Interconnection of Biological Knowledge Using NikkajiRDF and Interlinking Ontology for Biological Concepts

Tatsuya Kushida¹ · Kouji Kozaki² · Takahiro Kawamura³ · Yuka Tateisi¹ · Yasunori Yamamoto⁴ · Toshihisa Takagi^{1,5}

Received: 10 February 2019 / Accepted: 14 September 2019 / Published online: 27 September 2019
© The Author(s) 2019

Abstract

We investigated the interconnection on knowledge of biological molecules, biological phenomena, and diseases to efficiently collect information regarding the functions of chemical compounds and gene products, roles, applications, and involvements in diseases using knowledge graphs (KGs) developed from Resource Description Framework (RDF) data and ontologies. NikkajiRDF linked open data provide information on approximately 3.5 million chemical compounds and 694 application examples. We integrated NikkajiRDF with Interlinking Ontology for Biological Concepts (IOBC), including approximately 80,000 concepts, information on gene products, drugs, and diseases. Using IOBC's ontological structure, we confirmed that this integration enabled us to infer new information regarding biological and chemical functions, applications, and involvements in diseases for 5038 chemical compounds. Furthermore, we developed KGs from IOBC and added protein, biological phenomena, and disease identifiers used in major biological databases: UniProt, Gene Ontology, and MeSH to the KGs. Using the extended KGs and federated search to the DisGeNET, we discovered more than 60 chemicals and 700 gene products, involved in 32 diseases.

Keywords Chemical compound · Disease · Gene product · Inference · Knowledge Graph · Ontology

Introduction

Information on functions and physicochemical qualities of biological molecules, such as chemical compounds and gene products, is essential for not only elucidating and recognizing biological phenomena but also the development of various biobased products, for example, drugs, foods, and materials. A simple method to

✉ Tatsuya Kushida
kushida@biosciencedbc.jp

Extended author information available on the last page of the article

Fig. 1 Integration of NikkajirDF with major databases of chemical compounds using InChI/InChIKey

collect and leverage information was studied to prepare a rich research environment for researchers, developers, and engineers. We investigated the interconnection of biological knowledge for the development of chemical compounds, drugs, gene products, diseases, and biological phenomena. The goal was to retrieve reliable information on chemical compounds, drugs, and gene products using knowledge graphs (KGs) developed from biological ontology and the Resource Description Framework (RDF) data.

We developed NBDC NikkajirDF from the Japan Chemical Substance Dictionary (Nikkaji) [1], which is one of the largest databases of chemical compounds in Japan [2, 3]. Nikkaji includes 3.5 million chemical compounds, of which 6,454 have at least one of the 694 application examples (e.g., “hypotensive drug,” “artificial colorant”). NikkajirDF uses InChI and InChIKey as unique chemical identifiers and InChI, developed by the International Union of Pure and Applied Chemistry and National Institute of Standards and Technology, as a non-proprietary identifier of chemical compounds [4]. InChIKey is a hashed version of the full InChI. The InChI/InChIKey assists in simplifying mapping between chemical database IDs and facilitates a collection of the corresponding chemicals. NikkajirDF uses standard ontologies in PubChem [5] and ChEMBL [6]. These ontologies include Chemical Information Ontology [7] and SemanticScience Integrated Ontology (SIO) [8]. Consequently, users can perform SPARQL searches with these ontologies (Fig. 1). NikkajirDF links chemical compounds of more than 30 other databases that share the same InChIKey. We developed the RDF triples to link these compounds following UniChem [9] work and skos:closeMatch. Users can download the RDF data from

the Life Science Database Archive [2] and NBDC RDF Portal [3] websites. The SPARQL search can be performed using the endpoint [10].

Interlinking Ontology for Biological Concepts (IOBC), previously referred to as the “Refined JST thesaurus” [11], contains approximately 80,000 biological concepts, including biological phenomena, diseases, molecular functions, gene products, chemical compounds, drugs, and medical procedures. It also contains approximately 20,000 related concepts in basic chemistry and environmental science [12]. The concepts are structured by “subclass of” and 35 additional relations, for example, “has function,” “has role,” “has quality,” and “is participant in.” Each concept is labeled in both English and Japanese. We can browse and download the ontology from the BioPortal [13] homepage [12] to prepare the SPARQL endpoint [14].

Information on chemical compounds, drugs, and gene product functions/roles/applications is crucial in developing pharmaceutical products and discovering new materials for medical treatment. NikkajiRDF consists of a significant number of InChI/InChIKey chemical compounds. However, it lacks information on the functions/roles/applications. On the contrary, IOBC contains various biological phenomena, including diseases, chemical compounds, drugs, and gene products. However, these items lack the unique identifiers, such as InChI/InChIKey and Protein IDs (e.g., UniProtKB accession number [15]), used for easy mapping of biological molecules and drugs of other data resources. These data sources should be combined to efficiently collect the functions/roles/applications.

In addition to information on chemical compounds [16], this study¹ aimed to collect and interconnect biochemical and genomic knowledge to find drugs and biological molecules, such as gene product information by combining NikkajiRDF, IOBC, and other open-source data. Using ontological knowledge and unique identifiers, InChI/ InChIKey, UniProtKB accession number, and GeneID, helps infer the functions/roles/applications of a larger number of chemical compounds and gene products.

The rest of the paper is organized as follows: Sect. 2 reviews related works that describe representative open-source knowledge and ontologies to collect information on the functions/roles/applications of biological molecules. Section 3 describes the inference of the chemical compounds’ functions/roles/applications through combinations of NikkajiRDF, ChEBI, and IOBC. Section 4 presents a method of creating KGs from IOBC and extending the KGs using existing external databases, thesauri, and ontologies. We also demonstrate the inference of chemical compounds and gene products in biological phenomena and diseases using the KGs. Section 5 summarizes our conclusions and discusses future work.

¹ This study is the extension of our previous work which was published in JIST2018 [16].

Related Works

ChEBI is a major chemical database and ontology [17] of approximately 90,000 chemical compounds, identified through InChI and InChIKey, performing 1,000 roles and applications. ChEBI is used frequently in annotating and classifying chemical compounds through InChI/InChIKey in various databases: PubChem and ChEMBL. However, the number of chemical compounds in ChEBI is lacking even in comparison with that of other chemical databases such as NikkajiRDF, which contains information on approximately 3.5 million chemical compounds. Thus, preparing the knowledge bases and establishing a method to infer the functions/roles/applications of many chemical compounds is necessary.

