

Constructing Knowledge Graphs and Their Biomedical Applications

This manuscript ([permalink](#)) was automatically generated from [greenelab/knowledge-graph-review@d1d51c8](#) on December 26, 2019.

Authors

- **David Nicholson**

 [0000-0003-0002-5761](#) ·  [danich1](#)

Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552 and T32 HG000046

- **Jane Roe**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [janeroe](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

1. Give high level description of review as it pertains to knowledge graphs (creation and application)

Introduction

Knowledge graphs are a practical resource for many real world applications. They have been used in social medial mining to classify nodes [1] or to create a recommendation system [2]. Knowledge graphs have also been used to understand natural language via interpreting simple questions and using relational information to provide answers [3,4]. In a biomedical setting these graphs have been used to prioritize genes relevant to disease [5,6,7,8], perform drug repurposing [9] and identify drug-target interactions [10].

Despite their utility, precisely defining a knowledge graph is a difficult task because there are multiple conflicting definitions [11]. For this review, we define a knowledge graph as the following: a resource that integrates single or multiple sources of information into the form of a graph. This graph allows for the capacity to make semantic interpretation, continuously incorporate new information and uncover novel hidden knowledge through computational techniques and algorithms. Based on this definition resources like Hetionet [9] would be considered a knowledge graph. Hetionet integrates multiple sources of information into the form of a graph (example shown in Figure 1) and was used to derive novel information concerning unique drug treatments [9]. We do not consider databases like DISEASES [12] and DrugBank [13] to be knowledge graphs. These resources contain essential information, but do not represent their data in graph form.

Knowledge graphs are often constructed from manually curated databases [9,14,15,16]. These sources provide previously established information that can be incorporated into a graph. For example, a graph using DISEASES [12] as a resource would have genes and diseases as nodes, while edges would be added between nodes that have an association. This example shows a single type of relationship; however, there are graphs that use databases with multiple relationships. Other approaches have used natural language processing techniques to build knowledge graphs [17,18]. One example used a text mining system to extract sentences that indicated a protein interacting with another protein [19]. Once these sentences have been identified, they are incorporated as evidence for establishing edges in a knowledge graph.

In this review we describe various approaches for constructing and applying knowledge graphs in a biomedical setting. We discuss the pros and cons of constructing a knowledge graph via manually curated databases and via text mining systems. We also compare assorted approaches for applying knowledge graphs to solve biomedical problems. Lastly, we conclude on the practicality of knowledge graphs and point out future applications that have yet to be explored.

