Zhang et al. BMC Medical Informatics and Decision Making 2020, 20(Suppl 4):283
https://doi.org/10.1186/s12911-020-01274-z

BMC Medical Informatics and Decision Making

RESEARCH Open Access

A semantic relationship mining method among disorders, genes, and drugs from different biomedical datasets

Li Zhang[1], Jiamei Hu[1], Qianzhi Xu[1], Fang Li[2], Guozheng Rao[3,4*] and Cui Tao[2*]

From The 4th International Workshop on Semantics-Powered Data Analytics
Auckland, New Zealand. 27 October 2019

Abstract

Background: Semantic web technology has been applied widely in the biomedical informatics field. Large numbers of biomedical datasets are available online in the resource description framework (RDF) format. Semantic relationship mining among genes, disorders, and drugs is widely used in, for example, precision medicine and drug repositioning. However, most of the existing studies focused on a single dataset. It is not easy to find the most current relationships among disorder-gene-drug relationships since the relationships are distributed in heterogeneous datasets. How to mine their semantic relationships from different biomedical datasets is an important issue.

Methods: First, a variety of biomedical datasets were converted into RDF triple data; then, multisource biomedical datasets were integrated into a storage system using a data integration algorithm. Second, nine query patterns among genes, disorders, and drugs from different biomedical datasets were designed. Third, the gene-disorder-drug semantic relationship mining algorithm is presented. This algorithm can query the relationships among various entities from different datasets.

Results and conclusions: We focused on mining the putative and the most current disorder-gene-drug relationships about Parkinson's disease (PD). The results demonstrate that our method has significant advantages in mining and integrating multisource heterogeneous biomedical datasets. Twenty-five new relationships among the genes, disorders, and drugs were mined from four different datasets. The query results showed that most of them came from different datasets. The precision of the method increased by 2.51% compared to that of the multisource linked open data fusion method presented in the 4th International Workshop on Semantics-Powered Data Mining and Analytics (SEPDA 2019). Moreover, the number of query results increased by 7.7%, and the number of correct queries increased by 9.5%.

Keywords: Semantic relationship mining, Data integration, Disorder-gene-drug relationship

* Correspondence: rgz@tju.edu.cn; cui.tao@uth.tmc.edu
[3]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]School of Biomedical Informatics, University of Texas Health Science Center at Houston, 7000 Fannin St Suite 600, Houston, TX 77030, USA
Full list of author information is available at the end of the article

Background

Semantic web technology has been applied widely in the biomedical informatics field. The resource description framework (RDF) data model is commonly used to represent data in the database. A uniform resource identifier (URI) and character strings are used to represent different entities and the relationships between entities. These semantic datasets are published online and can be accessed via the HTTP protocol and are also known as linked open datasets [1]. For example, the Life Sciences dataset is one of the most important parts of Linked Open Data Cloud [2]. This database consists of 339 RDF datasets, including 234 BioPortal datasets, 35 Bio2RDF datasets, and 70 other datasets. Together, they contain over 30 billion semantic relationships. Furthermore, a vast number of semantic relationships has been extracted from biomedical literature databases with unstructured natural language texts (e.g., MEDLINE) [3, 4]. The other existing biomedical datasets include gene-related, disorder-related, and drug-related databases. For example, PharmGKB (https://www.pharmgkb.org) [5] is a database consisting of drugs, clinical guidelines, and gene-drug and gene-phenotype relationships. The UniProt (https://www.uniprot.org/) [6] database aims to provide comprehensive and high-quality resources on protein sequences and functional information. This database comprises UniProtKB, UniParc, UniRef, and the Proteomes dataset. The Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg) database is a professional knowledge base for the biological interpretation of large-scale molecular datasets, such as genomic and metagenomic sequences [7]. The Semantic MEDLINE Database (SemMedDB) [3] (https://skr3.nlm.nih.gov/SemMedDB/index.html) is a repository of semantic predications (subject-predicate-object triples) from MEDLINE citations (titles and abstracts). This database currently contains approximately 98 million predictions from all PubMed citations (approximately 29.1 million citations, processed using MEDLINE BASELINE 2019) [8]. Over 3000 papers are added to MEDLINE every day. Therefore, new semantic relationships are constantly added to SemMedDB.

In recent decades, continuous effort has been directed to mining semantic relationships from biomedical literature text with machine learning approaches. Conditional random field (CRF) and support vector machines (SVM) have been used to mine relationships [9–11]. In [12], a new semisupervised learning method based on hidden Markov models is proposed to extract the disease candidate genes from the human genome. This method predicts genes by positive-unlabeled learning (PU-Learning). In [13], a verb-centric approach is proposed to extract relationships without a training dataset. In [14], Kilicoglu H et al. extend a rule-based, compositional approach that uses lexical and syntactic information to predict relationships.

An increasing number of graph-based mining techniques are being applied to characterize the semantic relations in semantic relation extraction tasks [15–17]. In [18], graph theory and natural language processing techniques are applied to construct a molecular interaction network to extract relationships automatically.

Deep learning models have been adapted to extract semantic relations for the biomedical domain. Moreover, this approach achieves high performance on different biomedical datasets [19]. For example, in [20], unsupervised deep learning models discovered 32% of new relationships not originally known in the UMLS. In [21], recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are fused to learn the features. RNNs and CNNs are combined for high-quality biomedical relationship extraction.

