Title: CBKH: The Cornell Biomedical Knowledge Hub

Chang Su[#,1], Yu Hou[#,1], Winston Guo[2], Fayzan Chaudhry[3], Gregory Ghahramani[3], Haotan Zhang[3], Fei Wang*[,1]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY
[2]Department of Medicine, Weill Cornell Medicine, New York, NY
[3]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY
[#]Equal Contribution
*Corresponding author

Corresponding Author:

Fei Wang,
425 E 61 St. New York City. NY 10065. USA
few2001@med.cornell.edu

Abstract

The rapidly increasing biomedical knowledge, derived from biological experiments or gained from clinical practice, has become the important treasure in the biomedical research. The emerging knowledge graphs (KGs) provide an efficient and effective way to organize and retrieval the huge and increasing volume of biomedical knowledge. A biomedical KG (BKG) typically stores and represents knowledge by constructing a semantic network describing entities and the relationships between them. Previous efforts have been conducted to construct and curate BKGs by comprehensively integrating various biomedical data resources. Though the resulting BKGs have made a significant progress in this filed in advancing biological and medical research, there remain a big gap to a perfect one that is comprehensive and fine-grained enough. To this end, in the present study, we collected and integrated data from diverse well-curated biomedical knowledge bases and BKGs to curate a more comprehensive one, named the Cornell Biomedical Knowledge Hub (CBKH). To enhance the usage in accelerating biomedical research, we deployed CBKH using the famous graph database, Neo4j. This is a continuing effort and we are adding in more and more contents in CBKH to support the various complex needs in biomedical data analysis. Please contact us if you have better ideas and suggestions.

Introduction

Biomedicine is a discipline with lots of highly specialized knowledge accumulated from biological experiments and clinical practice. These knowledges are usually buried in massive biomedical literature and textbooks. This makes the effective knowledge organization and efficient knowledge retrieval a challenging task. Knowledge graph is a recently emerged concept aiming at achieving this goal. A knowledge graph (KG) stores and represents knowledge by constructing a semantic network describing entities and the relationships between them. The basic elements that comprising a knowledge graph are a set of biomedical entities and a set of different types of semantic relationships among the entities. In biomedicine, the typical entities could be diseases, drugs, and genes, etc., and the relationships could be treats (drug-treats-disease), binds (drug-binds-target protein), interactions (drug-drug interaction), etc. Large scale biomedical KG (BKG) makes efficient knowledge retrieval and inference possible.

Typically, construction and curation of a BKG is done via integrating publicly available biomedical knowledge bases and knowledge extracted from biomedical literature. For example, Hetionet [1], released in 2017, is a well-curated BKG that was constructed by integrating 29 publicly available data resources, such as DrugBank [2], GWAS Catalog [3], DISEASES [4], DisGeNET [5], etc. Similar to Hetionet, Drug Repurposing Knowledge Graph (DRKG) [6] was built by integrating data from six different existing databases, with a specific focus on drug repurposing for COVID-19. It contains 13 types of about 100K entities and 107 types of over 5 million relationships. PreMedKB [7] includes the information of disease, genes, variants, and drugs by integrating existing resources. The Clinical Knowledge Graph (CKG) [8] was constructed by combining relevant existing biomedical databases integration and texts extracted from scientific literature, containing over 16 million nodes and over 220 million relationships. Compared to other BKGs, CKG includes entities representing biological information at a finer granularity, such as metabolite, modified protein, molecule function, transcript, genetic variant, food, clinical variable, etc. In addition, some BKGs were built with a focus on specific diseases or conditions. For example, COVID-KG [9] extracted COVID-19 specific information from biomedical literature and constructed a knowledge graph containing diseases, chemicals, and genes, along with their relationships. KGHC [10] is a knowledge graph focused on hepatocellular carcinoma. It extracted information from literature and contents on the internet, as well as structured triples from SemMedDB [11].

Though significant progress has been achieved by these efforts, they are not perfect or comprehensive enough to incorporate all biomedical knowledge. For example, Parkinson's disease (PD) is associated genetic mutation like G2019S in LRRK2, but only gene-level information is saved in most existing BKGs, such as Hetionet. PD is associated with brain lesion detected by MRI, but such information is not incorporated in current BKGs. In addition, entities at finer granularity, such as molecules, which have been demonstrated to be important in biomedical research, are not included in most existing BKGs like Hetionet. Therefore, there still is the need for curation of a comprehensive BKG. To this end, in the present study, we collected and integrated data from multiple well-curated biomedical knowledge bases and BKGs to curate

a more comprehensive one, named the Cornell Biomedical Knowledge Hub (CBKH). We deployed our CBKH using Neo4j (https://neo4j.com). If you are interested in accessing CBKH, please contact us.

