IDF weight of the keywords q_i. $f(q_i, D)$ indicates the occurrence frequency of keywords q_i in the description D. |D| is the number of words in the description D. $avgl$ denotes the average number of words of all entity descriptions extracted from Drugbank.

Given the set of all drugs $Drugs$, the KG-based textual similarity between drug d_1 and drug d_2 can by given by:

$$\mathrm{KTS}(d_1, d_2) = \left(\left(socre(d_1, d_2) - min\left\{score(x, y)\,\middle|\, x, y{\in}Drugs\right\}\right)\right)/ \\ \left(\left(max\left\{score(x, y)\,\middle|\, x, y{\in}Drugs\right\} - min\left\{score(x, y)\,\middle|\, x, y{\in}Drugs\right\}\right)\right) \tag{8}$$

Attention mechanisms and softmax layer

Given a drug and all aforementioned features, we apply attention mechanism to assign different weights according to the specific role each feature plays when interacting with other features. The representation of side effect feature v_s are calculated as:

$$M_w = \tanh(W_{sw}H_w) \tag{9}$$

$$\alpha_w = \mathrm{softmax}\left(w_w^T M_w\right) \tag{10}$$

$$v_s = H_w \alpha_w^T \tag{11}$$

where $M_w \in R^{dlxm}$ is a nonlinear mapping function, $W_{sw} \in R^{dlxdl}$ and $w_w \in R^{dl}$ are projection parameters, $\alpha_w \in R^m$ is the normalized attention, H_w refers to the side effect vector sequence of a drug. Other four types of features are processed by the same attention mechanism.

Afterwards, there is a joint layer to join the final drug representations of drug 1 and drug 2. The outputs of the softmax layer is to predict the similarity of a drug pair:

$$\mathrm{y} = \mathrm{softmax}(\mathrm{W_o pr} + \mathrm{b_o}), \tag{12}$$

where each dimension of y denotes the normalized the similarity of a drug pair. $W_o \in R^{2xdl}$ is the projection matrix, and $b_o \in R^2$ is the offset vector.

Table 1 Results of KG Enrichment

	Class	Axiom	Logical axiom	Annotation axiom	SubClassOf	Object property	Hidden GCI
IDO	507	4668	1064	2.973	582	39	81
HPO	936	5937	840	4.143	840	1	0
DO (infectious diseases)	11,280	118,726	7509	99.876	7.507	0	0
DrugBank (antibiotics)	341	3871	275	2.856	287	0	0
Website and guidelines	415,533	2,498,168	1,696,510	878.270	423.244	0	0
KGDDS	1,267,004	7,608,725	2,540,731	2.533.987	1.266.993	39	0

Experiment and results

Dataset and metrics

Data collection KGDDS conducts the drug similarity evaluation mainly based on the antibiotic-relevant information in DrugBank.[8] We study the relationships between antibiotics and their corresponding side effects from SIDER,[9] explore the mechanism of essential pharmacologic properties of medications from NDF-RT[10] and extract textual feature from more than 500,000 papers about medicine provided by PubMed.[11]

Antibiotic pairs labeling To verify the effectiveness of KGDDS, we conduct experiments on 1326 pairs most commonly used antibiotics. Referring to [27], doctors score the similarity between two antibiotics, which ranges in [0, 1], according to both antibacteria spectrum and efficacies of medicine (see www.iasokg.com). 0 indicates that there is no similarity between two antibiotics, while 1 implies that the two antibiotics are extremely similar. To make antibiotic pairs labeling more accurate, each pair is labeled by at least 3 doctors and the average is taken as the final result. The Pearson coefficient between the scores issued by each doctor and the average score ranges from 0.827 to 0.864 while Spearman coefficient ranges from 0.792 to 0.888, both proving the reliability of doctors' assessment. The labeled antibiotic pairs are divided into training set and test set.

Metrics For the CNN model, the kernel and the depth are set to 5 and 20 respectively. A Fully connected layer whose size is 500 is added after the CNN layer. In the BM25 implementation, k_1 and b are set to 2 and 0.75, respectively. The maximum length of entity description is set to 400 words. For the DDS on Drugbank, Pearson correlation coefficient and Spearman rank correlation coefficient are adopted to evaluate the correlation between doctors' assessment and experiment results.