However, various associations between different datasets are likely to exist. For example, a gene in KEGG could be associated with a gene in PharmGKB. Since KEGG stores data in a different way than PharmGKB, it is time-consuming and arduous to combine the two databases directly. Overall, gene, drug, and disorder information has been stored in different heterogeneous datasets. These different datasets contain essential pieces of information for the identification of potential disorder biomarkers. Heterogeneity and fragmentation of these biomedical datasets make it challenging to quickly obtain essential information regarding particular genes, drugs, and disorders of interest. Furthermore, searching these enormous datasets and integrating the findings across the heterogeneous sources is costly and complicated [22]. Drug repositioning is one of the urgent issues that requires semantic relationship mining among genes, disorders, and drugs from different biomedical datasets for precision medicine.

Generally, these datasets provide query access for users through an application programming interface. Querying the relationships among genes, drugs, and disorders has become a research topic of increasing interest. The research on linked datasets capitalizes on the storage, management, and querying of information and promotes in-depth data analysis and data mining [23]. Semantic relationship mining among genes, disorders, and drugs is widely used, for example, in precision medicine and drug repositioning. For example, semantic relationships among diseases, drugs, genes, and variants are used to automatically identify potential drugs for precision medicine in the Precision Medicine Knowledgebase (PreMedKB) [24]. The semantic relationships between any two or more entities are queried to obtain comprehensive information. The semantic relationships among genes, disorders, drugs, and other concepts in a knowledge base can also be exploited for prioritizing drug repurposing or repositioning [25–27].

Drug repositioning is a relatively inexpensive and fast alternative to the lengthy and financially onerous task of new drug development [28]. Semantic relationship mining between a drug and other molecules or entities can also be used for drug-related knowledge discovery [29] and cooccurring entities analysis [30]. However, because these datasets could be stored in different places and in different ways, with different data formats and inconsistent representations of the same entity, the power of data mining across multiple datasets is far from being realized.

In this paper, a semantic relationship mining method among genes, disorders, and drugs from different biomedical datasets is presented. Semantic relationship mining across different biomedical datasets was performed to address this problem.

Parkinson's disease (PD) is a pervasive neurodegenerative disorder that affects approximately 6 million people worldwide. Genes play an essential role in the development of PD. Monogenic forms account for approximately 10% of all PD cases [31], while the other cases are multifactorial. An increasing number of PD loci have been identified [32]. We used PD as a case study and focused on mining the putative and most current disorder-gene-drug relationships of PD from four different biomedical datasets. We addressed some of the current challenges in the field, such as integration with different existing medical datasets and the exploitation of semantic relationship mining in real-case scenarios. This approach transcends the limitations of distributed heterogeneous data sources and results in more complete datasets in such a way that medical researchers can freely access multiple datasets across platforms. This study will impact future translational medical research.

Methods

Multisource data integration

The following life science datasets were studied in this paper: SemMedDB, KEGG, Uniprot, and PharmGKB. Different organizations publish these datasets. UMLS Metathesaurus was introduced to solve the morphology and polysemy problems. These datasets contain domain patterns for disorders (disorder), chemicals and drugs (drug) and genes and molecular sequences (gene). Figure 1 shows nine drug-disorder, gene-disorder, and drug-gene relationships.

Before mining, we converted the relational databases (including PharmGKB, KEGG, Uniprot, and SemMedDB) into the RDF data format using the D2R tool [33] to obtain the SemMedRDF, KEGGRDF, UniprotRDF and PharmGKBRDF datasets. We constructed Algorithm I to mine the semantic relationship types between SemMedRDF and other life science linked open data datasets.