DBpedia is a project that extracts structure information from Wikipedia [18] using RDF [19]. Wikidata is a knowledge base that allows every user to extend and edit stored information [20]. Although DBpedia and Wikidata are used widely for cross-domain knowledge, they have recently been attempting to integrate chemical information [19, 21]. DBpedia and Wikidata contain information on approximately 18,000 and 150,000 chemical compounds, respectively. However, these numbers are fewer than those of NikkajiRDF, PubChem, or ChEMBL.

From the DBpedia [22] and Wikidata [23] public SPARQL endpoints, users can collect information on biological and chemical functions/roles/applications to perform SPARQL queries. However, DBpedia uses only annotation properties: “dcterms:subject,” and “rdfs:seeAlso” to describe the information, instead of specific properties such as “has function (sio²:SIO_000225)” and “has role (sio:SIO_000228).” For example, the roles and applications of “Caffeine,” such as “Anxiogenics” and “Effect_of psychoactive_drugs_on_animals,” are described as objects of “dcterms:subject” and “rdfs:seeAlso,” respectively. Moreover, these properties include information outside their functions/roles/applications, for example, the categories.

This shows that users must select the information manually. Wikidata also faces the same problems as it uses a property “wdt³:P31 (instance of)” to describe the functions/roles/applications of compounds. The objects of the property include information outside its functions/roles/applications. Hence, we are incorporating some specific properties: “has function” and “has role,” into Wikidata to describe information. Therefore, DBpedia and Wikidata are currently neither reasonable nor suitable for the efficient collection of functions/roles/applications of chemical compounds.

ChEMBL and PubChem are the major chemical compounds’ databases that offer downloadable RDF data. ChEMBL provides the public SPARQL endpoint to collect the original data. Currently, PubChem does not provide the public SPARQL endpoints; however, the PubChem Classification Browser [24] and PUG REST [25] are available to search for and collect information.

² <http://semanticscience.org/resource/>.

³ <http://www.wikidata.org/prop/direct/>.

UniProt is a large protein knowledge base, providing information on functions, subcellular locations, molecular interactions, structures, amino acid sequences, similar proteins, and so on. Many biological databases adopt the protein identifier: UniProtKB accession number. The public SPARQL endpoint offers available data [26], and RDF data are effective in exploring life sciences.

DisGeNET [27] is a database that contains gene–disease associations, collected by expert human curation and through text-mining methods from many public data sources and the scientific literature. We can retrieve RDF data of the gene–disease associations from the SPARQL endpoint under the Open Database License [28].

Open Pharmacological Concepts Triple Store (Open PHACTS) [29], Bio2RDF [30], Chem2Bio2RDF [31], and RIKEN MetaDatabase [32] are databases for research and development to collect information on chemical compounds and gene products, integrated using semantic technologies such as RDF. Researchers and engineers can retrieve and leverage innovative drug discovery information from these databases.

Open PHACTS is an open innovation platform for drug discovery. Using semantic approaches, several linked open data, such as ChEMBL, Human Disease Ontology [33], and WikiPathways [34], are integrated. Information on chemical targets, assays, biological activities, and diseases is retrieved using keyword search, API, and Apps. Data are provided in various formats: RDF/Turtle, JSON, and XML. However, the Open PHACTS Linked Data API and associated services were closed in March 2019.

Bio2RDF applies semantic web technology to integrate life-science databases. Public databases, such as NCBI's Entrez Gene [35], Online Mendelian Inheritance in Man (OMIM) [36], Kyoto Encyclopedia of Genes and Genomes (KEGG) [37], and DrugBank [38], are converted to the RDF format through RDF conversion programs from XML, SQL, and TEXT. On the project page, RDF data are accessible from the SPARQL endpoint [39].

Chem2Bio2RDF is a project that collects information on chemical compounds/drugs and proteins/genes through the chemogenomics approach. The datasets include information on protein–protein interactions, diseases, side effects, and literature, linked to Bio2RDF, and Linked Open Drug Data [40]. This was designed for polypharmacology, pathway inhibition, and adverse drug reaction analysis. At present, Semantic Link Association Prediction [41] for drug target prediction based on Chem2Bio2RDF datasets is available; however, its SPARQL endpoint is unavailable.

RIKEN MetaDatabase is an RDF platform containing Riken's original databases, Bioresources (e.g., FANTOM [42], mouse resources [43]), and external databases (e.g., PDB [44]). Using standard ontologies (e.g., SIO, and Phenotype and Trait Ontology [45]), users can collect the metadata linked to other datasets.

Comparing our IOBC-leveraged project and other datasets with the above-mentioned projects, our proposed datasets have the following features: (1) They contain the relationships between instance-level information on various types of life-science knowledge (e.g., a relationship between a biological phenomenon: Fibrinolysis, and the succeeding disease: Fibrinolytic purpura, in Fig. 4 in Sect. 4.1). (2) In our dataset, IOBC serves as a hub for integrating various life-science concepts, such as

chemical compounds, gene products, biological phenomena, and diseases (Fig. 5 in Sect. 4.2). (3) Using the ontological structures and KGs (Figs. 6, 7, 8, 9, and 10 in Sect. 4.3), IOBC and other ontologies can infer new facts (e.g., biological molecular functions) from the integrated information.

Inference of Chemicals' Functions/Roles/Applications Using Ontological Structure

Inference of Chemical Compounds for Functions/Roles/Applications Using ChEBI and NikkajiRDF

In this section, we infer the functions/roles/applications of chemical compounds using linked open data and ontologies. NikkajiRDF has approximately 3.5 million chemical compounds; however, most of them lack application examples. We attempt to integrate NikkajiRDF with ChEBI using InChIKey to add information to NikkajiRDF chemical compounds based on ChEBI's roles and applications. Prior to that, ChEBI [46] and NikkajiRDF data [47] were stored in a triple store and the SPARQL execution was prepared. Consequently, 280 ChEBI roles and application terms could be assigned to 2,926 NikkajiRDF chemical compounds. Next, ChEBI's roles/applications were inferred to NikkajiRDF's chemical compounds using ChEBI's ontological structure. The following SPARQL query was performed.

```
prefix obo: <http://purl.obolibrary.org/obo/>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix nikkaji: <http://nikkaji.biosciencedbc.jp/>
SELECT distinct ?nikkaji ?chebi ?role
WHERE
{
  # NikkajiRDF.ttl
  graph <nikkaji:link2OtherDBs_basedOnUniChem20180515> {
    ?nikkaji skos:closeMatch ?chebi .
  }
  # ChEBI.owl
  graph <nikkaji:ChEBI2018001.owl.gz> {
    ?upperchebi rdfs:subClassOf
      [owl:someValuesFrom ?role ;
       owl:onProperty obo:RO_0000087] .
    ?chebi rdfs:subClassOf* ?upperchebi .
  }
}
```