Materials and Methods

Our ultimate goal was to build a biomedical knowledge graph via comprehensively incorporating biomedical knowledge as much as possible. To this end, we collected and integrated 15 publicly available data sources to curate a comprehensive one. Details of the used data resources were listed in Table 1.

Raw data processing and information extraction

Given the data resources, the first step was to pre-process the raw files of them and extract knowledge, including entity information and relationship information, from them. Generally, the data bases release their raw data files in various format, such as comma-separated values (CSV), tab-separated values (TSV), TXT, EXCEL tablet, Hypertext Markup Language (HTML), Resource Description Framework (RDF), and Web Ontology Language (OWL). To this end, for each data base, we parsed the raw files and extracted structured data, i.e., the descriptive files for each type of biomedical entity and the files of each type of relationship. Such procedure varies by data bases or even by files within the same data base.

Term normalization

For normalization of the entity terms, we utilized a greedy strategy. Specifically, we first chose a data base to initialize the vocabulary for each type of entity. Next, we used multiple identifiers as the linkage pool for entity normalization and incorporate and integrate entities from all data bases to enrich the entity vocabulary one by one.

For **gene** entities, we used HGNC gene repository [12] as the initialization vocabulary of gene entities, as it sets a standard nomenclature for human the genes. The linkage pool for normalization included HGNC ID, HGNC symbol and NCBI ID.

For **drug** entities, we initialized our vocabulary using DrugBank [2] as it provides the up-to-date list of approved drugs and investigational drugs under clinical trials. The linkage pool for drug entity normalization included MeSH term, MeSH term ID, Unified Medical Language System (UMLS) Concept Unique Identifier (CUI), and the drug name in UMLS.

For **disease** entities, we used the Disease Ontology [13] for initializing the vocabulary, as it is a structured database of diseases based on etiological classification. The linkage pool for the disease entities normalization included MeSH term, MeSH term ID, UMLS CUI and the disease name in UMLS.

For **anatomy** entities, we used the Uberon [14] for initializing the vocabulary, as it is a cross-species anatomical ontology based on traditional anatomical classification. The linkage pool for the anatomy entities normalization included MeSH term, MeSH term ID, UMLS CUI and the anatomy name in UMLS.

For **molecule** entities, we used the ChEMBL [15] for initializing the vocabulary, as it is a manually curated database of molecules with drug properties. The linkage pool for the molecule entities normalization included International Chemical Identifier (InChi).

For **symptom** entities, we collected the symptom entities from the Hetionet and described them by using the MeSH term and MeSH term ID. We used UMLS CUI as the linkage for symptom entities normalization.

CBKH deployment

To enhance the usability of CBKH in accelerating biomedical research, we deployed it using a graph database, Neo4j (https://neo4j.com), which provides the easy-to-use interface for query and visiting knowledge in the KG. By using the Cypher statement on the Neo4j platform, CBKH can be retrieved efficiently and flexibly.

Results

CBKH integrates data from 15 publicly available biomedical databases. The current version of CBKH (Figure 1 and Table 2) contains a total of 2,231,297 entities of 6 types. Specifically, the CBKH includes 22,963 anatomy entities, 18,503 disease entities, 36,436 drug entities, 87,942 gene entities, 2,065,015 molecule entities and 438 symptom entities. For the relationships in the CBKH (Table 3), there are 91 relation types within 8 kinds of entity pairs, including Anatomy-Gene, Drug-Disease, Drug-Drug, Drug-Gene, Disease-Disease, Disease-Gene, Disease-Symptom and Gene-Gene. In total, CBKH contains 48,678,651 relations. More specifically, there are 3 types of relations between the Anatomy-Gene pair, including such as ‘Express’ and ‘Absent’; 11 relation types between Drug-Disease pair, such as ‘Treat’ and ‘Effect’; 2 relation types between the Drug-Drug pair including ‘Interaction’ and ‘Resemble’; 25 relation types between the Drug-Gene pair, such as ‘Target’, ‘Upregulates’, and ‘Downregulates’; 2 relation types between the Disease-Disease pair including ‘is_a’ and ‘Resemble’; 16 relation types between the Disease-Gene pair, such as ‘Association’; the ‘Presents’ relation type between the Disease-Symptom pair; and 31 relation types between the Gene-Gene pair, such as ‘Covaries’ and ‘Interacts’. Since some resources are generated by text mining methods, they use the form of phrases to express the relations (Text-semantic relation). For example, the relation 'role in disease pathogenesis' between the Drug-Disease pair and the relation 'enhances expression/production' between the Drug-Gene pair. The CBKH relations were derived by integrating candidate resources, so some relationships connecting the two entities may have overlap. For example, there are 16,961 'Target_DrugBank' relationships and 11,801 'Binds_Hetionet' relationships in Drug-Gene. In these two relationships, a total of 4,745