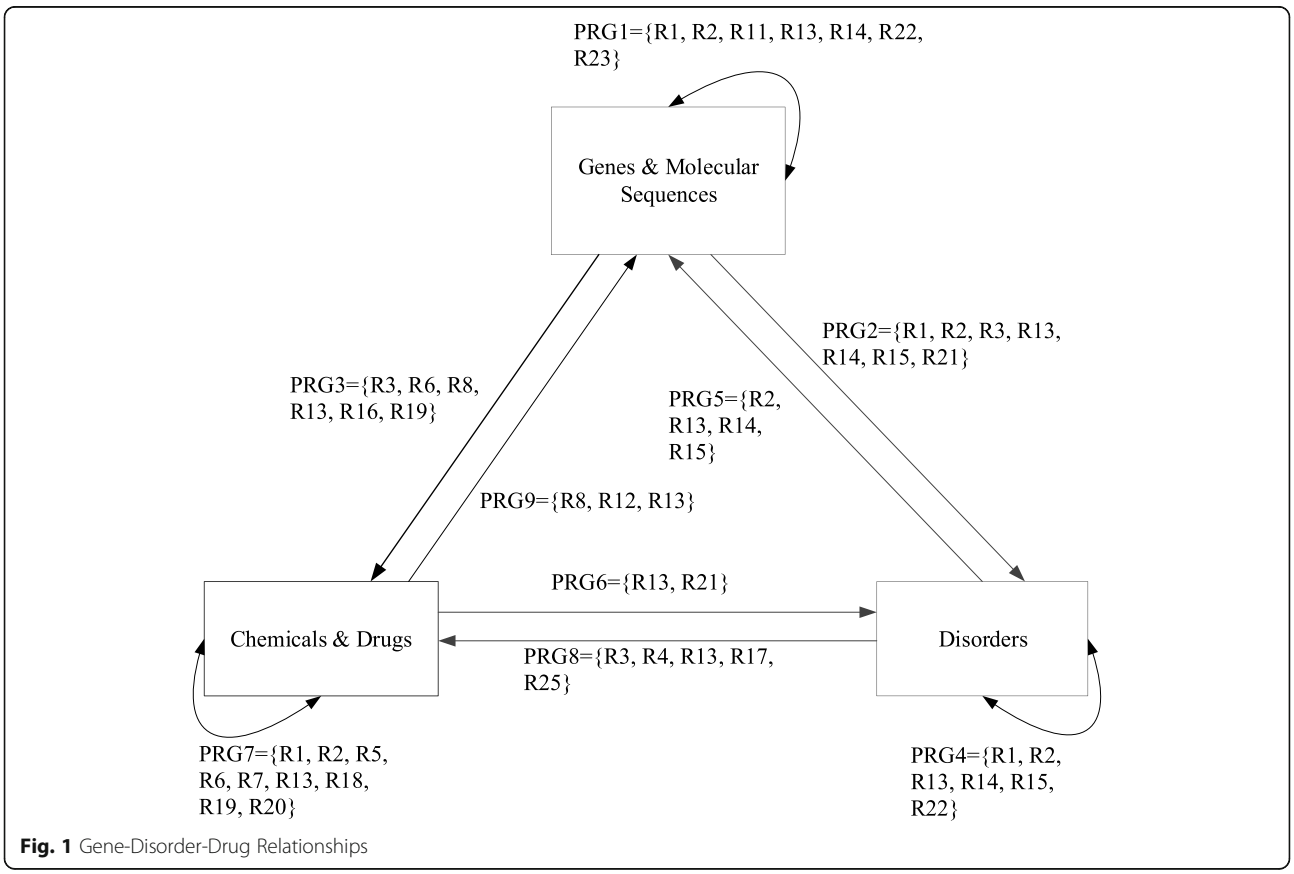

Fig. 1 Gene-Disorder-Drug Relationships

***Algorithm I*- Multi-Source Datasets Integration**

1: **Let** $\Sigma = \{\text{SemMed}RDF, \text{KEGG}RDF, \text{Uniprot}RDF, \text{PharmGKB}RDF\}$

$D2DLinks = \varnothing$, $AllPreds = \varnothing$;

2: // *Initialization: initialize myprop:UMLS-Label, myclass:UMLS-Drug, myclass:UMLS-Gene myclass:UMLS-Disorders*

All of the predicate is replaced with myprop:UMLS-Label while(getTriple(?s, rdfs:label, ?o) || getTriple(?s, SemMedRDF:name, ?o) || getTriple(?s, KEGGRDF:name, ?o) || getTriple(?s, UniprotRDF:name, ?o) || getTriple(?s, PharmGKBRDF:name, ?o))

3: // *"Predicates" extension.*

If a predicate can be found in the Metathesaurus of UMLS

Add concepts to *Allpreds* and mark it with the CUI found in the Metathesaurus of UMLS

Index *Allpreds* by predicate;

All of the objects are replaced with myclass:UMLS-Drug while (getTriple(?s, a, SemMedRDF:drug)||getTriple(?s, a, KEGGRDF:drug) || getTriple(?s, a, PharmGKBRDF: PharmGKB_Drugs))

All of the objects are replaced with myclass:UMLS-Gene while (getTriple(?s, a, SemMed*RDF*:gene) || getTriple(?s, a, KEGG*RDF*:gene) || getTriple(?s, a, Uniprot*RDF*:Gene) || getTriple(?s, a, PharmGKB*RDF*: PharmGKB_Genes))

All of the objects are replaced with myclass: UMLS-Disorder while (getTriple(?s, a, SemMedRDF:disorder) ||getTriple(?s, a, KEGGRDF: disorder) || getTriple(?s, a, PharmGKBRDF:PharmGKB_Disorders))

EndIf

4: Index Σ on (p,s,o);

5: Get the first *triple* ($triple \in \Sigma$), *Pred* (*Pred* is the Predicate of the first *triple*)

6: // *Loop 1 is designed to get all predicates: Allpreds*

Loop 1

If the first *triple* is the last triple, then Break;

ELSE

If $Pred \in Allpreds$, then Break;

ELSE

Add *Pred* to *Allpreds*;

Next triple;

Pred = getPredicate(*triple*);

EndIf

EndIf

End Loop 1

7: // *Loop 2 is designed to get all links between datasets*

Loop 2

Get the first *pred*, where $pred \in allpreds$;

If the first *pred* is the last predicate of *Allpreds*, then Break;

ELSE

Get all triples with *pred* s (where they have the same CUI);

If there are any two triples that have the same CUI of the subject and object but the namespace of the two triples is different;

Add the *pred* s and triples to *D2DLinks*;

ELSE

Get Next *pred* s with the same CUI;

EndIf

EndIf

End Loop 2

8: Add *D2DLinks* to Σ.

9: **Output** *D2DLinks*.

//: If '//' appears on the line as the first character, any data to the right of it will be ignored, as a comment.

Algorithm *I* is described step by step as follows.