Figure 2 shows the inference of the roles and applications of NikkajiRDF's chemical compound "Aspirin" using ChEBI. The inference process is as follows: (1) it was found that ChEBI's chemical compounds had the same InChIKey as NikkajiRDF's chemical compounds using the property skos:closeMatch (e.g., ChEBI's "acetylsalicylic acid" and NikkajiRDF's "Aspirin") in the NikkajiRDF

-draw. chebi: <http://purl.obolibrary.org/obo/>

At least one of the 1062 ChEBI role and application terms was assigned to each of the 18,386 NikkajiRDF chemical compounds through the ChEBI ontological structure. This indicates that the number of NikkajiRDF chemical compounds and roles/applications increased by approximately three times after inference. The reason is that 6,454 chemical compounds had at least one of the 694 applications, corresponding to ChEBI's roles/applications. This result is downloadable [48].

Inference of Chemical Compounds for Functions/Roles/Applications Using IOBC and NikkajiRDF

As mentioned previously, NikkajiRDF has approximately 3.5 million chemical compounds; however, IOBC has 17,180 organic chemicals, inorganic chemicals, and drugs, which do not contain InChI/InChIKey. A total of 5,781 of these chemical compounds has information on biological and chemical functions (e.g., Apoptosis [iobc:200906039143928462]), roles (e.g., antirheumatic drug [riobc:200906008284879667]), and chemical involvements in biological phenomena and diseases (e.g., hepatitis B [iobc:200906000547096041]). In particular, information on the chemical compounds in biological phenomena is unique to IOBC.

We implemented a Lexical OWL Ontology Matcher (LOOM) algorithm [49] to match the labels between the NikkajiRDF and IOBC chemical compounds. LOOM is a simple lexical algorithm to produce mappings. It takes two ontologies from a Semantic Web ontology language and produces pairs of related concepts from two ontologies. The label-comparison function removed delimiters such as spaces, underscores, and parentheses. Then, it used an approximate string comparison technique to mismatch one character in strings with length greater than four and no mismatches for shorter strings [49]. The LOOM algorithm is widely used in the field of life sciences such as BioPortal because it exhibits high performance in terms of the precision of the mappings [49], and it is also easy to implement in systems.

In our project, two life-science experts reviewed the algorithm results. If they found false-positive errors, they removed them. If their opinions were divided, they discussed them, and selected one of the opinions. In contrast, we did not evaluate false-negative errors of the mapping, because acquiring the information to calculate them was difficult.

As a result of executing the LOOM algorithm, in total, 10,576 NikkajiRDF chemical compounds were incorporated into IOBC. Two experts reviewed the results of the mapping algorithm, and they found 68 false-positives, which were subsequently removed. In this case, there was no difference of opinion among experts. The precision rate of the LOOM algorithm was 0.99 (10,508/10,576). For example, NikkajiRDF contained two entries whose labels were “HMDP,” namely, sti:200907088719956119 and sti:200907015329956587, whereas IOBC contained an entry whose label was “HMDP,” iobc:200906046710073151. In this case, the experts confirmed that iobc:200906046710073151 corresponded to sti:200907015329956587 through database descriptions such as using their structure information.

Euzenat and Shvaiko [50] have classified ontology matching (mapping) algorithms into two types: element-level techniques and structure-level techniques. Moreover, they have subclassified the former into five categories including string-based techniques and formal resource-based techniques; in contrast, the latter has been subclassified into four categories including graph-based techniques and

⁴ <http://purl.jp/bio/4/id/>.

⁵ <http://stirdf.jst.go.jp/id/>.

taxonomy-based techniques. Harrow [51] demonstrated some applications of ontology mapping in the fields of biomedical science.

We focused on taxonomy-based techniques, which utilized information on the upper concepts in ontological structures. Then, we conducted a preliminary experiment to compare the performance of only the LOOM algorithm with that of the combination of LOOM and taxonomy-based techniques to gauge any improvement in the ontology mapping. If chemical compounds with defined structures in NikkajiRDF and IOBC comprise basic chemical structures, such as phosphonic acid and polynuclear aromatic compounds, the chemical compounds can be related to the basic chemical structures using *skos:broader*. In this preliminary experiment, we examined whether the number of 68 earlier false-positives produced by only performing the LOOM algorithm would be effectively decreased using not only label information but also basic chemical structures.

Consequently, we confirmed that the utilization of both chemical structures and label information decreased the number of errors to 59, removing 9 false-positives, which is an indicative of the improved precision rate. For example, as mentioned earlier, this improvement can be seen in the case of two NikkajiRDF chemical compounds that have the label HMDP, namely *sti:200907088719956119* and *sti:200907015329956587*, and an IOBC chemical compound that has the same label namely *iobc:200906046710073151*. In addition, both *sti:200907015329956587* and *iobc:200906046710073151* have a common basic chemical structure “phosphonic acid”; in contrast, *sti:200907088719956119* does not have the mentioned structure. Therefore, we have confirmed that both, the NikkajiRDF chemical compound “*sti:200907015329956587*” and the IOBC chemical compound “*iobc:200906046710073151*,” were the same chemical compound. Results obtained using chemical compound mapping were equivalent to those derived from expert manual curation based on the structural information on these chemical compounds.

Furthermore, by appropriately leveraging ontology mapping algorithms mentioned above for biomedical concepts, we would be able to discover new relations among biomedical concepts, such as those of equivalent and overlapping relations, which could not be identified using only string comparison techniques, such as the LOOM. For example, there is an ontology mapping system “AgreementMaker-Light (AML) [52],” which implements some matching algorithms: (1) “The LexicalMatcher” to find literal full name matches between the lexicon entries of two ontologies, (2) “The ThesaurusMatcher,” to find literal full name matches involving synonyms inferred from an automatically generated thesaurus, and (3) “The XRefMatcher,” which uses cross-reference information among data sources. In the AML’s matching tasks using anatomy, phenotype, and disease datasets, they have demonstrated that not only the precision rate but also recall rate and F-measure were improved, simply by optimizing the algorithm parameters or combining some algorithms [52].