Knowledge comprehensiveness of knowledge graph construction

The antibiotics KG was constructed based on the DO, IDO, NCBI, HPO and DrugBank databases. Table 1 indicates different KG metrics as reported in Protégé. We deleted irrelevant object properties and hidden GCIs because they are sporadic and thereby better suited to the manual investigation. All other components of Class, Axiom and SubClassOf were retained.

Experimental results of KG-based drug-drug similarity measures

The experimental results on Drugbank are summarized in Table 2. Several state-of-the-art baselines are adopted for comparison: (1) GADES: a graph-based semantic similarity measure approach [8]; (2) Res: a model using information content to evaluate semantic similarity in a taxonomy [6]; (3) Wpath: a model which computes semantic similarity of concepts in knowledge graphs by considering both path information and information content [28]; (4) Hybrids: an information retrieval method based on Wpath and takes medical properties into account to calculate the drug similarity [29]. (5) MedSim: a semantic similarity method in bio-medical knowledge graphs using random forest regression model [30]. (6) Tiresias, a large-scale similarity-based framework that takes in various

Table 2 Result on Drugbank (with ablation study)

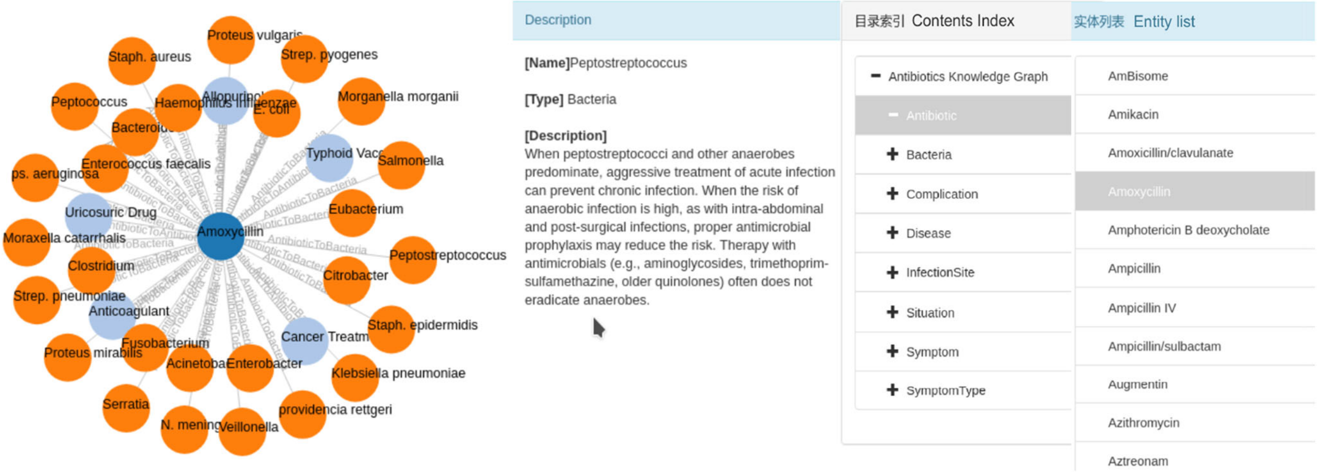
Model	Pearson	Spearman
GADES: Traverso et al. 2016 [8]	0.251	0.202
Res: Resnik et al. 2005 [6]	0.211	0.223
Wpath: Zhu et al. 2017 [28]	0.250	0.203
Hybrids: Hliaoutakis 2005 [29]	0.257	0.278
MedSim: Lei et al. 2018 [30]	0.585	0.523
Tiresias: Abdelaziz et al. 2017 [31]	0.591	0.532
KGDDS	0.623	0.589
w/o pharmacology feature	0.336	0.348
w/o KG structural feature	0.578	0.511
w/o KG textual feature	0.551	0.489

[8] https://www.drugbank.ca/

[9] sideeffects.embl.de/

[10] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/

[11] https://www.ncbi.nlm.nih.gov/pubmed/



Description

[Name]Peptostreptococcus

[Type] Bacteria

[Description]
When peptostreptococci and other anaerobes predominate, aggressive treatment of acute infection can prevent chronic infection. When the risk of anaerobic infection is high, as with intra-abdominal and post-surgical infections, proper antimicrobial prophylaxis may reduce the risk. Therapy with antimicrobials (e.g., aminoglycosides, trimethoprim-sulfamethazine, older quinolones) often does not eradicate anaerobes.