The first step is variable initializations, where Σ is all data sets, including SemMedRDF, KEGGRDF, UniprotRDF and PharmGKBRDF. *Links* is a variable that saves a mined semantic relationship. Variable *AllPreds* stores the predicate of the datasets;

A compound index of *BMRDFs* is built on the predicate, subject, and object and will reduce the processing time;

The first triple is obtained from *BMRDFs*;

All of the predicates *Allpreds* are obtained from *BMRDFs*;

"Predicates" extension: If a predicate can be found in the Metathesaurus of UMLS, there will be several concepts with the same concept unique identifier (CUI), e.g., when searching the Predicate: "TREATS" in the Metathesaurus. The results are shown in Fig. 2. All of the concepts are added to *Allpreds* marking the CUI;

*Allpreds*is indexed on predicate;

The first *pred* of *Allpreds* is obtained;

If any two triples have the same CUI of the subject, predicate, and object while the namespace of the subject or object is different, this predicate will be one of the *Links*;

All of the *Links* will be added to *BMRDFs*. It will link the SemMedRDF to other biomedical datasets.

Gene-disorder-drug semantic relationship mining

GENE-DISORDER-DRUG SEMANTIC RELATIONSHIP MINING

***Algorithm II*-** Semantic Relationship Mining among Genes, Drugs and Disorders.

Requirements: Gene, Drug, Disorder is the set of all entities that belong to the gene, drug and disorder groups. Relation_gene2disorder is the set of relationships between genes and disorders. Relation_gene2drug is the set of relationships between genes and drugs, and so on.

```
For each entity g in Gene
    For each g.neighborEntity
        e = g.neighborEntity
        p = g.predication(e)    // get the predication p: (g, p, e)
        If (e ∈ Gene)
            Relation_gene2gene.add(p)
        EndIf
        If (e ∈ Drug)
            Relation_gene2drug.add(p)
        EndIf
        If (e ∈ Disorder)
            Relation_gene2disorder.add(p)
        EndIf
    EndFor
EndFor
// the same steps for each entity in Drug and Disorder
```

Fig. 2 The search results extension of Predicate: "TREATS" in UMLS

To fully understand the relationships among genes, disorders, and drugs, the following algorithm was designed to mine the attribute relationships among the three.

In Algorithm II, three entity sets are defined first: Gene, Drug, and Disorder. The relationships are defined among the three: the relational dataset from gene to disorder is called Relation_gene2disorder; the relational dataset from a gene to a drug is called Relation_gene2drug; other relational datasets can be named similarly. The algorithm to accomplish relationship querying is described as follows:

Traverse every entity in the Gene dataset;

Traverse the adjacent entity e of each entity and the predicate relationship p between the two;

If the adjacent entity e belongs to the element of Gene dataset, add the relationship p to Relation_gene2gene; if it belongs to the Drug dataset, add the relationship p to Relation_gene2drug; if it belongs to Disorder dataset, add the relationship p to Relation_gene2disorder.

Traverse each entity in the Drug and Disorder datasets to obtain the corresponding relational dataset.

Query pattern design

Nine types of relational query patterns were designed based on the gene-drug-disorder relationships in Fig. 1.

Table 1 Query patterns

No.	Query pattern
Q1	Query all genes related to a specific gene
Q2	Query all disorders caused by a specific gene
Q3	Query all drugs targeting a specific gene
Q4	Query all disorders related to a specific disorder
Q5	Query all genes causing a specific disorder
Q6	Query all drugs treating a specific disorder
Q7	Query all drugs related to a specific drug
Q8	Query all disorders treated by a specific drug
Q9	Query all genes targeted by a specific drug

These query patterns are used in many research fields [25, 26, 34]. They are shown in Table 1.

It is necessary to know the possible paths from a disorder to a drug to query the relevant drugs for a particular disorder, as shown in the relationship path in Fig. 1. For example, the algorithm designed for querying all drugs that treat a specific disorder is shown in Algorithm III. The remaining query processes can be performed in the same manner.

Algorithm III- Query all drugs that treat a specific disorder.

Require: S is the set of URIs of a disorder. Temp is the set of possible results.

```
For each triple t matchs (?s, myprop:UMLS-Label, disorder)
    S.add(getSubject(t))
End for
For each URI s in S
    For each predicate p in PRG6
        For each triple t matchs (s, p, ?o)
            If(myclass:UMLS-Drug.contains(?o))
                Temp.add(getObject(t))
            EndIf
        EndFor
    EndFor
EndFor
return Temp
```

The algorithm to query all drugs that treat a specific disorder is described as follows:

Take the disorder name entered by the user as the object, and use the customized myprop: Label as the predicate to find the subject URI set S;

The relational set from disorder to drug analyzed in the previous section is the following: Traverse each URI in set S, and use each element in as predicate to query. The object set of the query is Temp;

Traverse temp to remove the elements that are not in myclass: Drug;

Output the remaining results in Temp.

Other algorithms for related queries are similar, except that the relational set changes.

Experiments and results

Experiment dataset

Overall, any biomedical datasets can be used to mine the semantic relationships among them. Here, we demonstrated how semantically integrated RDF datasets, extracted from structured biomedical databases or linked open data, can be used to automatically mine the semantic relationships among them. SemMedDB, KEGG, Uniprot, and PharmGKB were used in the experiment.