Furthermore, using the IOBC ontological structure, at least one of the 432 biological and chemical functions, roles, and chemical involvements in biological phenomena could be inferred for 5038 extended chemical compounds (Fig. 3 and Table 1). Inference using the ontology enabled the assignment of more chemical compound

Fig. 3 Inference of the biological and chemical functions, roles, and chemical involvements in the biological phenomena of IOBC's chemicals derived from NikkajiRDF. It is inferred that “Dopamine” would be involved with “Catecholamine cardiomyopathy” with which the upper class “catecholamine” is involved. This diagram is visualized on a web service: <https://www.kanzaki.com/works/2009/pub/graph-draw>

Table 1 Inference results of IOBC chemical compounds' functions, roles, and involvements in biological phenomena of the inheritance from upper-class chemical compounds

Inference	Has function	Has role	Is participant in
No	9 functions to 15 chemical compounds	343 roles to 4814 chemical compounds	32 biological phenomena to 33 chemical compounds
Yes	15 functions to 187 chemical compounds ^a	368 roles to 4839 chemical compounds ^b	49 biological phenomena to 594 chemical compounds ^c

“No” of the column “Inference” indicated that a chemical's functions, roles, or involvements were assigned to properties, “has function,” “has role,” and “in the participant of” the chemical compounds. “Yes” indicated that a chemical's functions, roles, or involvements were inferred using them of upper chemical compounds in addition to the results of the case of the “No.”

^aE.g., “Barbital” inherits a function “hypnotic action,” which the upper class “sedative hypnotic” originally possessed

^bE.g., “Lysergic Acid” inherits a role “antiparkinson drug,” which the upper class “ergot alkaloid” originally possessed

^cCf. Fig. 3

functions, roles, and involvements in biological phenomena than that obtained by not using the ontology. For the cases of “is participant in” and Inference: Yes in Table 1, the SPARQL query and result are available in [53].

Creating KGs from IOBC

Aside from the “is-a” and “whole-part” relationships, we leverage more than 30 relations within IOBC for functions/roles/applications/qualities of chemical compounds, drugs, and gene products. The primary focus was on the relationships between a preceding biological phenomenon (e.g., Fibrinolysis [iobc:200906057747871335]) and the succeeding disease (e.g., Fibrinolytic purpura [iobc:200906056051568500]). The relationships were described using a property “precedes [rxkos: precedes]” within the IOBC (Fig. 4). Gene products, which regulated or promoted a biological phenomenon and preceded a disease, were claimed to be potential candidates for disease-related gene products. IOBC has 35 properties,

such as “has function,” “precedes,” and “is participant in,” to describe the relationships between the concepts [11, 33]. It is possible to precisely discover potential candidate genes by performing a SPARQL search.

In another study [55], we developed KGs: Fibrinolysis network (Fig. 4) [56] and Bone metabolic turnover network (BMT network) [57] from IOBC. A SPARQL query was performed to create the KGs. Each of the KGs was constructed as collections of concepts connected with “Fibrinolysis” and “BMT [iobc:200906094913122330]” within three steps, respectively. Next, we stored them in a triple store. Then, we inferred chemical compounds with diseases from both the KGs.

In Sect. 4, in addition to chemical compounds, we inferred gene product involvements in biological processes and diseases using the Fibrinolysis network and BMT network. The involvements of diseases in any chemical compound and gene product can be inferred using disease information preceding biological phenomena.

Extending the KGs using existing databases, thesauri and ontologies

IOBC contains various biological concepts, such as chemical compounds, gene products, proteins, biological processes, and diseases. However, these concepts did not have sufficient external links to other databases, thesauri, and ontologies. Thus, in the Fibrinolysis, and the BMT network, which comprised 181 IOBC’s concepts in total, we executed the LOOM algorithm [49] (see Sect. 3.2) to match the labels and synonyms of resources between the IOBC and major RDF data (e.g., ChEBI, PubChem, ChEMBL, Medical Subject Headings (MeSH) [58] using UniProt and Gene Ontology (GO) [59]) with a SPARQL search. Two experts confirmed the results, and manually removed 1 false-positive. In this case, there were no differences in opinion among experts. The precision rate of the LOOM algorithm was 0.99 (461/462). From the true-positive data, we created triples between IOBC and other RDF resources using a property `skos:exactMatch`. We used both original URIs as identifiers of the resources (e.g., <http://purl.uniprot.org/uniprot/P02675>) and URIs corresponding to the original ones, provided by Identifiers.org [60] (e.g., <http://identifiers.org/uniprot/P02675>) in the triples. Next, we stored them in the triple store that contained IOBC.

We collected relationships between GO concepts such as biological processes, and the related human proteins provided by UniProt from AmiGO 2 [61]. From the collected relationships, we created triples using a property “has function [rsio:SIO_000225] (e.g., “uniprotkb:P05155 [rSERPING1]” “has function” “Fibrinolysis [go:GO_0042730]”). For the resource’s URIs, we used both the original URIs and the URIs provided by Identifiers.org. Finally, we stored the triples into the triple store. Consequently, the KGs consisted of IOBC’s concepts and the corresponding concepts derived from other RDF data (e.g., UniProt) (Fig. 5). By performing a federated SPARQL search to the endpoints (e.g., UniProt SPARQL endpoint [26]), we interconnected the IOBC’s KGs and other RDF data.