relationships overlaps, which means that both 'Target' and 'Binds' relationships exist in these corresponding entities.

Future work

KG quality control

The procedures of constructing and curating a BKG include sophisticated efforts on raw data file extraction and pre-processing, data annotation, as well as terminology normalization, which may result in quality issues. In general, there are two categories of quality issues in KGs: the incorrectness and incompleteness.

Incorrectness refers to incorrect facts in the KG, e.g., a relation connecting two entities exists in the BKG but inconsistent with real-world evidence. To address this, a common strategy is manual annotation with sampled small subsets. Such procedure is time- and cost-consuming, if one wants to evaluate sufficient triplets to reach the statistic criteria. To address this, for example, Gao et al. [16] proposed an iterative evaluation framework for KG accuracy evaluation. Specifically, inspired by the properties of the annotation cost function observed in practice, the authors developed a cluster sampling strategy with unequal probability theory. Their framework resulted in a 60% shrunk annotation cost and can be easily extended to address evolving KG. In addition, the use of well-designed biomedical vocabularies such as the Unified Medical Language System (UMLS) will improve entity term normalization and hence reduce the risk of errors caused by the ambiguous biomedical entities. Moreover, learning based on KG structure to refine the KG is also a potential way to solve this issue. Early efforts, such as Zhao et al. [17] has been focused on this field.

Incompleteness mainly refers to the missing of biologically or clinically meaningful triplets in the KG. To address the incompleteness in biomedical KG, we integrated multiple data resources, biomedical data bases, and biomedical KGs to construct and curate a more comprehensive one. However, there is no guarantee the included resources are combined comprehensive enough to cover all biomedical knowledge. In addition, today's largely available biomedical literature and medical data (e.g., EHRs) are great treasure of biomedical knowledge. In this context, previous studies have been focused on deriving knowledge from biomedical literature [18-21] and EHR data [22, 23], and the derived knowledge could be a good complement for the biomedical KGs. Moreover, the computational methods such as the KG embedding models (e.g., TransE and TransH) and the GNNs (e.g., R-GCN) have been used in KG completion [24], which predict missing relations within a KG according to its structure properties.

Focus on specific diseases on health conditions

Similar to most existing BKGs, like Hetionet and CKG, our CBKH focus on general biomedical knowledge. However, for the sake of precision medicine on some specific human diseases or

health conditions, there is the need for very fine-grained knowledge with a specific focus on them. In this context, COVID-KG [9] included biomedical knowledge with a specific focus on COVID-19; KGHC [10] is a knowledge graph constructed focusing on addressing hepatocellular carcinoma. Following this idea, to adapt our KG to addressing problems in specific complex diseases and health conditions like Alzheimer's disease, Parkinson's disease, and mental illness, we will focus on collecting fine-grained data, such as genotype-phenotype associations and brain region atrophy-phenotype associations and incorporating them to enrich our BKG, for the specific usage of these diseases.

Keeping KG up to date

Thanks to the advances in the high throughput techniques, biomedical data have been continuously produced. Meanwhile, a rapid increasing amount of biomedical literature are being published. As most existing studies gather knowledge from the experimental data and biomedical literature manually, more human involvement is required. In this context, we would highlight the usage of the computational methods, such as Natural Language Processing (NLP) techniques, which can automatically and efficiently extract knowledge from the raw data files, such as biomedical literature and clinical trial documentations. In the future, we may incorporate such kind of technique to assist us in KG maintenance.

Acknowledgement

The work is supported by NSF 1750326 and 2027970.