Fig. 3 Visualization of “amoxycillin” subgraph. From left to right: tree overview, detail description, and taxonomy demonstration

sources of drug-related data and knowledge as inputs, and provides DDI predictions as outputs [31].

In order to analyze the effectiveness of the different features of KGDDS, we also report the ablation test in terms of discarding pharmacology feature (w/o pharmacology feature) and KG feature (w/o KG structural feature and w/o KG textual feature).

There are multiple interesting observations from Table 2 are shown as follows: (1) KGDDS substantially and consistently outperforms the existing methods by a noticeable margin with respect to different correlations. (2) Generally, all features contribute in similarity measure, and it makes larger performance boosting to measure medical semantic similarity. (3) Although the MedSim and Tiresias employ more features, the KG we used is more domain-specific and completed, thus providing more knowledge for the similarity computation. This is within our expectation since medical knowledge can enhance the knowledge representational learning of a specific domain, while KG can further introduce structural and textual knowledge to enrich overall knowledge representations.

Data visualization and availability

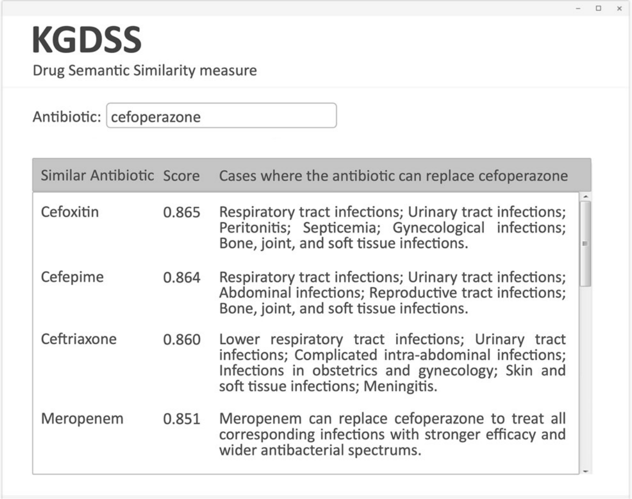
KGDDS is a visual environment for browsing KG represented as directed graphs. Graphs are visualized using circle and arcs between them. Nodes are class and instance nodes, with relations represented as the edges linking these nodes.

Overview, zoom and details on demand Most current knowledge bases, e.g. DBPedia, Yago and Drugbank, are release without visualization tool and present information in textual formats. KGDDS enables different forms of visualization on the proposed KG, including intuitive vision and textual formats.

Users can either consult the entities in a graphical display or browse the waterfall chart which visualizes the hierarchy of entities with respect to their biomedical classification. The left panel “Graph” in Fig. 3 shows results of entity relation query and subgraph query, while the middle panel “Description” presents results of entity facts query. KGDDS also allows the user to explore the KG hierarchy (see the right panel in Fig. 3).

Different from the most common used OntoViz visualization tool [32], KGDDS allows keyword search and enables users to explore the hierarchy or graph starting from a specific node. The KG is presented as a 2D graph with the capability for each class to present, apart from the name, type, description, taxonomy and relations. The entities are displayed in different colors.

Case study To study drug substitution, we employ KGDDS to predict the similarity scores between cefoperazone and other antibiotics. Referring to [33], two antibiotics whose similarity scores over 0.85 can be replaced with each other under normal circumstances.