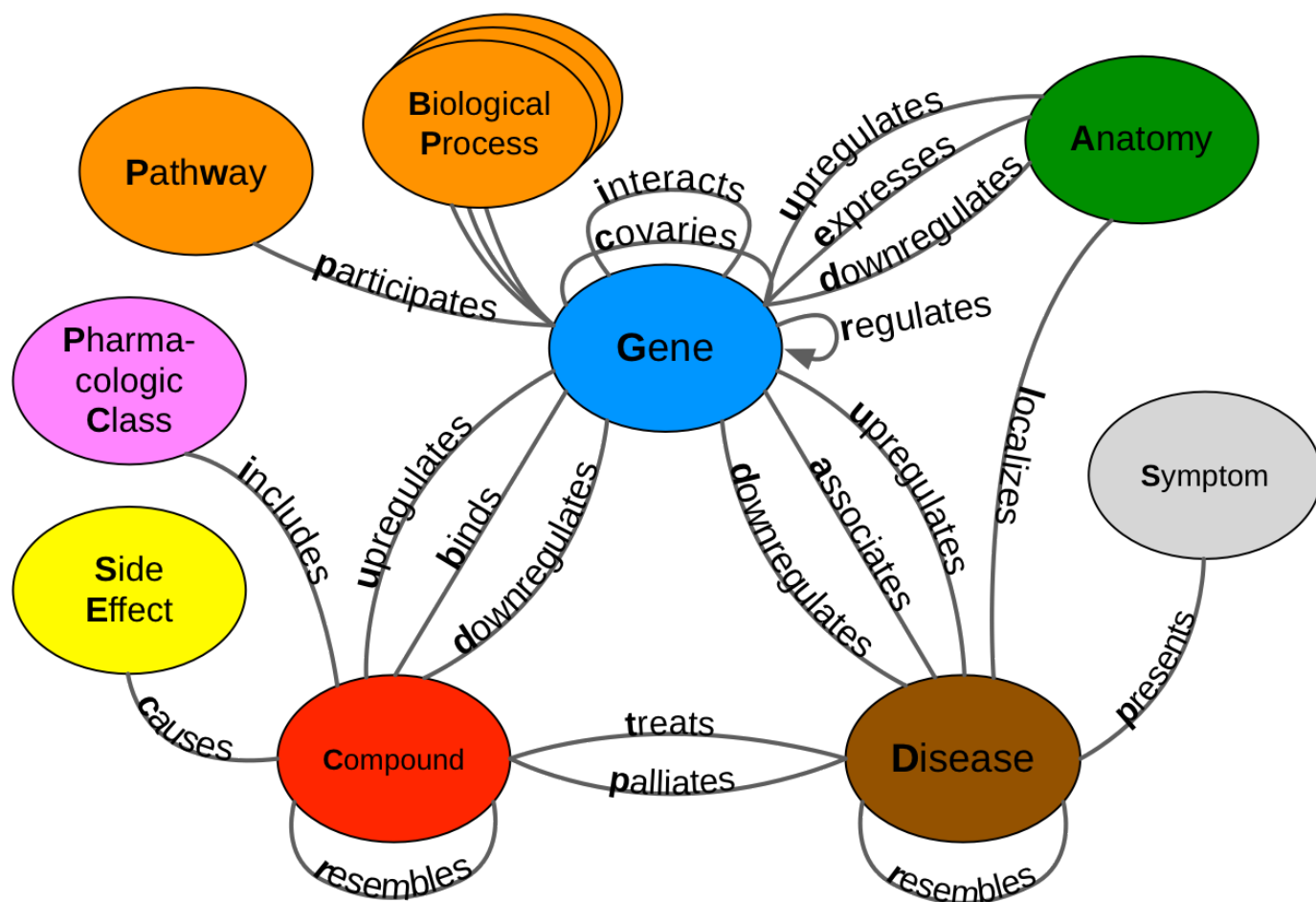


Figure 1: A metagraph (schema) of the heterogeneous network used in the Rephetio project [9]. This undirected network depicts pharmacological and biomedical information. The nodes (circles) represent entities and edges (lines) depict relational information between two entities.

Building Biomedical Knowledge Graphs

Knowledge graphs can be constructed in many ways using resources such as text or pre-existing databases. Usually, knowledge graphs are constructed using pre-existing databases. These databases are constructed by domain experts using approaches ranging from manual curation to automated techniques, such as text mining systems. Manual curation is a process that involves extensive use of domain experts to read papers and detect sentences that assert a relationship. Automated approaches involve the use of machine learning or natural language processing techniques to rapidly detect sentences of interest. We categorize these automated approaches into the following groups: rule-based extraction, unsupervised machine learning, and supervised machine learning. We discuss examples of each type of approach and synthesize the strengths and weaknesses of each.

Constructing Databases and Manual Curation

Database construction can date back all the way to 1956 where the first database contained a protein sequence of the insulin molecule [20]. This process involves gathering relevant text such as journal articles, abstracts, or web-based text. At this point curators can read gathered text and detect relationship asserting sentences (i.e. relationship extraction). An alternative to use a text mining system to filter out extraneous sentences, then incorporate curators to perfect the system's findings. This semi-automatic approach is way to augment curators throughout the curation process. We discuss the pros and cons of using manual curation for relationship extraction and mention databases that use this method to populate their fields.

Notable databases have been constructed via manual curation (Table {???}). For example, COSMIC [21] was constructed via a group of domain experts scanning the literature for key cancer related genes. This database has reached close to 35M entries in 2016 [21] and grew to a total of 45M entries in 2019 [22]. Studies have shown that these databases contain relatively precise data, but in low quantities [23,24,25,26,27,28,29]. This happens because the high publication rate is too much for curators to keep up [30]. This findings highlight a critical need for future approaches to be fast enough to compete with an increasing publication rate.

Semi-automatic methods are a way to augment curators during the curation process [27,31,32,33,34,35,36]. First step in this context is to use an automatic system to initially extract sentences from text. This process filters out irrelevant sentences, which means less text for curators to sift through. After the pre-filtering step curators can approve or remove the identified sentences. This semi-automatic process was found to speed up the curation process compared to manual approach [31,37]. Curators in [37] saved an average of 2.8 hours of overall time while curators in [31] saved about the same amount of time (2 hours). Despite the speed up, this process is prone to produce bias results. As automated systems excel in identifying sentences for commonly occurring relationships, they miss out on lessor known relationships [31]. Plus, these systems have a hard time parsing ambiguous sentences that naturally occur in text. This complication results in curators have a difficult time correcting these systems [31]. Given these caveats, a future direction would be using or creating approaches that can mitigate the relationship bias. Furthermore, future approaches should look into using techniques that simplify sentences to solve the ambiguity issue [38,39].

Despite the negatives of manual curation, it is still an essential process for relationship extraction approaches. This process can be used to generate gold standard datasets that automated systems use for validation [40,41]. Furthermore, manual curation can be used during the training process of automated systems (i.e. active learning) [42]. It is important to remember that manual curation alone is precise, but results in low recall rates [29]. Future databases should consider initially relying on automated methods to obtain sentences at an acceptable recall level, then incorporate manual curation as a way to fix or remove irrelevant results.

Database [Reference]	Short Description	Number of Entries	Entity Types	Relationship Types	Method of Population
Entrez-Gene [43]	NCBI's Gene annotation database that contains information pertaining to genes, gene's organism source, phenotypes etc.	7,883,114	Genes, Species and Phenotypes	Gene-Phenotypes and Genes-Species mappings	Semi-automated curation
UniProt [44]	A protein protein interaction database that contains proteomic information.	560,823	Proteins, Protein sequences	Protein-Protein interactions	Manual and Automated Curation
PharmGKB [45]	A database that contains genetic, phenotypic, and clinical information related to pharmacogenomic studies.	43,112	Drugs, Genes, Phenotypes, Variants, Pathways	Gene-Phenotypes, Pathway-Drugs, Gene-Variants, Gene-Pathways	Manual Curation and Automated Methods