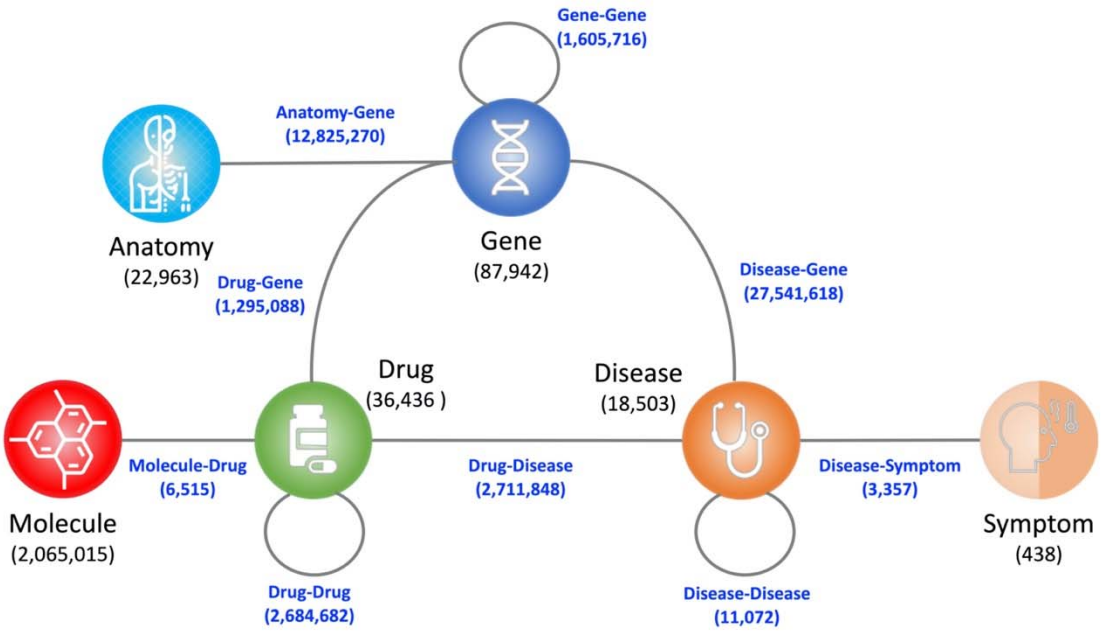

Figure 1. Schema of CBKH

Table 1. Data resources used for integration

Source	Entity		Relation		URL	License
	Types	Number	Types	Number		
Bgee[25]	Anatomy, Gene	60072	Anatomy-Express Present-Gene, Anatomy-Express Absent-Gene	11731369	https://bgee.org/	https://creativecommons.org/publicdomain/zero/1.0/
Brenda Tissue Ontology[26]	Tissue	6478	-	-	https://www.brenda-enzymes.org/index.php	https://creativecommons.org/licenses/by/4.0/
Cell Ontology[27]	Cell	2,200	-	-	http://obofoundry.org/ontology/cl.html	https://creativecommons.org/licenses/by/4.0/
Comparative Toxicogenomics Database[28]	Disease, Gene, Chemical, Pathway	73922	Chemical-Gene, Chemical-Disease, Chemical-Pathway, Gene-Disease, Gene-Pathway, Disease-Pathway	38344568	http://ctdbase.org/	https://creativecommons.org/licenses/by/4.0/
ChEMBL[15]	Molecular	1940733	-	-	https://www.ebi.ac.uk/chembl/	https://creativecommons.org/licenses/by-sa/3.0/
ChEBI[29]	Molecular	155342	-	-	https://www.ebi.ac.uk/chebi/init.do	https://creativecommons.org/licenses/by/4.0/
Drug Repurposing Knowledge Graph[6]	Anatomy, Atc, Biological process, Cellular component, Compound, Disease, Gene, Molecular function, Pathway, Pharmacologic class, Side effect, Symptom, Tax	97238	Gene-Gene, Compound-Gene, Disease-Gene, Atc-Compound, Compound-Compound, Compound-Disease, Gene-Tax, Biological process-Gene, Disease-Symptom, Anatomy-Disease, Disease-Disease, Anatomy-Gene, Gene-Molecular function, Compound-Pharmacologic class, Cellular component-Gene, Gene-Pathway, Compound-Side effect	5874261	https://github.com/gnn4dr/DRKG	https://www.apache.org/licenses/LICENSE-2.0
Disease Ontology[13]	Disease	10648	-	-	https://disease-ontology.org/	https://creativecommons.org/publicdomain/zero/1.0/
DrugBank[2]	Drug	15128	Drug-Target, Drug-Enzyme, Drug-Carrier, Drug-Transporter	28014	https://go.drugbank.com/	http://creativecommons.org/licenses/by-nc/4.0/