Table 2 Predicates and their corresponding numbers

No.	Predicates
R1	sem:coexists_with
R2	sem:interacts_with
R3	sem:causes
R4	sem:prevents
R5	sem:manifestation_of
R6	sem:affects
R7	sem:occurs_in
R8	sem:associated_with
R9	kegg:hasDisease
R10	kegg:hasDrug
R11	uniprot:externalLink
R12	pharmgkb: Related_Genes
R13	pharmgkb:associated
R14	sem:stimulates
R15	sem:inhibits
R16	sem:disrupts
R17	sem:treats
R18	sem:complicates
R19	sem:predisposes
R20	sem:augments
R21	sem:produces
R22	kegg:hasPathway
R23	kegg:hasGene
R24	pharmgkb: Related_Drugs
R25	pharmgkb:c2b2r_Related_Diseases

Semantic relationship mining results

As shown in Table 2, 25 new relationships between the gene, disorder, and drug were mined from the SemMedRDF, KEGGRDF, UniprotRDF, and PharmGKBRDF datasets. As there are many relationships, the relationships in Fig. 1 were replaced by numbers, and each relationship set is represented by nine predicate relationship groups (PRG1-PRG9) in Table 3. For example, in row 2 of Table 3, the new relationships R1, R2, R11, R13, R14, R22, and R23 belong to PRG1. These relationships are also associated with the query patterns Q1. The new relationships can help us to mine more semantic relationships.

Query results

1. Q1: Query all of the genes that are related to a specific gene, PARK2. There were 95 results (genes,

Table 3 Query patterns

No.	Related predicates	PRG (Predicates relationship group) No.
Q1	R1, R2, R11, R13, R14, R22, R23	PRG1
Q2	R1, R2, R3, R13, R14, R15, R21	PRG2
Q3	R3, R6, R8, R13, R16, R19	PRG3
Q4	R1, R2, R13, R14, R15, R22	PRG4
Q5	R2, R13, R14, R15	PRG5
Q6	R13, R21	PRG6
Q7	R1, R2, R5, R6, R7, R13, R18, R19, R20	PRG7
Q8	R3, R4, R13, R17, R25	PRG8
Q9	R8, R12, R13	PRG9

proteins, and molecular sequences) related to PARK2, including PARK7, GCH1, PACRG, FBXW8, PINK1, and NBR1 (Table 4). Among them, 61 results were from SemMedDB, 23 results belonged to PharmGKB, and 11 results were from Uniprot.

2. Q2: Query all of the disorders caused by a specific gene, PARK2. There were 123 results (disorders) caused by PARK2. Some results were autosomal recessive juvenile Parkinson disease, leukemia, chronic myeloid leukemia, carcinoma of the large intestine, chronic obstructive airway disease, and chromosomal translocation. SemMedDB yielded 81 results, and another 42 results belonged to PharmGKB.
3. Q3: Query all drugs that target a specific gene, PARK2. There were 68 results (Chemicals & Drugs) that target PARK2. Some results were Cholesterol, multicatalytic endopeptidase complex, ubiquitin-protein ligase, FBXW8, and Reactive Oxygen Species. SemMedDB yielded 55 results, and another 13 results belonged to PharmGKB.
4. Q4: Query all disorders involved in a specific disorder, Parkinson's. There were 66 results (disorders) involved in Parkinson's. Some results were encephalitis, tremor, depressive disorder, hypokinesia, cognitive deficit, respiratory failure, equilibration disorder, and Lewy body disease. All of the results belonged to SemMedDB.
5. Q5: Query all of the genes that cause a specific disorder, Parkinson's. There were 28 results (Genes, protein, and molecular sequences) involved in Parkinson's. Some results were PARK1, PARK2, and CHCHD2. PharmGKB yielded 25 results, and another 3 results belonged to SemMedDB.
6. Q6: Query all of the drugs that treat a specific disorder, Parkinson's. There were 51 results (Chemicals & Drugs) involved in Parkinson's. Some results were dopamine, levodopa, dopamine transporter, and multicatalytic endopeptidase complex. SemMedDB yielded 40 results, and another 11 results belonged to PharmGKB.
7. Q7: Query all of the drugs involved in a specific drug, Levodopa. There were 79 results (Chemicals & Drugs) involved in Levodopa. Some results were Reserpine, Acetylcholine, Antipsychotic Agents, Monoamine Oxidase, Serotonin, and Isoproterenol. SemMedDB yielded 67 results, and another 12 results were from KEGG.
8. Q8: Query all of the disorders treated by a specific drug, Levodopa. There were 47 results (disorders) involved in Levodopa. Some results are Parkinson's Disease, Seborrheic dermatitis, Hepatic Encephalopathy, Hepatic Coma, Hypotension, Secondary hyperprolactinemia due to prolactin-secreting tumor, Striatonigral Degeneration, nervous system disorder, and Hypokinesia.