Database [Reference]	Short Description	Number of Entries	Entity Types	Relationship Types	Method of Population
COSMIC [21]	A database that contains high resolution human cancer genetic information.	35,946,704	Genes, Variants, Tumor Types	Gene-Variant Mappings	Manual Curation
BioGrid [46]	A database for major model organisms. It contains genetic and proteomic information.	572,084	Genes, Proteins	Protein-Protein interactions	Semi-automatic methods
Comparative Toxicogenomics Database [47]	A database that contains manually curated chemical-gene-disease interactions and relationships.	2,429,689	Chemicals (Drugs), Genes, Diseases	Drug-Genes, Drug-Disease, Disease-Gene mappings	Manual curation and Automated systems
Comprehensive Antibiotic Resistance Database [48]	Manually curated database that contains information about the molecular basis of antimicrobial resistance.	174,443	Drugs, Genes, Variants	Drug-Gene, Drug-Variant mappings	Manual curation
OMIM [49]	A database that contains phenotype and genotype information	25,153	Genes, Phenotypes	Gene-Phenotype mappings	Manual Curation

Table. A table of databases that used a form of manual curation to populate entries. Reported number of entities and relationships are relative to time of publication. {#tbl:manual-curated-databases}

Text Mining for Relationship Extraction

Rule-Based Natural Language Processing

1. Mention papers on hand written rules and expressions

Extracting Relationships Without Labels

Unsupervised methods of extraction involve drawing inferences from data without the use of labels. These methods involve some form of clustering or statistical calculations. In this section we discuss methods that use unsupervised learning to detect relationship asserting sentences from text.

An unsupervised method to extract relationships exploits the fact that two entities can appear together in text. This kind of event is called co-occurrence and studies that use this phenomenon can be found in table 1. Two databases DISEASES [12] and STRING [50] were populated using a co-occurrence scoring method on PubMed abstracts. Both databases used the same scoring method that measured the frequency of co-mention pairs within individual sentences as well as the abstracts themselves. This method assumes independence between each individual occurrence. Under this

assumption mention pairs that occur more than expected were presumed to indicate the presence of an association or interaction. This approach was able to identify 543,405 disease gene associations [12] and 792,730 high confidence protein protein interactions [50], but is limited to only using PubMed abstracts.

Full text articles are able to drastically amplify text mining power to detect relationships [51, 52]. Westergaard et al. used a co-occurrence approach, similar to DISEASES [12] and STRING [50], to mine full articles for protein-protein interactions and other protein related information [51]. The authors discovered that full text provided better prediction power than using abstracts alone. This improvement suggests that future text mining approaches should consider using full text to increase detection power.

Unsupervised methods have been focused on treating multiple biomedical relationships as multiple isolated problems. These methods repeatedly use the same model for each biomedical relationship type. An alternative to this perspective is to capture all different relationship types at once. Clustering is an approach that accomplish this concept of simultaneous extraction. Percha et al. used a biclustering algorithm on generated dependency parse trees to group PubMed abstract sentences [53]. Each cluster was manually curated to determine which relationship they represented. This approach captured 4,451,661 dependency paths for 36 different groups [53]. Despite the success, this approach suffered from technical issues such as dependency tree parsing errors. This type of error resulted in sentences not being grouped by the clustering algorithm [53]. Future clustering approaches should consider simplifying sentences to prevent this type of issue.

Overall unsupervised methods provide a means to rapidly find relationship asserting sentences without the need of annotated text. Approaches in this category range from using co-occurrence scores to clustering sentences. These methods provide a generalizable framework that can be used on large repositories of text. Future methods can improve detection power by considering the use of methods that simplify sentences and use datasets that include full text articles.

Table 1: Table of approaches that mainly use a form of co-occurrence.

Study	Relationship of Interest
[54]	Protein-Protein Interactions, Disease-Gene and Tissue-Gene Associations
[55]	Drug Disease Treatments
[56]	Drug, Gene and Disease interactions
[51]	Protein-Protein Interactions
[12]	Disease-Gene associations
[57]	Protein-Protein Interactions
[58]	Genotype-Phenotype Relationships

Supervised Machine Learning

1. Mention the availability of publically available data

1. PPI - 5 datasets

1. 10.1016/j.artmed.2004.07.016

2. 10.1186/1471-2105-8-50

3. Learning language in logic - genic interaction extraction challenge

4. 10.1093/bioinformatics/btl616

5. <http://helix-web.stanford.edu/psb02/ding.pdf>

2. DaG - 3 datasets

1. 10.1016/j.jbi.2012.04.004
2. 10.1186/s12859-015-0472-9
3. 10.1186/1471-2105-14-323
4. 10.1186/1471-2105-13-161
3. CiD
4. 10.1093/database/baw068
5. CbG
6. Biocreative VI track 5 - raw citation
7. more if exists talk about deep learning methods
2. Mention the use of Support Vector Machines and other non deep learning classifiers
 1. Will have to mention that field has moved to deep learning.
 2. 10.1186/s13326-017-0168-3
 3. 10.1371/journal.pcbi.1004630
3. Mention deep learning methods
 1. 1901.06103v1
 2. 10.1016/j.knosys.2018.11.020
 3. 10.1177/0165551516673485
 4. 1706.01556v2
 5. ^^ A few papers here but a lot more will be put into place
 6. Mention caveat which is the need for large annotated datasets
 7. Mention a direction the field is moving to which is weak supervision and more that info that will come in time.

Applying Knowledge Graphs to Biomedical Challenges

1. Mention that these graphs can be used for discovery
2. Mention representation learning (aka representing a graph as dense vectors for nodes and/or edges)
- 3.