Fig. 5 Data schema of the IOBC's KG extended by other data sources

Inference of Chemicals for Functions/Roles/Applications Using KG

In the extended KGs, the Fibrinolysis network, and the BMT network, we performed the following SPARQL search to infer chemical compounds and gene products' involvement in diseases.

```

prefix dcterms: <http://purl.org/dc/terms/>
prefix sio: <http://semanticscience.org/resource/>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix subject: <http://purl.jp/bio/4/subject/>
prefix xkos: <http://rdf-vocabulary.ddialliance.org/xkos#>
SELECT distinct ?label_chem ?label_event ?label_disease
WHERE
{
  #finding pre-biological events of diseases
  {?event xkos:precedes ?disease .}
  UNION {?event sio:SIO_000062 ?disease .}
  #finding chemicals regulating biological events
  {?chem sio:SIO_000062 ?event .}
  UNION {?chem sio:SIO_000225 ?event .}
  ?disease dcterms:subject subject:LS51 . #diseases
  ?chem skos:prefLabel ?label_chem.
  ?event skos:prefLabel ?label_event.
  ?disease skos:prefLabel ?label_disease.
}

```

Fig. 6 Associations between chemicals, namely chemical compounds, drugs and gene products, and diseases in the Fibrinolysis network (1/2)

Fig. 7 Associations between chemicals, namely chemical compounds, drugs and gene products, and diseases in the Fibrinolysis network (2/2)

Table 2 PubChem SIDs, ChEBI IDs, and MeSH Unique IDs for the disease-related chemical compounds inferred from the Fibrinolysis network

Chemical name	PubChem SID	ChEBI	MeSH Unique ID
Vitamin K	53787905		D014812
Aminocaproic acid	53788266	16586	D015119
Vitamin K 2	53789135		D024482
Vitamin K 3	53790403		D024483
Vitamin K 1	56365821		D010837
4-Aminomethylbenzoic acid	134223847		C015668
Tranexamic acid	134224929	48669	D014148
Anagrelide		142290	
Anagrelide hydrochloride		55345	
Antifibrinolytic drug		48675	
Fibrin fragment D			C036309
Antifibrinolytic agents			D000933
Alpha-2-antiplasmin			D000979
Blood coagulation factors			D001779
Fibrin			D005337
Blood coagulation factor inhibitors			D019774

Consequently, we discovered 7 PubChem substances, 5 ChEBI compounds and drugs, 13 MeSH chemicals, 325 UniProt proteins (e.g., uniprotkb:P05155), and 7 CompexPortal complexes strongly involved in 16 kinds of diseases (e.g., Fibrinolytic purpura) in the Fibrinolysis network (Figs. 6 and 7) based on the ontological structures, and relationships (e.g. *rdfs:subClassOf*). In the BMT network (Figs. 8, 9, and 10), we discovered 39 PubChem substances, 1 ChEBML compound, 2 ChEBI compounds and drugs, 51 MeSH chemicals, 377 UniProt proteins (e.g., uniprotkb:Q99572), and 6 RNAcentral ncRNAs strongly involved in 15 kinds of diseases (e.g., Osteolysis).

We discovered 5 chemical compounds related to Fibrinolytic purpura, namely, anagrelide (chebi:CHEBI_142290), anagrelide hydrochloride (chebi:CHEBI_55345), 6-aminohexanoic acid (chebi:CHEBI_16586), and Tranexamic acid (chebi:CHEBI_48669) in the Fibrinolysis network. However, we did not confirm that these relationships were from Bio2RDF, Chem2BioRDF, or RIKEN MetaDatabase. The chemical compounds and gene products discovered are the potential candidates. Future studies should validate these inferred results biologically and clinically.

Furthermore, we investigated whether these disease-related chemical compounds, which were inferred in the Fibrinolysis network (Fig. 6 and 7), have been authorized as disease drugs using the comparative toxicogenomics database (CTD) [62] and PubChem. Consequently, we confirmed that all of the disease-related chemical compounds were not authorized as drugs for the inferred diseases, such as Fibrinolytic purpura (Fig. 6), and in which clinical trials were also not conducted. Table 2 shows that the PubChem SIDs, ChEBI IDs, and

Table 3 Information on the therapeutic uses and clinical trials for disease-related chemical compounds inferred from the Fibrinolysis network (1/2)

Chemical name	PubChem CID	Therapeutic uses	Clinical trials
Vitamin K	5280483	Yes	Yes
Aminocaproic acid	564	Yes	Yes
Vitamin K 2	5282367		Yes
Vitamin K 3	4055	Yes	Yes
Vitamin K 1	5284607	Yes	Yes
4-Aminomethylbenzoic acid	65526		Yes
Tranexamic acid	5526		Yes
Anagrelide	135409400	Yes	
Anagrelide hydrochloride	135413494	Yes	
Antifibrinolytic drug			
Fibrin fragment D			
Antifibrinolytic agents			
alpha-2-Antiplasmin			
Blood coagulation factors			
Fibrin			
Blood coagulation factor inhibitors			

In the PubChem therapeutic uses and clinical trials columns, “Yes” indicates that the corresponding information has been reported for at least any one disease (except for inferred diseases) and blanks indicate that the information has not been reported yet

MeSH Unique IDs for the disease-related chemical compounds inferred from the Fibrinolysis network. Moreover, life-science experts manually collected the chemical identifiers that linked to the information on therapeutic uses and clinical trials such as PubChem CIDs, DrugBank IDs, and CTD IDs from the internal and external links of the PubChem SIDs, PubChem CIDs, and MeSH Unique IDs, respectively. Tables 3 and 4 summarize the information on disease-related chemical compounds on the PubChem therapeutic uses and clinical trials, whether they are categorized as Approved (A) or Investigational (I) in the DrugBank, and Therapeutic (T) or Marker/Mechanism (M) in the CTD (see Table 4 legend) for at least any one disease (except for inferred diseases.)

Chemical compounds that have information on the PubChem therapeutic uses and categorized as “A” in the DrugBank or “T” in the CTD have been used in medical treatment. Thus, confirming the medical efficacy, we expect to decrease the drug development cost and the period because human toxicity tests and pharmacokinetic studies have been already performed on the chemical compounds. Such information about the disease-related chemical compounds, that is, drug candidates, which the KG infers, would be useful for the drug repositioning that refers to the development of existing drugs for new medical indications.

Table 4 Information on the therapeutic uses and clinical trials for disease-related chemical compounds inferred from the Fibrinolysis network (2/2)

Chemical name	DrugBank ID	DrugBank: group	CTD ID	CTD: direct evidence
Vitamin K			D014812	T, M
Aminocaproic acid	DB00513	A, I	D015119	T, M
Vitamin K 2	DB12148	I	D024482	T, M
Vitamin K 3	DB00170	A, N	D024483	T, M
Vitamin K 1	DB01022	A, I	D010837	T, M
4-Aminomethylbenzoic acid			C015668	T
Tranexamic acid	DB00302	A	D014148	T, M
Anagrelide	DB00261	A		
Anagrelide hydrochloride				
Antifibrinolytic drug				
Fibrin fragment D				
Antifibrinolytic agents				
Alpha-2-antiplasmin			D000979	
Blood coagulation factors				
Fibrin				
Blood coagulation factor inhibitors				

In the “DrugBank: Group” column, “A” indicates the authorized drug, “I” indicates the investigational drug, and “N” indicates the nutraceutical drug for at least any one disease (except for inferred diseases). In the “CTD: Direct Evidence” column, “T” is an abbreviation for therapeutic, and it means that a chemical has a known or potential therapeutic role in at least any one disease. “M” is an abbreviation for marker/mechanism, and it means that a chemical correlates with at least any one disease or may play a role in the etiology of at least any one disease. Blanks indicate that the information has not been reported yet

Some diseases in IOBC contain external links: MeSH, International Statistical Classification of Diseases and Related Health Problems, 10th Revision [63], OMIM, National Drug File—Reference Terminology [64], National Cancer Institute Thesaurus (NCIt) [65], and DisGeNET. Using the Fibrinolysis network (Fig. 7), we found 325 UniProt proteins as thromboembolism-related gene products.