Table 4 Some genes related to PARK2

No.	Predicate	Object
1	<http://www4.wiwiss.fu-berlin.de/semdb/PREDICATE#COEXISTS_WITH>	<http://www4.wiwiss.fu-berlin.de/semdb/OBJECT_NAME#PARK7>
2	<http://www4.wiwiss.fu-berlin.de/semdb/PREDICATE#COEXISTS_WITH>	<http://www4.wiwiss.fu-berlin.de/semdb/OBJECT_NAME#GCH1>
3	<http://www4.wiwiss.fu-berlin.de/semdb/PREDICATE#COEXISTS_WITH>	<http://www4.wiwiss.fu-berlin.de/semdb/OBJECT_NAME#PACRGgene|PACRG>
4	<http://www4.wiwiss.fu-berlin.de/semdb/PREDICATE#COEXISTS_WITH>	<http://www4.wiwiss.fu-berlin.de/semdb/OBJECT_NAME#FBXW8>
5	<http://www4.wiwiss.fu-berlin.de/pharmgkb/ASSOCIATION#ASSOCIATED>	<http://www4.wiwiss.fu-berlin.de/pharmgkb/Entity2_NAME#PINK1>
…	…	…
95	<http://www4.wiwiss.fu-berlin.de/uniprot/EXTERNALLINK>	<http://www4.wiwiss.fu-berlin.de/uniprot#NBR1>

SemMedDB yielded 36 results, and another 11 results belonged to PharmGKB.

9. Q9: Query all of the genes that are targeted by a specific drug, Levodopa. There were 26 results (Genes, protein, and molecular sequences) involved in Levodopa. Some results were PARK1, PARK2, and CHCHD2. All of the results belonged to SemMedDB.

For the nine relationships between genes, disorders, and drugs, nine queries (Q1-Q9) were designed. Tables 5 and 6 record the source and respective proportions of each query result. To evaluate the results to improve the accuracy, we invited three professionals as domain experts to evaluate the query results. Two of these experts evaluated the results independently. The three experts provided their confidence levels ("Yes," or "No") in the query results. Each query result received the label "the correct query result" if it received more than two "Yes". Otherwise, it was labeled "a false query result". The analysis of the query results is shown in Tables 5 and 6: the column of "No" represents the nine queries. In the column of "(The number of correct queries results): (The number of queries results),", for example, in Table 4, "48: 56" means that there were 56 query results from SemMedDB for Q1 in total. Forty-eight of them received the "correct results" label. The column "Precision" means that the "The number of correct query results" out of the total "The number of query results." For example, in Table 4, "91.11" means that the "The number of correct query results" of Q1 was 91.11% (82/90).

In Tables 5 and 6, the results are mainly from SemMedDB and PharmGKB. Furthermore, some of the results are from KEGG and Uniprot. The precision of PharmGKB, KEGG, and Uniprot was 100%. The precision of SemMedDB using the method in the paper published in the ISWC SEPDA 2019 workshop [35] was 83.08% (329: 396). The precision of SemMedDB using the method in this paper was 86.44% (376: 435), which was an increase of 4.04%.

The precision of the method published in the ISWC SEPDA 2019 workshop [35] was 87.68% (477/544). The precision of the method presented in this paper was 89.88% (524/583). The precision increased by 2.51%. Furthermore, the number of query results increased by 7.7% ((583–544)/583), and the number of correct query results increased by 9.5% ((524–477)/524). That means that the method in this paper can help mine more results with increased precision.

Discussion

Strengths

It is crucial to integrate SemMedDB with other databases in this method. SemMedDB is a database of semantic predictions (subject-predicate-object triples) from MEDLINE citations (titles and abstracts). SemMedDB currently contains approximately 98 million predictions from all PubMed citations (approximately 29.1 million citations, processed using MEDLINE BASELINE 2019) [8]. Over 3000 papers are added to MEDLINE every day. Therefore, new semantic relationships are added continuously to SemMedDB. The latest relationships can help to discover new relationships for related research. Some potential recommended drugs reported in the recent literature for PD have been found in the preliminary step work on drug repositioning based on this method.

In this paper, the semantic relationship mining method is used to explore interesting, hidden, or previously unknown biomedical relationships. Twenty-five new relationships are extracted in the verification experiment. It helps to improve the results with quantity and quality. Furthermore, interesting, hidden, or previously unknown biomedical relationships can help to detect the potential relationships between drugs and diseases [20, 36].

Table 5 Analysis of the query results from [35]

No.	(The number of correct query results): (The number of query results)					Precision (%)
	SemMedDB	PharmGKB	KEGG	Uniprot	Total	
Q1	48: 56	23: 23	–	11:11	82: 90	91.11
Q2	56: 73	42: 42	–	–	98: 115	85.22
Q3	44: 52	13: 13	–	–	57: 65	87.69
Q4	54: 63	–	–	–	54: 63	85.71
Q5	–	25: 25	–	–	25: 25	100
Q6	29: 36	11: 11	–	–	40: 47	85.11
Q7	54: 61	–	12: 12	–	66: 73	90.41
Q8	25: 32	11: 11	–	–	36: 43	83.72
Q9	19: 23	–	–	–	19: 23	82.61
Total	329: 396	125: 125	12: 12	11: 11	477: 544	87.68

Table 6 Analysis of the query results from this paper

No.	(The number of correct query results): (The number of query results)					Precision (%)
	SemMedDB	PharmGKB	KEGG	Uniprot	Total	
Q1	53: 61	23: 23	–	11:11	87: 95	91.58
Q2	67: 81	42: 42	–	–	109: 123	88.62
Q3	48: 55	13: 13	–	–	61: 68	89.71
Q4	58: 66	–	–	–	58: 66	87.88
Q5	2: 3	25: 25	–	–	27: 28	96.43
Q6	34: 40	11: 11	–	–	45: 51	88.24
Q7	60: 67	–	12: 12	–	72: 79	91.14
Q8	31: 36	11: 11	–	–	42: 47	89.36
Q9	23: 26	–	–	–	23: 26	88.46
Total	376: 435	125: 125	12: 12	11: 11	524: 583	89.88

The nine types of common query patterns are proposed in the baseline method. This approach covers all semantic relationships between genes, disorders and drugs. Compared with the other models, our method can be extended to be used in more applications without a training dataset. Moreover, the method can also meet the requirements of processing large-scale data without high computational cost. The processing time increases with the size of the data linearly. It is more effective than the machine learning method, such as SemRep. In SemMedDB, the weighted average precision of the predictions is based on the number of predictions evaluated, which was approximately 0.79 [37–40]. In this paper, we used the approach in [34] to extract high-quality triples from SemMedDB. The precision increased by 2.27%.