Unifying Techniques

1. Set up the problem that maps a knowledge graph into a low dimensional space

Matrix Factorization

1. Mention techniques for these with some papers

Deep Learning

1. Define node neighborhoods
2. Talk about random walks
3. Talk about auto encoders random walk independent approaches

Unifying Applications

1. Mention how the previous section is used in a biomedical setting

Disease and Gene Interactions

1. Mention disease gene prioritization
2. Mention Disease gene associations

Protein Protein Interactions

1. Mention predicting genes interacting genes

Drug Interactions

1. Talk about drug side effects
2. Drug repurposing
3. Drug-Disease Interactions

Clinical applications

1. Can mention EHR use and other related applications
2. Mention Tiffany's work on private data embeddings

Conclusion

1. Summarize discussed positives and pitfalls
2. Leave some open ended questions yet to be explored
3. Will come into play as I write this review paper

References

1. Node Classification in Social Networks

Smriti Bhagat, Graham Cormode, S. Muthukrishnan
arXiv (2011-01-17) <https://arxiv.org/abs/1101.3291v1>
DOI: [10.1007/978-1-4419-8462-3_5](https://doi.org/10.1007/978-1-4419-8462-3_5)

2. Network Embedding Based Recommendation Method in Social Networks

Yufei Wen, Lei Guo, Zhumin Chen, Jun Ma
Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18 (2018)
<https://doi.org/gf6rtt>
DOI: [10.1145/3184558.3186904](https://doi.org/10.1145/3184558.3186904)

3. Open Question Answering with Weakly Supervised Embedding Models

Antoine Bordes, Jason Weston, Nicolas Usunier
arXiv (2014-04-16) <https://arxiv.org/abs/1404.4326v1>

4. Neural Network-based Question Answering over Knowledge Graphs on Word and Character Level

Denis Lukovnikov, Asja Fischer, Jens Lehmann, Sören Auer
Proceedings of the 26th International Conference on World Wide Web - WWW '17 (2017)
<https://doi.org/gfv8hp>
DOI: [10.1145/3038912.3052675](https://doi.org/10.1145/3038912.3052675)

5. Towards integrative gene prioritization in Alzheimer's disease.

Jang H Lee, Graciela H Gonzalez
Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2011)
<https://www.ncbi.nlm.nih.gov/pubmed/21121028>
DOI: [10.1142/9789814335058_0002](https://doi.org/10.1142/9789814335058_0002) · PMID: [21121028](https://pubmed.ncbi.nlm.nih.gov/21121028/)

6. PhenoGeneRanker: A Tool for Gene Prioritization Using Complete Multiplex Heterogeneous Networks

Cagatay Dursun, Naoki Shimoyama, Mary Shimoyama, Michael Schläppi, Serdar Bozdogan
Cold Spring Harbor Laboratory (2019-05-27) <https://doi.org/gf6rtr>
DOI: [10.1101/651000](https://doi.org/10.1101/651000)

7. Biological Random Walks: Integrating heterogeneous data in disease gene prioritization

Michele Gentili, Leonardo Martini, Manuela Petti, Lorenzo Farina, Luca Becchetti
2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (2019-07) <https://doi.org/gf6rts>
DOI: [10.1109/cibcb.2019.8791472](https://doi.org/10.1109/cibcb.2019.8791472)

8. Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes

Mona Alshahrani, Robert Hoehndorf
Bioinformatics (2018-09-01) <https://doi.org/gd9k8n>
DOI: [10.1093/bioinformatics/bty559](https://doi.org/10.1093/bioinformatics/bty559) · PMID: [30423077](https://pubmed.ncbi.nlm.nih.gov/30423077/) · PMCID: [PMC6129260](https://pubmed.ncbi.nlm.nih.gov/PMC6129260/)

9. Systematic integration of biomedical knowledge prioritizes drugs for repurposing

Daniel Scott Himmelstein, Antoine Lizée, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini

eLife (2017-09-22) <https://doi.org/cdfk>
DOI: [10.7554/elife.26726](https://doi.org/10.7554/elife.26726) · PMID: [28936969](https://pubmed.ncbi.nlm.nih.gov/28936969/) · PMCID: [PMC5640425](https://pubmed.ncbi.nlm.nih.gov/PMC5640425/)

10. Assessing Drug Target Association Using Semantic Linked Data

Bin Chen, Ying Ding, David J. Wild

PLoS Computational Biology (2012-07-05) <https://doi.org/rn6>

DOI: [10.1371/journal.pcbi.1002574](https://doi.org/10.1371/journal.pcbi.1002574) · PMID: [22859915](https://pubmed.ncbi.nlm.nih.gov/22859915/) · PMCID: [PMC3390390](https://pubmed.ncbi.nlm.nih.gov/PMC3390390/)

11. Towards a definition of knowledge graphs

Lisa Ehrlinger, Wolfram Wöß

SEMANTiCS (2016)

12. DISEASES: Text mining and data integration of disease-gene associations

Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafo, Janos X. Binder, Lars Juhl Jensen

Methods (2015-03) <https://doi.org/f3mn6s>

DOI: [10.1016/j.ymeth.2014.11.020](https://doi.org/10.1016/j.ymeth.2014.11.020) · PMID: [25484339](https://pubmed.ncbi.nlm.nih.gov/25484339/)

13. DrugBank 5.0: a major update to the DrugBank database for 2018

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, ... Michael Wilson

Nucleic Acids Research (2017-11-08) <https://doi.org/gcwtzk>

DOI: [10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037) · PMID: [29126136](https://pubmed.ncbi.nlm.nih.gov/29126136/) · PMCID: [PMC5753335](https://pubmed.ncbi.nlm.nih.gov/PMC5753335/)

14. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information

Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, Jianyang Zeng

Nature Communications (2017-09-18) <https://doi.org/gbxwrc>

DOI: [10.1038/s41467-017-00680-8](https://doi.org/10.1038/s41467-017-00680-8) · PMID: [28924171](https://pubmed.ncbi.nlm.nih.gov/28924171/) · PMCID: [PMC5603535](https://pubmed.ncbi.nlm.nih.gov/PMC5603535/)

15. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks

Hui Liu, Yinglong Song, Jihong Guan, Libo Luo, Ziheng Zhuang

BMC Bioinformatics (2016-12) <https://doi.org/gf6v27>

DOI: [10.1186/s12859-016-1336-7](https://doi.org/10.1186/s12859-016-1336-7) · PMID: [28155639](https://pubmed.ncbi.nlm.nih.gov/28155639/) · PMCID: [PMC5259862](https://pubmed.ncbi.nlm.nih.gov/PMC5259862/)

16. Finding disease similarity based on implicit semantic similarity

Sachin Mathur, Deendayal Dinakarpandian

Journal of Biomedical Informatics (2012-04) <https://doi.org/b7b3tw>

DOI: [10.1016/j.jbi.2011.11.017](https://doi.org/10.1016/j.jbi.2011.11.017) · PMID: [22166490](https://pubmed.ncbi.nlm.nih.gov/22166490/)

17. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences

Patrick Ernst, Amy Siu, Gerhard Weikum

BMC Bioinformatics (2015-05-14) <https://doi.org/gb8w8d>

DOI: [10.1186/s12859-015-0549-5](https://doi.org/10.1186/s12859-015-0549-5) · PMID: [25971816](https://pubmed.ncbi.nlm.nih.gov/25971816/) · PMCID: [PMC4448285](https://pubmed.ncbi.nlm.nih.gov/PMC4448285/)

18. Constructing biomedical domain-specific knowledge graph with minimum supervision

Jianbo Yuan, Zhiwei Jin, Han Guo, Hongxia Jin, Xianchao Zhang, Tristram Smith, Jiebo Luo

Knowledge and Information Systems (2019-03-23) <https://doi.org/gf6v26>

DOI: [10.1007/s10115-019-01351-4](https://doi.org/10.1007/s10115-019-01351-4)

19. Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein-protein interaction

Shweta Yadav, Asif Ekbal, Sriparna Saha, Ankit Kumar, Pushpak Bhattacharyya
Knowledge-Based Systems (2019-02) <https://doi.org/gf4788>
DOI: [10.1016/j.knosys.2018.11.020](https://doi.org/10.1016/j.knosys.2018.11.020)

20. Biological Databases- Integration of Life Science Data

Nishant Toomula, Arun Kumar, Sathish Kumar D, Vijaya Shanti Bheemidi
Journal of Computer Science & Systems Biology (2012) <https://doi.org/gf8qcb>
DOI: [10.4172/jcsb.1000081](https://doi.org/10.4172/jcsb.1000081)

21. COSMIC: somatic cancer genetics at high-resolution

Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, ... Peter J. Campbell
Nucleic Acids Research (2016-11-28) <https://doi.org/f9v865>
DOI: [10.1093/nar/gkw1121](https://doi.org/10.1093/nar/gkw1121) · PMID: [27899578](https://pubmed.ncbi.nlm.nih.gov/27899578/) · PMCID: [PMC5210583](https://pubmed.ncbi.nlm.nih.gov/PMC5210583/)

22. COSMIC: the Catalogue Of Somatic Mutations In Cancer

John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, ... Simon A Forbes
Nucleic Acids Research (2018-10-29) <https://doi.org/gf9hxxg>
DOI: [10.1093/nar/gky1015](https://doi.org/10.1093/nar/gky1015) · PMID: [30371878](https://pubmed.ncbi.nlm.nih.gov/30371878/) · PMCID: [PMC6323903](https://pubmed.ncbi.nlm.nih.gov/PMC6323903/)

23. Recurated protein interaction datasets

Lukasz Salwinski, Luana Licata, Andrew Winter, David Thorneycroft, Jyoti Khadake, Arnaud Ceol, Andrew Chatr Aryamontri, Rose Oughtred, Michael Livstone, Lorrie Boucher, ... Henning Hermjakob
Nature Methods (2009-12) <https://doi.org/fgvkmmf>
DOI: [10.1038/nmeth1209-860](https://doi.org/10.1038/nmeth1209-860) · PMID: [19935838](https://pubmed.ncbi.nlm.nih.gov/19935838/)

24. Literature-curated protein interaction datasets

Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-François Rual, Heather Borick, Pascal Braun, Matija Dreze, ... Marc Vidal
Nature Methods (2008-12-30) <https://doi.org/d4j62p>
DOI: [10.1038/nmeth.1284](https://doi.org/10.1038/nmeth.1284) · PMID: [19116613](https://pubmed.ncbi.nlm.nih.gov/19116613/) · PMCID: [PMC2683745](https://pubmed.ncbi.nlm.nih.gov/PMC2683745/)

25. Curation accuracy of model organism databases

I. M. Keseler, M. Skrzypek, D. Weerasinghe, A. Y. Chen, C. Fulcher, G.-W. Li, K. C. Lemmer, K. M. Mladinich, E. D. Chow, G. Sherlock, P. D. Karp
Database (2014-06-12) <https://doi.org/gf63jz>
DOI: [10.1093/database/bau058](https://doi.org/10.1093/database/bau058) · PMID: [24923819](https://pubmed.ncbi.nlm.nih.gov/24923819/) · PMCID: [PMC4207230](https://pubmed.ncbi.nlm.nih.gov/PMC4207230/)

26. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders

Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, Ada Hamosh
Nucleic Acids Research (2014-11-26) <https://doi.org/gf8qgb6>
DOI: [10.1093/nar/gku1205](https://doi.org/10.1093/nar/gku1205) · PMID: [25428349](https://pubmed.ncbi.nlm.nih.gov/25428349/) · PMCID: [PMC4383985](https://pubmed.ncbi.nlm.nih.gov/PMC4383985/)

27. Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature

H.-M. Müller, K. M. Van Auken, Y. Li, P. W. Sternberg
BMC Bioinformatics (2018-03-09) <https://doi.org/gf7rbz>
DOI: [10.1186/s12859-018-2103-8](https://doi.org/10.1186/s12859-018-2103-8) · PMID: [29523070](https://pubmed.ncbi.nlm.nih.gov/29523070/) · PMCID: [PMC5845379](https://pubmed.ncbi.nlm.nih.gov/PMC5845379/)

28. Text mining and expert curation to develop a database on psychiatric diseases and their genes

Alba Gutiérrez-Sacristán, Àlex Bravo, Marta Portero-Tresserra, Olga Valverde, Antonio Armario, M. C. Blanco-Gandía, Adriana Farré, Lierni Fernández-Ibarrondo, Francina Fonseca, Jesús Giraldo, ... Laura I. Furlong

Database (2017-01-01) <https://doi.org/gf8qb5>

DOI: [10.1093/database/bax043](https://doi.org/10.1093/database/bax043) · PMID: [29220439](https://pubmed.ncbi.nlm.nih.gov/29220439/) · PMCID: [PMC5502359](https://pubmed.ncbi.nlm.nih.gov/PMC5502359/)

29. Manual curation is not sufficient for annotation of genomic databases

William A. Baumgartner Jr, K. Bretonnel Cohen, Lynne M. Fox, George Acquah-Mensah, Lawrence Hunter

Bioinformatics (2007-07-01) <https://doi.org/dtck86>

DOI: [10.1093/bioinformatics/btm229](https://doi.org/10.1093/bioinformatics/btm229) · PMID: [17646325](https://pubmed.ncbi.nlm.nih.gov/17646325/) · PMCID: [PMC2516305](https://pubmed.ncbi.nlm.nih.gov/PMC2516305/)

30. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index

Peder Olesen Larsen, Markus von Ins

Scientometrics (2010-03-10) <https://doi.org/c4hb8r>

DOI: [10.1007/s11192-010-0202-z](https://doi.org/10.1007/s11192-010-0202-z) · PMID: [20700371](https://pubmed.ncbi.nlm.nih.gov/20700371/) · PMCID: [PMC2909426](https://pubmed.ncbi.nlm.nih.gov/PMC2909426/)

31. Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction

Aurélié Névél, Rezarta Islamaj Doğan, Zhiyong Lu

Journal of Biomedical Informatics (2011-04) <https://doi.org/bq34sj>

DOI: [10.1016/j.jbi.2010.11.001](https://doi.org/10.1016/j.jbi.2010.11.001) · PMID: [21094696](https://pubmed.ncbi.nlm.nih.gov/21094696/) · PMCID: [PMC3063330](https://pubmed.ncbi.nlm.nih.gov/PMC3063330/)

32. Assisting manual literature curation for protein-protein interactions using BioQRator

D. Kwon, S. Kim, S.-Y. Shin, A. Chatr-aryamontri, W. J. Wilbur

Database (2014-07-22) <https://doi.org/gf7hm3>

DOI: [10.1093/database/bau067](https://doi.org/10.1093/database/bau067) · PMID: [25052701](https://pubmed.ncbi.nlm.nih.gov/25052701/) · PMCID: [PMC4105708](https://pubmed.ncbi.nlm.nih.gov/PMC4105708/)

33. Argo: an integrative, interactive, text mining-based workbench supporting curation

R. Rak, A. Rowley, W. Black, S. Ananiadou

Database (2012-03-20) <https://doi.org/h5d>

DOI: [10.1093/database/bas010](https://doi.org/10.1093/database/bas010) · PMID: [22434844](https://pubmed.ncbi.nlm.nih.gov/22434844/) · PMCID: [PMC3308166](https://pubmed.ncbi.nlm.nih.gov/PMC3308166/)

34. CurEx

Michael Loster, Felix Naumann, Jan Ehmueller, Benjamin Feldmann

Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18 (2018) <https://doi.org/gf8qb8>

DOI: [10.1145/3269206.3269229](https://doi.org/10.1145/3269206.3269229)

35. Re-curation and rational enrichment of knowledge graphs in Biological Expression Language

Charles Tapley Hoyt, Daniel Domingo-Fernández, Rana Aldisi, Lingling Xu, Kristian Kolpeja, Sandra Spalek, Esther Wollert, John Bachman, Benjamin M Gyori, Patrick Greene, Martin Hofmann-Apitius

Database (2019-01-01) <https://doi.org/gf7hm4>

DOI: [10.1093/database/baz068](https://doi.org/10.1093/database/baz068) · PMID: [31225582](https://pubmed.ncbi.nlm.nih.gov/31225582/) · PMCID: [PMC6587072](https://pubmed.ncbi.nlm.nih.gov/PMC6587072/)