Hetionet[1]	Anatomy, Biological process, Cellular component, Compound, Diease, Gene, Molecular function, Pathway, Pharmacologic class, Side effect, Symptom	47031	Anatomy–downregulates–Gene, Anatomy–expresses–Gene, Anatomy–upregulates–Gene, Compound–binds–Gene, Compound–causes–Side Effect, Compound–downregulates–Gene, Compound–palliates–Disease, Compound–resembles–Compound, Compound–treats–Disease, Compound–upregulates–Gene, Disease–associates–Gene, Disease–downregulates–Gene, Disease–localizes–Anatomy, Disease–presents–Symptom, Disease–resembles–Disease, Disease–upregulates–Gene, Gene–covaries–Gene, Gene–interacts–Gene, Gene–participates–Biological Process, Gene–participates–Cellular Component, Gene–participates–Molecular Function, Gene–participates–Pathway, Gene→regulates→Gene, Pharmacologic Class–includes–Compound		https://github.com/hetio/hetionet	https://creativecommons.org/publicdomain/zero/1.0/
HUGO Gene Nomenclature Committee[12]	Gene	41439	-	-	https://www.genenames.org/	http://creativecommons.org/licenses/by/4.0/
KEGG[30]	Drug, Disease, Gene, Variant, Compound, Pathway	33756186	Drug-Gene, Disease-Gene, Gene-Pathway	43464	https://www.kegg.jp/	http://creativecommons.org/licenses/by-nc/2.0/uk/
PharmGKB[31]	Genes, Variant, Drug/Chemical, Phenotype	43112	Disease-Gene, Drug/Chemical - Gene, Gene-Gene, Gene-Variant, Disease-Variant, Drug/Chemical-Variant	61616	https://www.pharmgkb.org/	https://creativecommons.org/licenses/by-sa/4.0/
TISSUE[32]	Tissue, Gene	26260	Tissue-Express-Gene	6788697	https://tissues.jensenlab.org/	https://creativecommons.org/licenses/by/4.0/
Uberon[14]	Anatomy	14944	-	-	https://www.ebi.ac.uk/ols/ontologies/uberon	http://creativecommons.org/licenses/by/3.0/

Table 2. Statistics of biomedical entities in CBKH

Entity Type	Number	Included Identifiers[1]
Anatomy	22,963	Uberon ID, BTO ID, MeSH ID, Cell Ontology ID
Disease	18,503	Disease Ontology ID, KEGG ID, PharmGKB ID, MeSH ID, OMIM ID
Drug	36,436	DrugBank ID, KEGG ID, PharmGKB ID, MeSH ID
Gene	87,942	HGNC ID, NCBI ID, PharmGKB ID
Molecule	2,065,015	CHEMBL ID, CHEBI ID
Symptom	438	MeSH ID

[1] The identifiers used for entity term normalization.

Table 3. Statistics of relationships among entities in CBKH

Entity pair	Total number of relations between the entity pairs	Relation types	Number of relations of the specific type
Anatomy-Gene	12,825,270	Anatomy-Express_Bgee-Gene	4,931,924
		Anatomy-Absent_Bgee-Gene	3,338,908
		Anatomy-Express_TISSUE-Gene	6,783,256
Drug-Disease	2,711,848	Drug-Palliates_Hetionet-Disease	390
		Drug-Treats_Hetionet-Disease	755
		Drug-Effect_KEGG-Disease	1,527
		Drug-Association_CTD-Disease	2,682,510
		Drug-Treat_DrugBank-Disease	4,717
		Drug-Text_Semantic-Disease[1]	50,419
Drug-Drug	2,684,682	Drug-Interaction_DrugBank-Drug	2,682,157
		Drug-Resemble_Hetionet-Drug	6,486
Drug-Gene	1,295,088	Drug- Target_DrugBank-Gene	16,961
		Drug-Transporter_DrugBank-Gene	3,066
		Drug-Enzyme_DrugBank-Gene	5,243
		Drug-Carrier_DrugBank-Gene	853
		Drug-Downregulates_DrugBank-Gene	46,467
		Drug-Upregulates_DrugBank-Gene	54,204