Limitations and future effort

Since the fact that the quality of the datasets will affect the semantic relationship mining, the method has some limitations: (1) The quality of the SemMedDB should be improved in future research. (2) The quality of the other datasets depends on their creators. Thus, high-quality datasets will be selected carefully. Alternatively, we will try our best to improve the quality of the datasets selected. (3) Currently, mining semantic relationships among genes, disorders, and drugs from different biomedical datasets is the first step for precision medicine and drug repositioning. It would be desirable to mine repositioning drugs based on semantic relationships for more disorders, such as PD, Alzheimer's Disease, cancer.

Conclusions

In this paper, a semantic relationship mining method among genes, disorders, and drugs was developed. In this method, data from various biomedical datasets were first converted into RDF triples and then integrated into a system for querying nine types of common query patterns. We focused on mining the putative and latest gene-disorder-drug relationships about PD. The experiment was conducted on four different datasets. The results showed that our method has significant advantages in integrating multisource heterogeneous biomedical data. Twenty-five new relationships among genes, disorder, and drugs were identified, and most of them came from different datasets. Moreover, the precision of our method increased by 2.51%. The number of query results increased by 7.7%, and the number of correct queries increased by 9.5%. These findings demonstrate that our method is robust and reliable in mining important gene-disorder-drug relationships.

Abbreviations

CUI: Concept unique identifier; KEGG: the Kyoto Encyclopedia of Genes and Genomes; PD: Parkinson's disease; RDF: Resource description framework; SemMedDB: Semantic MEDLINE Database; URI: Uniform resource identifier

Acknowledgments

We thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments. We also thank Dr. Irmgard Willcockson for language editing.

Disclaimer

The content is solely the responsibility of the authors.

About this supplement

This article has been published as part of BMC Medical Informatics and Decision Making Volume 20 Supplement 4 2020: Selected articles from the Fourth International Workshop on Semantics-Powered Data Analytics (SEPDA 2019). The full contents of the supplement are available at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-4.

Authors' contributions

LZ and GR designed the study and drafted the original manuscript. JH and QX collected the data and performed the experiments. FL revised the manuscript. GR and CT supervised the study. All authors have read and approved the manuscript.

Funding

This research was partially supported by the National Natural Science Foundation of China (NSFC) (61373165, 61672377), Tianjin Educational Science Planning Project (HE3049), and the China Scholarship Council (CSC). This publication costs are funded by the Tianjin Educational Science Planning Project (HE3049).

Availability of data and materials

The PharmGKB is available at https://www.pharmgkb.org. The UniProt is available at https://www.uniprot.org. The KEGG is available at https://www.genome.jp/kegg. The SemMedDB is available at https://skr3.nlm.nih.gov/SemMedDB. The query results are available from the corresponding author upon request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Economics and Management, Tianjin University of Science and Technology, Tianjin 300457, China. [2]School of Biomedical Informatics, University of Texas Health Science Center at Houston, 7000 Fannin St Suite 600, Houston, TX 77030, USA. [3]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China. [4]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300350, China.

Received: 21 September 2020 Accepted: 22 September 2020
Published: 14 December 2020