36. LocText: relation extraction of protein localizations to assist database curation

Juan Miguel Cejuela, Shrikant Vinchurkar, Tatyana Goldberg, Madhukar Sollepura Prabhu Shankar, Ashish Baghudana, Aleksandar Bojchevski, Carsten Uhlig, André Ofner, Pandu Raharja-Liu, Lars Juhl Jensen, Burkhard Rost

BMC Bioinformatics (2018-01-17) <https://doi.org/gf8qb9>
DOI: [10.1186/s12859-018-2021-9](https://doi.org/10.1186/s12859-018-2021-9) · PMID: [29343218](https://pubmed.ncbi.nlm.nih.gov/29343218/) · PMCID: [PMC5773052](https://pubmed.ncbi.nlm.nih.gov/PMC5773052/)

37. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements

Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, Imre Solti

Journal of the American Medical Informatics Association (2014-05) <https://doi.org/f5zggh>
DOI: [10.1136/amiajnl-2013-001837](https://doi.org/10.1136/amiajnl-2013-001837) · PMID: [24001514](https://pubmed.ncbi.nlm.nih.gov/24001514/) · PMCID: [PMC3994857](https://pubmed.ncbi.nlm.nih.gov/PMC3994857/)

38. iSimp in BioC standard format: enhancing the interoperability of a sentence simplification system

Y. Peng, C. O. Tudor, M. Torii, C. H. Wu, K. Vijay-Shanker

Database (2014-05-21) <https://doi.org/gf9hxf>
DOI: [10.1093/database/bau038](https://doi.org/10.1093/database/bau038) · PMID: [24850848](https://pubmed.ncbi.nlm.nih.gov/24850848/) · PMCID: [PMC4028706](https://pubmed.ncbi.nlm.nih.gov/PMC4028706/)

39. BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction.

Siddhartha Jonnalagadda, Graciela Gonzalez

AMIA ... Annual Symposium proceedings. AMIA Symposium (2010-11-13)
<https://www.ncbi.nlm.nih.gov/pubmed/21346999>
PMID: [21346999](https://pubmed.ncbi.nlm.nih.gov/21346999/) · PMCID: [PMC3041388](https://pubmed.ncbi.nlm.nih.gov/PMC3041388/)

40. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships

Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhua Nieto, Gianluca Trifiro, Jan A. Kors, Laura I. Furlong

Journal of Biomedical Informatics (2012-10) <https://doi.org/f36vn6>
DOI: [10.1016/j.jbi.2012.04.004](https://doi.org/10.1016/j.jbi.2012.04.004) · PMID: [22554700](https://pubmed.ncbi.nlm.nih.gov/22554700/)

41. Comparative experiments on learning information extractors for proteins and their interactions

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, Yuk Wah Wong

Artificial Intelligence in Medicine (2005-02) <https://doi.org/dhztptn>
DOI: [10.1016/j.artmed.2004.07.016](https://doi.org/10.1016/j.artmed.2004.07.016) · PMID: [15811782](https://pubmed.ncbi.nlm.nih.gov/15811782/)

42. A Unified Active Learning Framework for Biomedical Relation Extraction

Hong-Tao Zhang, Min-Lie Huang, Xiao-Yan Zhu

Journal of Computer Science and Technology (2012-11) <https://doi.org/gf8qb4>
DOI: [10.1007/s11390-012-1306-0](https://doi.org/10.1007/s11390-012-1306-0)

43. Entrez Gene: gene-centered information at NCBI

D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova

Nucleic Acids Research (2010-11-28) <https://doi.org/fsjqczz>
DOI: [10.1093/nar/gkq1237](https://doi.org/10.1093/nar/gkq1237) · PMID: [21115458](https://pubmed.ncbi.nlm.nih.gov/21115458/) · PMCID: [PMC3013746](https://pubmed.ncbi.nlm.nih.gov/PMC3013746/)

44. UniProt: a worldwide hub of protein knowledge *Nucleic Acids Research* (2018-11-05)

<https://doi.org/gfwqck>
DOI: [10.1093/nar/gky1049](https://doi.org/10.1093/nar/gky1049) · PMID: [30395287](https://pubmed.ncbi.nlm.nih.gov/30395287/) · PMCID: [PMC6323992](https://pubmed.ncbi.nlm.nih.gov/PMC6323992/)

45. Pharmacogenomics Knowledge for Personalized Medicine

M Whirl-Carrillo, EM McDonagh, JM Hebert, L Gong, K Sangkuhl, CF Thorn, RB Altman, TE Klein

Clinical Pharmacology & Therapeutics (2012-10) <https://doi.org/gdnfzr>
DOI: [10.1038/clpt.2012.96](https://doi.org/10.1038/clpt.2012.96) · PMID: [22992668](https://pubmed.ncbi.nlm.nih.gov/22992668/) · PMCID: [PMC3660037](https://pubmed.ncbi.nlm.nih.gov/PMC3660037/)

46. The BioGRID interaction database: 2013 update

Andrew Chatr-aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, ... Mike Tyers
Nucleic Acids Research (2012-11-30) <https://doi.org/f4jnz4>
DOI: [10.1093/nar/gks1158](https://doi.org/10.1093/nar/gks1158) · PMID: [23203989](https://pubmed.ncbi.nlm.nih.gov/23203989/) · PMCID: [PMC3531226](https://pubmed.ncbi.nlm.nih.gov/PMC3531226/)