We performed the following federated SPARQL search via the DisGeNET SPARQL endpoint to integrate information on gene–disease associations in DisGeNET with the KG (Fig. 11).

Fig. 8 Associations between chemicals, namely chemical compounds, drugs and gene products, and diseases in the BMT network (1/3)

Fig. 9 Associations between chemicals, namely chemical compounds, drugs and gene products, and diseases in the BMT network (2/3)

```

prefix sio: <http://semanticscience.org/resource/>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix xkos: <http://rdf-vocabulary.ddialliance.org/xkos#>
prefix mesh: <http://id.nlm.nih.gov/mesh/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
select distinct ?prot
where {
  service <http://rdf.disgenet.org/sparql/> {
    SELECT distinct ?prot
    where {
      ?gda sio:SIO_000628 ?disease . # refers to
      ?disease skos:exactMatch mesh:D013923 . #Thromboembolism
      ?gda sio:SIO_000628 ?gene .
      ?gene rdf:type ncit:C16612 . #gene
      ?gene sio:SIO_010078 ?prot . # encodes
    }
  }
}
graph <http://nikkaji.biosciencedbc.jp/KG_FibrinolysisNetworkStep123extendedGO_20190209>
{
  {?event xkos:precedes ?disease .}
  UNION
  {?event sio:SIO_000062 ?disease .} # is participant in
  OPTIONAL {?event skos:exactMatch ?event2 .}
  {?uniprot sio:SIO_000062 ?event .}
  UNION
  {?uniprot sio:SIO_000225 ?event .} # has function
  UNION
  {?uniprot sio:SIO_000062 ?event2 .}
  UNION
  {?uniprot sio:SIO_000225 ?event2 .}
  ?uniprot skos:exactMatch ?prot .
}
}

```

Consequently, we discovered 13 disease-related proteins via both, the IOBC's KG and DisGeNET (e.g., uniprotkb:P00734) [66]. Moreover, we also found 18 disease-related proteins suggested only by DisGeNET (e.g., uniprotkb:P08519) and 312 disease-related proteins suggested only by the IOBC's KG (e.g., uniprotkb:Q9P126). This shows that the gene products suggested by IOBC's KG and DisGeNET may be stronger disease candidates than those suggested only by IOBC's or DisGeNET.

Conclusions

The Semantic Automated Discovery and Integration is a framework that assists in extracting chemical information using SPARQL [67]. Further, RDF and KG machine learning to find drug targets and predict side effects has been performed [68]. The results are actively being discussed; however, researchers with low specialized knowledge and skill sets may face challenges to prepare the execution environments of these drugs.

Fig. 10 Associations between chemicals, namely chemical compounds, drugs and gene products, and diseases in the BMT network (3/3)

Fig. 11 Interconnection between KG developed from IOBC and DisGeNET RDF using MeSH disease

We integrated biological knowledge: chemical compounds, gene products, biological processes, and diseases. We constructed KGs, from NikkajiRDF and IOBC, to facilitate the easy collection of biochemical and genomic information on the Internet, particularly information on chemical compounds' and gene products' functions and roles, as well as involvements in biological processes, including diseases. Valuable biochemical and genomic data sources dispersed globally should be findable,

accessible, interoperable, and reusable based on the FAIR principle [69]. The InChI/InChIKey as a chemical identifier based on the steric structure and other major identifiers in the biological database, thesauri, and ontologies such as UniProtKB accession number are necessary for integrating chemical compounds and gene products among different data sources. A federated search on SPARQL endpoints, such as the NBDC RDF portal, is also important. Conversely, the federated search from the public DBpedia SPARQL endpoint [22] to other SPARQL endpoints is currently unavailable.

We are evaluating the effectiveness of the knowledge expansion and inference using KGs, and ontologies in the field of bioresources. As a result, we confirm that they assist in finding new bioresource usages. For example, using KGs created from IOBC, NikkajiRDF, and other data sources, we can discover that coumarin (sti:200907007165179824), efficiently produced by a Tobacco cell, is not only a chemical compound related to oxidative stress, and plant defense responses [70], but also used in fluorescent dyes (chebi:CHEBI_51121), and as an anticoagulant (snomedct⁶:373307003).

In the future, the utilization of information on the interactions between chemical compounds, gene products, and metabolic and signal transduction pathways will facilitate more extensive and precise collection and prediction of chemical compounds' and gene products' associations with biological phenomena, along with the corresponding side effects. This will improve drug discovery, selection of effective medical treatments, and application of materials.

Acknowledgements This study was supported by an operating grant from the Japan Science and Technology Agency and JSPS KAKENHI Grant Number JP17H01789. A part of this study was progressed and discussed in Japan BioHackathon 2016 (BH16.12), which served as a research and development meeting. We are grateful to all participants who gave us their valuable advice and constructive comments.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Kimura, T., Kushida, T.: Openness of Nikkaji RDF data and integration of chemical information by Nikkaji acting as a hub. *J. Inf. Process. Manag.* **58**(3), 204–212 (2015)
2. NikkajiRDF Homepage in life science database archive. <http://doi.org/10.18908/lbdba.nbdc01530-02-000>. Accessed 25 Aug 2019
3. NikkajiRDF Homepage in NBDC RDF portal. <https://integbio.jp/rdf/?view=detail&id=nikkaji>. Accessed 25 Aug 2019
4. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., Pletnev, I.: InChI-the worldwide chemical structure identifier standard. *J. Cheminform.* **5**(1), 7 (2013)
5. Fu, G., Batchelor, C., Dumontier, M., Hastings, J., Willighagen, E., Bolton, E.: PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. *J. Cheminform.* **7**(1), 34 (2015)

⁶ <http://purl.bioontology.org/ontology/SNOMEDCT/>.