		Drug-Associated_KEGG-Gene	9,105
		Drug-Associated_PharmGKB-Gene	5,718
		Drug-Binds_Hetionet-Gene	11,801
		Drug-Downregulates_Hetionet-Gene	21,102
		Drug-Upregulates_Hetionet-Gene	18,756
		Drug-Interaction_CTD-Gene	1,181,456
		Drug-Text_Semantic-Gene[2]	65,382
Disease-Disease	11,072	Disease-is_a_DO-Disease	10,529
		Disease-Resemble_Hetionet-Disease	543
Disease-Gene	27,541,618	Disease-Associate_Hetionet-Gene	12,623
		Disease-Downregulates_Hetionet-Gene	7,623
		Disease-Upregulates_Hetionet-Gene	7,731
		Disease-Associate_KEGG-Gene	5,052
		Disease-Associate_PharmGKB-Gene	3,534
		Disease-Association_CTD-Gene	27,487,252
		Disease-Text_Semantic-Gene[3]	95,587
Disease-Symptom	3,357	Disease- Present_Hetionet-Symptom	3,357
Gene-Gene	1,605,716	Gene-Covaries_Hetionet-Gene	61,690
		Gene-Interacts_Hetionet-Gene	147,164
		Gene-Regulates_Hetionet-Gene	265,672
		Gene-Associate_PharmGKB-Gene	2,836
		Gene-Text_Semantic-Gene[4]	1,816,789

[1] Drug-Text_Semantic-Disease relation type includes: 'Compound treats the disease_DRUGBANK', 'treatment/therapy (including investigatory)_GNBR', 'inhibits cell growth (esp. cancers)_GNBR', 'alleviates, reduces_GNBR', 'biomarkers (of disease progression)_GNBR', 'prevents, suppresses_GNBR', 'role in disease pathogenesis_GNBR'.

[2] Drug-Text_Semantic-Gene relation type includes: 'affects expression/production (neutral)_GNBR', 'agonism, activation_GNBR', 'inhibits_GNBR', 'metabolism, pharmacokinetics_GNBR', 'antagonism, blocking_GNBR', 'increases expression/production_GNBR', 'binding, ligand (esp. receptors)_GNBR', 'decreases expression/production_GNBR', 'transport, channels_GNBR', 'enzyme activity_GNBR', 'direct interation_IntAct', 'physical association_IntAct', 'association_IntAct'.

[3] Disease-Text_Semantic-Gene relation type includes: 'improper regulation linked to disease_GNBR', 'causal mutations_GNBR', 'polymorphisms alter risk_GNBR', 'role in pathogenesis_GNBR', 'possible therapeutic effect_GNBR', 'biomarkers (diagnostic)_GNBR', 'promotes progression_GNBR', 'drug targets_GNBR', 'overexpression in disease_GNBR', 'mutations affecting disease course_GNBR'.

[4] Gene-Text_Semantic-Gene relation type includes: 'activates, stimulates_GNBR', 'production by cell population_GNBR', 'regulation_GNBR', 'binding, ligand (esp. receptors)_GNBR', 'signaling pathway_GNBR', 'increases expression/production_GNBR', 'same protein or complex_GNBR', 'enhances response_GNBR', 'affects expression/production (neutral)_GNBR', 'association_IntAct', 'physical association_IntAct', 'colocalization_IntAct', 'dephosphorylation reaction_IntAct', 'cleavage reaction_IntAct', 'direct interation_IntAct', 'phosphorylation reaction_IntAct', 'ADP ribosylation reaction_IntAct', 'ubiquitination reaction_IntAct', 'protein cleavage_IntAct', 'reaction_STRING', 'catalysis_STRING', 'activation_STRING', 'inhibition_STRING', 'other_STRING', 'binding_STRING', 'post-translational modification_STRING', 'expression_STRING'

Reference:

1. Himmelstein, D.S., et al., *Systematic integration of biomedical knowledge prioritizes drugs for repurposing.* Elife, 2017. **6**: p. e26726.
2. Wishart, D.S., et al., *DrugBank 5.0: a major update to the DrugBank database for 2018.* Nucleic acids research, 2018. **46**(D1): p. D1074-D1082.
3. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic acids research, 2019. **47**(D1): p. D1005-D1012.
4. Pletscher-Frankild, S., et al., *DISEASES: Text mining and data integration of disease–gene associations.* Methods, 2015. **74**: p. 83-89.
5. Piñero, J., et al., *The DisGeNET knowledge platform for disease genomics: 2019 update.* Nucleic acids research, 2020. **48**(D1): p. D845-D855.
6. Ioannidis, V.N., et al., *Drkg-drug repurposing knowledge graph for covid-19*. 2020, arxiv.
7. Yu, Y., et al., *PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs.* Nucleic acids research, 2019. **47**(D1): p. D1090-D1101.
8. Santos, A., et al., *Clinical knowledge graph integrates proteomics data into clinical decision-making.* bioRxiv, 2020.
9. Wang, Q., et al., *COVID-19 literature knowledge graph construction and drug repurposing report generation.* arXiv preprint arXiv:2007.00576, 2020.
10. Li, N., et al., *KGHC: a knowledge graph for hepatocellular carcinoma.* BMC Medical Informatics and Decision Making, 2020. **20**(3): p. 1-11.
11. Kilicoglu, H., et al., *SemMedDB: a PubMed-scale repository of biomedical semantic predications.* Bioinformatics, 2012. **28**(23): p. 3158-3160.
12. Braschi, B., et al., *Genenames. org: the HGNC and VGNC resources in 2019.* Nucleic acids research, 2019. **47**(D1): p. D786-D792.
13. Schriml, L.M., et al., *Human Disease Ontology 2018 update: classification, content and workflow expansion.* Nucleic acids research, 2019. **47**(D1): p. D955-D962.
14. Mungall, C.J., et al., *Uberon, an integrative multi-species anatomy ontology.* Genome biology, 2012. **13**(1): p. 1-20.
15. Mendez, D., et al., *ChEMBL: towards direct deposition of bioassay data.* Nucleic acids research, 2019. **47**(D1): p. D930-D940.
16. Gao, J., et al., *Efficient knowledge graph accuracy evaluation.* arXiv preprint arXiv:1907.09657, 2019.
17. Zhao, S., et al., *Biomedical Knowledge Graph Refinement with Embedding and Logic Rules.* arXiv preprint arXiv:2012.01031, 2020.
18. Zhao, S., et al., *Recent advances in biomedical literature mining.* Briefings in Bioinformatics, 2020.
19. Xu, R., L. Li, and Q. Wang, *Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature.* Bioinformatics, 2013. **29**(17): p. 2186-2194.
20. Zhang, Y., et al., *Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths.* Bioinformatics, 2018. **34**(5): p. 828-835.
21. Sahu, S.K. and A. Anand, *Drug-drug interaction extraction from biomedical texts using long short-term memory network.* Journal of biomedical informatics, 2018. **86**: p. 15-24.
22. Rotmensch, M., et al., *Learning a health knowledge graph from electronic medical records.* Scientific reports, 2017. **7**(1): p. 1-11.

23. Chen, I.Y., et al. *Robustly extracting medical knowledge from ehrs: A case study of learning a health knowledge graph*. in *Pac Symp Biocomput*. 2020. World Scientific.
24. Arora, S., *A Survey on Graph Neural Networks for Knowledge Graph Completion.* arXiv preprint arXiv:2007.12374, 2020.
25. Bastian, F.B., et al., *The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals.* Nucleic Acids Research, 2021. **49**(D1): p. D831-D847.
26. Gremse, M., et al., *The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources.* Nucleic acids research, 2010. **39**(suppl_1): p. D507-D513.
27. Diehl, A.D., et al., *The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability.* Journal of biomedical semantics, 2016. **7**(1): p. 1-10.
28. Davis, A.P., et al., *The comparative toxicogenomics database: update 2019.* Nucleic acids research, 2019. **47**(D1): p. D948-D954.
29. Hastings, J., et al., *ChEBI in 2016: Improved services and an expanding collection of metabolites.* Nucleic acids research, 2016. **44**(D1): p. D1214-D1219.
30. Ogata, H., et al., *KEGG: Kyoto encyclopedia of genes and genomes.* Nucleic acids research, 1999. **27**(1): p. 29-34.
31. Whirl-Carrillo, M., et al., *Pharmacogenomics knowledge for personalized medicine.* Clinical Pharmacology & Therapeutics, 2012. **92**(4): p. 414-417.
32. Palasca, O., et al., *TISSUES 2.0: an integrative web resource on mammalian tissue expression.* Database, 2018. **2018**.