References

1. Wang X, Rao G, Jiang L, Lyu X, Yang Y, Feng Z. TraPath: fast regular path query evaluation on large-scale RDF graphs; 2014.
2. Community TH on behalf of the LD. Linked Data; 2012. p. 2–6. http://linkeddata.org/. Accessed 7 July 2020.
3. Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: an advanced information management application for biomedicine. Inf Serv Use. 2011;31:15–21.
4. Percha B, Altman RB. A global network of biomedical relationships derived from text. Bioinformatics. 2018;34:2614–24.
5. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92:414–7.
6. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46:2699.
7. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2015;44:457–62.
8. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012;28:3158–60.
9. Carlson A, Betteridge J, Wang RC, Hruschka ER Jr, Mitchell TM. Coupled semi-supervised learning for information extraction. In: Proceedings of the third ACM international conference on web search and data mining; 2010. p. 101–10.
10. Hu X, Wu DD. Data mining and predictive modeling of biomolecular network from biomedical literature databases. IEEE/ACM Trans Comput Biol Bioinform. 2007;4:251–63.
11. Yousef A, Moghadam Charkari N. SFM: a novel sequence-based fusion method for disease genes identification and prioritization. J Theor Biol. 2015;383:12–9.
12. Nikdelfaz O, Jalili S. Disease genes prediction by HMM based PU-learning using gene expression profiles. J Biomed Inform. 2018;81:102–11.
13. Yang H, Swaminathan R, Sharma A, Ketkar V, D'Silva J. Mining biomedical text towards building a quantitative food-disease-gene network. In: Studies in Computational Intelligence. Berlin: Springer; 2011. p. 205–25.
14. Kilicoglu H, Rosemblat G, Rindflesch TC. Assigning factuality values to semantic relations extracted from biomedical research literature. PLoS One. 2017;12:1–20.
15. Luo Y, Uzuner Ö, Szolovits P. Bridging semantics and syntax with graph algorithms-state-of-the-art of extracting biomedical relations. Brief Bioinform. 2017;18:160–78.
16. Hu ZY, Zeng RQ, Qin XC, Wei L, Zhang Z. A method of biomedical knowledge discovery by literature mining based on SPO predications: a case study of induced pluripotent stem cells. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics); 2018. p. 383–93.
17. Workman TE, Fiszman M, Hurdle JF, Rindflesch TC. Biomedical text summarization to support genetic database curation: using semantic MEDLINE to create a secondary database of genetic information. J Med Libr Assoc. 2010;98:273–81. https://doi.org/10.3163/1536-5050.98.4.003.
18. Cairelli MJ, Fiszman M, Zhang H, Rindflesch TC. Networks of neuroinjury semantic predications to identify biomarkers for mild traumatic brain injury. J Biomed Semantics. 2015;6:1–14.
19. Peña-Torres JA, Gutiérrez RE, Bucheli VA, González FA. How to adapt deep learning models to a new domain: the case of biomedical relation extraction. TecnoLógicas. 2019;22 SPE:49–62.
20. Wang Y, Rastegar-Mojarad M, Komandur-Elayavilli R, Liu H. Leveraging word embeddings and medical entity extraction for biomedical dataset retrieval using unstructured texts. Database (Oxford). 2017;2017:1–13.
21. Zhang Y, Lin H, Yang Z, Wang J, Zhang S, Sun Y, et al. A hybrid model based on neural networks for biomedical relation extraction. J Biomed Inform. 2018;81:83–92.
22. Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics. 2015;16:55.
23. Li W, Zhang B, Rao G, Chen R, Feng Z. Hash tree indexing for fast SPARQL query in large scale RDF data management systems. In: CEUR workshop proceedings; 2017. p. 1–4.
24. Yu Y, Wang Y, Xia Z, Zhang X, Jin K, Yang J, et al. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Res. 2018;47:D1090–101.
25. Malas TB, Vlietstra WJ, Kudrin R, Starikov S, Charrout M, Roos M, et al. Drug prioritization using the semantic properties of a knowledge graph. Sci Rep. 2019;9:6281.
26. Tian Z, Teng Z, Cheng S, Guo M. Computational drug repositioning using meta-path-based semantic network analysis. BMC Syst Biol. 2018;12:134.
27. Vlietstra WJ, Zielman R, van Dongen RM, Schultes EA, Wiesman F, Vos R, et al. Automated extraction of potential migraine biomarkers using a semantic graph. J Biomed Inform. 2017;71:178–89.
28. Kumar R, Harilal S, Gupta SV, Jose J, Thomas DG, Uddin MS, et al. Exploring the new horizons of drug repurposing: a vital tool for turning hard work into smart work. Eur J Med Chem. 2019;182:111602.
29. Gachloo M, Wang Y, Xia J. A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition. Genomics Inform. 2019;17(2):e18.
30. Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In: Pacific Symposium on Biocomputing. Fairmont Orchid, Big Island of Hawaii: World Scientific; 2012. p. 422–33.
31. Jenner P, Morris HR, Robbins TW, Goedert M, Hardy J, Ben-Shlomo Y, et al. Parkinson's disease--the debate on the clinical phenomenology, aetiology, pathology and pathogenesis. J Parkinsons Dis. 2013;3:1–11.
32. Konovalova EV, Lopacheva OM, Grivennikov IA, Lebedeva OS, Dashinimaev EB, Khaspekov LG, et al. Mutations in Parkinson's disease-associated PARK2 gene are accompanied by imbalance in programmed cell death systems. Acta Nat. 2015;7:146–51.
33. D2R Server: Accessing databases with SPARQL and as Linked Data. http://d2rq.org/d2r-server. Accessed 7 July 2020.
34. Cong Q, Feng Z, Li F, Zhang L, Rao G, Tao C. Constructing Biomedical Knowledge Graph Based on SemMedDB and Linked Open Data. In: Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018. Madrid: IEEE; 2019. p. 1628–31.
35. Rao G, Zhang L, Zhang X, Li W, Li F, Tao C. A multi-source linked open data fusion method for gene disorder drug relationship querying. CEUR Workshop Proc. 2019;2427:31–5.
36. Rao G, Gao J, Li Z, Cong QFZ. QSICPM: a novel quantitative semantic interaction calculation prediction method for drug repositioning based on biomedical literature semantic data; 2019.
37. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. AMIA Annu Symp Proc. 2003;2003:554–8.
38. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. In: Scientific data: World Scientific; 2016. p. 209–20. https://doi.org/10.1038/sdata.2016.35.

39. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003;36:462–77.
40. Fiszman M, Demner-Fushman D, Lang FM, Goetz P, Rindflesch TC. Interpreting comparative constructions in biomedical text. In: ACL 2007 - Proc Work BioNLP 2007 Biol Transl Clin Lang Process, vol. June; 2007. p. 137–44. https://doi.org/10.3115/1572392.1572417.

Publisher's Note