6. Willighagen, E.L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A.J., Tkachenko, V., Hastings, J., Chen, B., Wild, D.J.: The ChEMBL database as linked open data. *J. Cheminform* **5**(1), 23 (2013)
7. Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., Dumontier, M.: The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS* **6**(10), e25513 (2011)
8. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., Klassen, D., McCusker, J.P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M.D., Hoehndorf, R.: The SemanticScience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J. Biomed. Semantics* **5**(1), 14 (2014)
9. Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S., Overington, J.P.: UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.* **5**(1), 3 (2013)
10. NBDC RDF portal SPARQL endpoint. <https://integbio.jp/rdf/sparql>. Accessed 25 Aug 2019
11. Kushida, T., Kozaki, K., Tateisi, Y., Watanabe, K., Masuda, T., Matsumura, K., Kawamura, T., Takagi, T.: Efficient construction of a new ontology for life sciences by sub-classifying related terms in the Japan Science and Technology Agency thesaurus. In: Proceedings of the 8th international conference on biomedical ontology (ICBO 2017), 1–6, vol. 2137 of CEUR-WS.org, Newcastle (2017)
12. IOBC Homepage in BioPortal. <http://purl.bioontology.org/ontology/IOBC>. Accessed 25 Aug 2019
13. Noy, N.F., Shah, N. H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Ru-bin, D.L., Storey, M.A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* **37**(suppl_2), W170–W173 (2009)
14. IOBC SPARQL endpoint. <http://lod.hozo.jp/repositories/IOBC>. Accessed 25 Aug 2019
15. Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B.E., Martin, M.J., McGarvey, P., Gasteiger, E.: Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinform* **10**(1), 136 (2009)
16. Kushida, T., Kozaki, K., Kawamura, T., Tateisi, Y., Yamamoto, Y., Takagi, T.: Inference of functions, roles, and applications of chemicals using linked open data and ontologies. In: Semantic Technology: 8th Joint International Semantic Technology Conference (JIST 2018). LNCS 11341, pp. 385–397. Springer, Awaji (2018)
17. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C.: The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **41**(D1), D456–D463 (2013)
18. Wikipedia. <https://www.wikipedia.org/>. Accessed 25 Aug 2019
19. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia—a crystallization point for the Web of Data. *Web Semant. Sci. Serv. Agents World Wide Web* **7**(3), 154–165 (2009)
20. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. *Commun. ACM* **57**(10), 78–85 (2014)
21. Ertl, P., Patiny, L., Sander, T., Rufener, C., Zasso, M.: Wikipedia chemical structure explorer: sub-structure and similarity searching of molecules from Wikipedia. *J. Cheminform.* **7**(1), 10 (2015)
22. DBpedia public SPARQL endpoint. <https://dbpedia.org/sparql>. Accessed 25 Aug 2019
23. Wikidata public SPARQL endpoint. <https://query.wikidata.org/>. Accessed 25 Aug 2019
24. PubChem Classification Browser. <https://pubchem.ncbi.nlm.nih.gov/classification/>. Accessed 25 Aug 2019
25. PUG REST. <https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest>. Accessed 25 Aug 2019
26. UniProt SPARQL endpoint. <http://sparql.uniprot.org/sparql>. Accessed 25 Aug 2019
27. Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., Furlong, L.I.: DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, 1–17 (2015)
28. DisGeNET SPARQL endpoint. <http://rdf.disgenet.org/sparql/>. Accessed 25 Aug 2019
29. Batchelor, C., Breninkmeijer, C.Y.A., Chichester, C., Davies, M., Digles, D., Dunlop, I., Evelo, C.T., Gaulton, A., Goble, C., Gray, A., Groth, P., Harland, L., Karapetyan, K., Loizou, A., Overington, J., Pettifer, S., Steele, J., Stevens, R., Tkachenko, V., Waagmeester, A., Williams, A.J., Willighagen, E.: Scientific lenses to support multiple views over linked chemistry data. In: The Semantic Web: 13th International Semantic Web Conference (ISWC 2014). Proceedings, Part I, pp. 98–113. Springer, Riva del Garda (2014)
30. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* **41**(5), 706–716 (2008)

31. Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinform.* **11**, 255 (2010)
32. Kobayashi, N., Lenz, K., Masuya, H.: RIKEN MetaDatabase: a database platform as a microcosm of linked open data cloud in the life sciences. In: *Semantic Technology: 6th Joint International Conference (JIST 2016)*. LNCS 10055, pp. 99–115. Springer, Gold Coast (2016)
33. Schriml, L.M., Mitaka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C.P., Kibbey, S., Sreekumar, P., Le, C., Giglio, M., Greene, C.: Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**(D1), D955–D962 (2019)
34. Slenter, D.N., Kutmon, N., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L.M.T., Evelo, C.T., Pico, A.R., Willighagen, E.L.: WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**(D1), D661–D667 (2018)
35. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R., Murphy, T.D.: Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**(Database issue), D36–42 (2015)
36. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**(Database issue), D514–517 (2005)
37. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000)
38. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., Wilson, M.: DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**(D1), D1074–D1082 (2018)
39. Bio2RDF SPARQL endpoint. <http://bio2rdf.org/sparql>. Accessed 17 May 2019
40. Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C.S., Willighagen, E., Hajagos, J., Marshall, M.S., Prud'hommeaux, E., Hassenzadeh, O., Pichler, E., Stephens, S.: Linked open drug data for pharmaceutical research and development. *J. Cheminform.* **3**(1), 19 (2011)
41. Chen, B., Ding, Y., Wild, D.J.: Assessing drug target association using semantic linked data. *PLoS Comput. Biol.* **8**(7), e1002574 (2012)
42. The FANTOM Consortium and the RIKEN PMI and CLST (DGT): a promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014)
43. Yoshiki, A., Ike, F., Mekada, K., Kitaura, Y., Nakata, H., Hiraiwa, N., Mochida, K., Ijuin, M., Kadota, M., Murakami, A., Ogura, A., Abe, K., Moriwaki, K., Obata, Y.: The mouse resources at the RIKEN BioResource center. *Exp. Anim.* **58**(2), 85–96 (2009)
44. Kinjo, A.R., Bekker, G.J., Suzuki, H., Tsuchiya, Y., Kawabata, T., Ikegawa, Y., Nakamura, H.: Protein Data Bank Japan (PDBj): Updated user interfaces, Resource Description Framework, analysis tools for large structures. *Nucleic Acids Res.* **45**(D1), D282–D288 (2017)
45. Gkoutos, G.V., Mungall, C., Dolken, S., Ashburner, M., Lewis, S., Hancock, J., Schofield, P., Kohler, S., Robinson, P.N.: Entity/quality-based logical definitions for the human skeletal phenome using PATO. In: *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 7069–7072 (2009)
46. ChEBI ontology files. <ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/>. Accessed 25 Aug 2019
47. link2OtherDBs_basedOnUniChem of NikkajiRDF. <http://doi.org/10.18908/lbdba.nbdc01530-02-006>. Accessed 25 Aug 2019
48. SPARQL query result in Section 3.1. <http://nikkaji-rdf.biosciencedbc.jp/download/quary24/chebi2nikkajiRDF/0,5000.html>. Accessed 25 Aug 2019
49. Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating mappings for ontologies in biomedicine: simple methods work. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, pp. 198–202 (2009)
50. Euzenat, J., Shvaiko, P.: *Ontology Matching*, 2nd edn. Springer, Heidelberg, New York, Dordrecht, London (2013)
51. Harrow, I., Balakrishnan, R., Jimenez-Ruiz, E., Jupp, S., Lomax, J., Reed, J., Romacker, M., Senger, C., Splendiani, A., Wilson, J., Woollard, P.: Ontology mapping for semantically enabled applications. *Drug Discov. Today* (18), S1359–6446 (2019)

52. Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F.M., Cruz, I.F.: Tackling the challenges of matching biomedical ontologies. *J. Biomed. Semantics* **9**(1), 4 (2018)
53. SPARQL query result in Section 3.2. http://nikkaji-rdf.biosciencedbc.jp/download/quary25/reasoning_inheritance/.html. Accessed 25 Aug 2019
54. Kushida, T., Masuda, T., Tateisi, Y., Watanabe, K., Matsumura, K., Kawamura, T., Kozaki, K., Takagi, T.: Refining JST thesaurus and discussing the effectiveness in life science research. In: Proc. of 5th Intelligent Exploration of Semantic Data Workshop (IESD 2016, co-located with ISWC 2016), pp. 1–14, Kobe (2016)
55. Kushida, T., Tateisi, Y., Masuda, T., Watanabe, K., Matsumura, K., Kawamura, T., Kozaki, K., Takagi, T.: Refined JST Thesaurus Extended with Data from Other Open Life Science Data Sources. In: *Semantic Technology: 7th Joint International Conference (JIST 2017)*. LNCS 10675, pp. 35–48. Springer, Gold Coast (2017)
56. SPARQL query result in Section 4.1, Fibrinolysis network. <http://nikkaji-rdf.biosciencedbc.jp/download/quary27/FibrinolysisNetwork20190208/.csv>. Accessed 25 Aug 2019
57. SPARQL query result in Section 4.1, BMT network. <http://nikkaji-rdf.biosciencedbc.jp/download/quary28/BMTNetwork20190208/.csv>. Accessed 25 Aug 2019
58. Bodenreider, O., Nelson, S.J., Hole, W.T., Chang, H.F.: Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proc. AMIA Sympos.* 815–819 (1998)
59. Gene Ontology Consortium: Creating the gene ontology resource: design and implementation. *Genome Res.* **11**(8), 1425–1433 (2001)
60. Wimalaratne, S.M., Bolleman, J., Juty, N., Katayama, T., Dumontier, M., Redaschi, N., Le Novère, N., Hermjakob, H., Laibe, C.: SPARQL-enabled identifier conversion with Identifiers.org. *Bioinformatics* **31**(11), 1875–1877 (2015)
61. Amigo 2. <http://amigo.geneontology.org/amigo>. Accessed 25 Aug 2019
62. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., McMorran, R., Wiegiers, J., Wiegiers, T.C., Mattingly, C.J.: The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.* **47**(D1), D948–D954 (2019)
63. ICD10. <https://www.who.int/classifications/icd/icdonlineversions/en/>. Accessed 25 Aug 2019
64. NDFRT. <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/>. Accessed 25 Aug 2019
65. NCIt. <https://ncit.nci.nih.gov/>. Accessed 25 Aug 2019
66. SPARQL query result in Section 4.3. <http://nikkaji-rdf.biosciencedbc.jp/download/quary29/ThromboembolismRelatedGeneProducts/.html>. Accessed 25 Aug 2019
67. Chepelev, L.L., Dumontier, M.: Semantic Web integration of Cheminformatics resources with the SADI framework. *J. Cheminform.* **3**(1), 16 (2011)
68. Alshahrani, M., Khan, M.A., Maddouri, O., Kinjo, A.R., Queralt-Rosinach, N., Hoehndorf, R.: Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* **33**(17), 2723–2730 (2017)
69. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Bouwman, J.: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. data* **3**, 160018 (2016)
70. Chong, J., Baltz, R., Schmitt, C., Beffa, R., Fritig, B., Saindrenan, P.: Downregulation of a pathogen-responsive tobacco UDP-Glc:phenylpropanoid glucosyltransferase reduces scopolin glucoside accumulation, enhances oxidative stress, and weakens. *Plant Cell* **14**(5), 1093–1107 (2002)

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

Affiliations

Tatsuya Kushida¹ · **Kouji Kozaki**² · **Takahiro Kawamura**³ · **Yuka Tateisi**¹ · **Yasunori Yamamoto**⁴ · **Toshihisa Takagi**^{1,5}

Kouji Kozaki

kozaki@osakac.ac.jp

Takahiro Kawamura

takahiro.kawamura@affrc.go.jp

Yuka Tateisi

tateisi@biosciencedbc.jp

Yasunori Yamamoto

yy@dbcls.rois.ac.jp

Toshihisa Takagi

tt@tuins.ac.jp

¹ National Bioscience Database Center, Japan Science and Technology Agency, 5-3, Yonbancho, Chiyoda-ku, Tokyo 102-8666, Japan

² Faculty of Information and Communication Engineering, Osaka Electro-Communication University, Osaka, Japan

³ National Agriculture and Food Research Organization, Ibaraki, Japan

⁴ Database Center for Life Science, Research Organization of Information and Systems, Chiba, Japan

⁵ Toyama University of International Studies, Toyama, Japan