KGDDS
Drug Semantic Similarity measure

Antibiotic: cefoperazone

Similar Antibiotic	Score	Cases where the antibiotic can replace cefoperazone
Cefoxitin	0.865	Respiratory tract infections; Urinary tract infections; Peritonitis; Septicemia; Gynecological infections; Bone, joint, and soft tissue infections.
Cefepime	0.864	Respiratory tract infections; Urinary tract infections; Abdominal infections; Reproductive tract infections; Bone, joint, and soft tissue infections.
Ceftriaxone	0.860	Lower respiratory tract infections; Urinary tract infections; Complicated intra-abdominal infections; Infections in obstetrics and gynecology; Skin and soft tissue infections; Meningitis.
Meropenem	0.851	Meropenem can replace cefoperazone to treat all corresponding infections with stronger efficacy and wider antibacterial spectrums.

Fig. 4 Drug similarity result provided by KGDDS

For the antibiotic cefoperazone, Fig. 4 presents antibiotics that are similar to it whose similarity score is over 0.85 and indicates the cases where they can replace each other. Take cefoperazone and ceftriaxone as an example. Ceftriaxone can replace cefoperazone in most cases except disease caused by a few bacteria such as *Pseudomonas aeruginosa* etc. In the absence of susceptibility testing, our method can help doctors to find the most appropriate drug substitution to treat most of Gram-negative bacteria infections, such as respiratory infection, pneumonia, and biliary infection.

Conclusions

KGDDS displayed excellent performance in calculating semantic similarity in antibiotics. The proposed method is extensible, reproducible and applicable to KG-based similarity calculation in medical field. In addition, this study developed a KGDDS with combined research on knowledge acquisition, KG construction and visualization in the infectious disease and antibiotic prescription fields. We believe that the most promising avenues for future research include incorporating other semantic similarity evaluation methods based on taxonomical relation, property information, relationships and other factors.