47. The Comparative Toxicogenomics Database: update 2019

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, Carolyn J Mattingly
Nucleic Acids Research (2018-09-24) <https://doi.org/gf8qb7>
DOI: [10.1093/nar/gky868](https://doi.org/10.1093/nar/gky868) · PMID: [30247620](https://pubmed.ncbi.nlm.nih.gov/30247620/) · PMCID: [PMC6323936](https://pubmed.ncbi.nlm.nih.gov/PMC6323936/)

48. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database

Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, ... Andrew G. McArthur
Nucleic Acids Research (2016-10-26) <https://doi.org/f9wbjs>
DOI: [10.1093/nar/gkw1004](https://doi.org/10.1093/nar/gkw1004) · PMID: [27789705](https://pubmed.ncbi.nlm.nih.gov/27789705/) · PMCID: [PMC5210516](https://pubmed.ncbi.nlm.nih.gov/PMC5210516/)

49. OMIM.org: leveraging knowledge across phenotype-gene relationships.

Joanna S Amberger, Carol A Bocchini, Alan F Scott, Ada Hamosh
Nucleic acids research (2019-01-08) <https://www.ncbi.nlm.nih.gov/pubmed/30445645>
DOI: [10.1093/nar/gky1151](https://doi.org/10.1093/nar/gky1151) · PMID: [30445645](https://pubmed.ncbi.nlm.nih.gov/30445645/) · PMCID: [PMC6323937](https://pubmed.ncbi.nlm.nih.gov/PMC6323937/)

50. STRING v9.1: protein-protein interaction networks, with increased coverage and integration

Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, Lars J. Jensen
Nucleic Acids Research (2012-11-29) <https://doi.org/gf5kcd>
DOI: [10.1093/nar/gks1094](https://doi.org/10.1093/nar/gks1094) · PMID: [23203871](https://pubmed.ncbi.nlm.nih.gov/23203871/) · PMCID: [PMC3531103](https://pubmed.ncbi.nlm.nih.gov/PMC3531103/)

51. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts

David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, Søren Brunak
PLOS Computational Biology (2018-02-15) <https://doi.org/gcx747>
DOI: [10.1371/journal.pcbi.1005962](https://doi.org/10.1371/journal.pcbi.1005962) · PMID: [29447159](https://pubmed.ncbi.nlm.nih.gov/29447159/) · PMCID: [PMC5831415](https://pubmed.ncbi.nlm.nih.gov/PMC5831415/)

52. STITCH 4: integration of protein-chemical interactions with user data

Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H. Blicher, Christian von Mering, Lars J. Jensen, Peer Bork
Nucleic Acids Research (2013-11-28) <https://doi.org/f5shb4>
DOI: [10.1093/nar/gkt1207](https://doi.org/10.1093/nar/gkt1207) · PMID: [24293645](https://pubmed.ncbi.nlm.nih.gov/24293645/) · PMCID: [PMC3964996](https://pubmed.ncbi.nlm.nih.gov/PMC3964996/)

53. A global network of biomedical relationships derived from text

Bethany Percha, Russ B Altman
Bioinformatics (2018-02-27) <https://doi.org/gc3ndk>
DOI: [10.1093/bioinformatics/bty114](https://doi.org/10.1093/bioinformatics/bty114) · PMID: [29490008](https://pubmed.ncbi.nlm.nih.gov/29490008/) · PMCID: [PMC6061699](https://pubmed.ncbi.nlm.nih.gov/PMC6061699/)

54. CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision

Alexander Junge, Lars Juhl Jensen

Bioinformatics (2019-06-14) <https://doi.org/gf4789>
DOI: [10.1093/bioinformatics/btz490](https://doi.org/10.1093/bioinformatics/btz490) · PMID: [31199464](https://pubmed.ncbi.nlm.nih.gov/31199464/)

55. A new method for prioritizing drug repositioning candidates extracted by literature-based discovery

Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Dingcheng Li, Rashmi Prasad, Hongfang Liu
2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2015-11)
<https://doi.org/gf479j>
DOI: [10.1109/bibm.2015.7359766](https://doi.org/10.1109/bibm.2015.7359766)

56. Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases

Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, Wynand Alkema
PLoS Computational Biology (2010-09-23) <https://doi.org/bhrw7x>
DOI: [10.1371/journal.pcbi.1000943](https://doi.org/10.1371/journal.pcbi.1000943) · PMID: [20885778](https://pubmed.ncbi.nlm.nih.gov/20885778/) · PMCID: [PMC2944780](https://pubmed.ncbi.nlm.nih.gov/PMC2944780/)

57. STRING v10: protein–protein interaction networks, integrated over the tree of life

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, ... Christian von Mering
Nucleic Acids Research (2014-10-28) <https://doi.org/f64rfn>
DOI: [10.1093/nar/gku1003](https://doi.org/10.1093/nar/gku1003) · PMID: [25352553](https://pubmed.ncbi.nlm.nih.gov/25352553/) · PMCID: [PMC4383874](https://pubmed.ncbi.nlm.nih.gov/PMC4383874/)

58. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine

Ayush Singhal, Michael Simmons, Zhiyong Lu
PLOS Computational Biology (2016-11-30) <https://doi.org/f9gz4b>
DOI: [10.1371/journal.pcbi.1005017](https://doi.org/10.1371/journal.pcbi.1005017) · PMID: [27902695](https://pubmed.ncbi.nlm.nih.gov/27902695/) · PMCID: [PMC5130168](https://pubmed.ncbi.nlm.nih.gov/PMC5130168/)