Acknowledgements This work was financially supported by the National Natural Science Foundation of China (No.61602013), and the Shenzhen Fundamental Research Project (No. JCYJ20170818091546869, JCYJ20160330095313861). Min Yang was sponsored by CCF-Tencent Open Research Fund.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Hliaoutakis, A., Varelas, G., Petrakis, E. G. M., Milios, E., MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity. In Proceeding of the International Conference on Theory and Practice of Digital Libraries. Springer, Berlin, 512–515, 2006.
2. Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G., Measures of semantic similarity and relatedness in the biomedical domain. J. Biomed. Inform. 40(3):288–299, 2007.
3. Nguyen, H. A., Al-Mubaid, H., New ontology-based semantic similarity measure for the biomedical domain. In: Proceeding of the 2006 IEEE International Conference on Granular Computing. IEEE, 623–628, 2006.
4. Batet, M., Sánchez, D., and Valls, A., An ontology-based measure to compute semantic similarity in biomedicine. J. Biomed. Inform. 44(1):118–125, 2011.
5. Li, Y., Bandar, Z., and McLean, D., An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. 15(4):871–882, 2003.
6. Resnik, P., Using information content to evaluate semantic similarity in a taxonomy. In Proceeding of IJCAI. 448–453, 2005.
7. Al-Mubaid, H., Nguyen, H. A., A cluster-based approach for semantic similarity in the biomedical domain. In Proceeding of the 28th Annual International Conference of the IEEE EMBS'06. IEEE, 2713–2717, 2006.
8. Traverso, I., Vidal, M. E., Kämpgen, B., Sure-Vetter, Y., GADES: A Graph-based Semantic Similarity Measure. In Proceeding of the 12th International Conference on Semantic Systems. ACM, 101–104, 2016.
9. Shiralkar, P., Flammini, A., Menczer, F., Ciampaglia, G. L., Finding streams in knowledge graphs to support fact checking. In Proceeding of the 2017 IEEE International Conference on Data Mining (ICDM 2017). IEEE, 859–864, 2017.
10. Jovic, A., Prcela, M., Gamberger, D., Ontologies in medical knowledge representation. In Proceeding of the 29th International Conference on Information Technology Interfaces (ITI 2007). IEEE, 535–540, 2007.
11. Ge, T., Wang, Y., De Melo, G., Li, H., Chen, B., Visualizing and curating knowledge graphs over time and space. In Proceeding of ACL-2016 System Demonstrations (ACL-2016). 25–30, 2016.
12. Monika, L., Sampson, J., and Rester, M., Ontology Visualization: Tools and Techniques for Visual Representation of Semi-Structured Meta-Data. J. Univ. Comput. Sci. 16(7):1036–1054, 2010.
13. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., and Giannopoulou, E., Ontology visualization methods-a survey. ACM Computing Surveys (CSUR) 39(4):10, 2007.
14. Abello, J., Van Ham, F., and Krishnan, N., Ask-graphview: A large scale graph visualization system. IEEE Trans. Vis. Comput. Graph. 12(5):669–676, 2006.
15. Auber, D., Tulip-a huge graph visualization framework. Graph Drawing Software:105–126, 2004.
16. Fu, B., Noy, N. F., Storey, M.A., Indented tree or graph? A usability study of ontology visualization techniques in the context of class mapping evaluation. In Proceeding of the International Semantic Web Conference: 117–134, 2013.
17. Zhang, P., Wang, F., Hu, J., Sorrentino, R., Towards personalized medicine: Leveraging patient similarity and drug similarity analytics. In Proceeding of the AMIA Summits on Translational Science Proceedings: 132, 2014.
18. Vilar, S., Uriarte, E., Santana, L., Tatonetti, N. P., and Friedman, C., Detection of drug-drug interactions by modeling interaction profile fingerprints. PLoS One 8(3):1–11, 2013.
19. Zhang, P., Agarwal, P., and Obradovic, Z., Computational drug repositioning by ranking and integrating multiple data sources. Machine Learning and Knowledge Discovery in Databases:579–594, 2013.
20. Baig, M. M., and Gholamhosseini, H., Smart health monitoring systems: an overview of design and modeling. J. Med. Syst. 37(2):9898, 2013.
21. Luo, H., Zhang, P., Huang, H., Huang, J., Kao, E., Shi, L., He, L., and Yang, L., Ddi-cpi, a server that predicts drug-drug interactions through implementing the chemical-protein interactome. Nucleic Acids Res. 42:W46–W52.
22. Zhang, P., Wang, F., Hu, J., and Sorrentino, R., Label propagation prediction of drug-drug interactions based on clinical side effects. Sci. Rep. 5:12339, 2015.
23. Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P., and Sadoghi, M., Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. Web Semant. Sci. Serv. Agents World Wide Web 44:104–117, 2017.
24. Holten, D., Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Trans. Vis. Comput. Graph. 12(5):741–748, 2006.
25. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q., Line: Large-scale information network embedding. In Proceedings of

the 24th International Conference on WWW. International World Wide Web Conferences Steering Committee: 1067–1077.

26. Robertson, S., and Zaragoza, H., The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval 3(4):333–389, 2009.

27. Ho, I. W., Lee, C. T., Chen, P. W., and Lo, Y. C., Impact of cumulative antibiograms sub-categorized by origins of infection acquisition on the selection of empirical antimicrobial Therapy. Journal of Biomedical & Laboratory Sciences 27(1):10–18, 2015.

28. Zhu, G., and Iglesias, C., Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Trans. Knowl. Data Eng. 29(1):72–85, 2017.

29. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G., and Milios, E., Information retrieval by semantic similarity. International Journal on Semantic Web and Information Systems (IJSWIS) 2(3):55–73, 2006.

30. Lei, K., Yuan, K., Zhang, Q., Shen, Y., MedSim: A Novel Semantic Similarity Measure in Bio-medical Knowledge Graphs. In International Conference on Knowledge Science, Engineering and Management. Springer, Cham 479–490, 2018.

31. Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C., and Tatonetti, N. P., Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat. Protoc. 9(9): 2147–2163, 2014.

32. Sintek, M. (2003) Ontoviz tab: Visualizing protégé ontologies.

33. Ho, P. L., Wong, S.S.Y. Reducing bacterial resistance with IMPACT-Interhospital Multi-disciplinary Programme on Antimicrobial ChemoTherapy. Centre for Health Protection, 2012.