Constructing Knowledge Graphs and Their Biomedical Applications

This manuscript (permalink) was automatically generated from greenelab/knowledge-graph-review@e15d016 on September 27, 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.

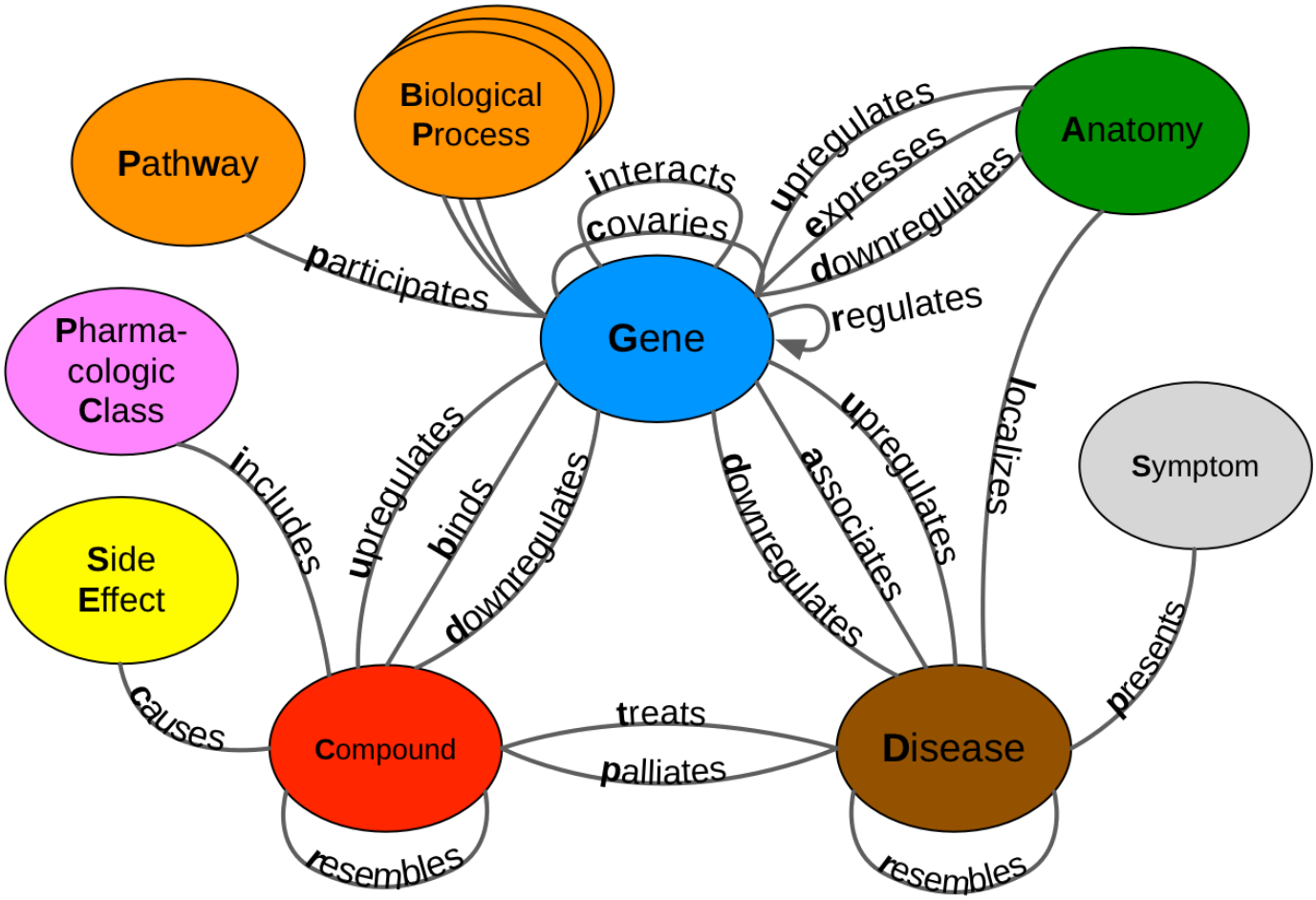

Figure 1: A metagraph (schema) of the heterogeneous network used in the Rephetio project [9]. This undirected network depicts pharmacological and biomedical information. The nodes (circles) represent entities and edges (lines) depict relational information between two entities.

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

1. Talk about papers that construct knowledge graphs without text mining approaches
2. Discuss the positives and negatives for these methods

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 [20] 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 [20], but is limited to only using PubMed abstracts.

Full text articles are able to drastically amplify text mining power to detect relationships [21,22]. Westergaard et al. used a co-occurrence approach, similar to DISEASES [12] and STRING [20], to mine full articles for protein-protein interactions and other protein related information [21]. 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 persepctive 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 [23]. Each cluster was manually curated to determine which relationship they represented. This approach captured 4,451,661 dependency paths for 36 different groups [23]. 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 [23]. 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
[24]	Protein-Protein Interactions, Disease-Gene and Tissue-Gene Associations
[25]	Drug Disease Treatments
[26]	Drug, Gene and Disease interactions
[21]	Protein-Protein Interactions
[12]	Disease-Gene associations
[27]	Protein-Protein Interactions
[28]	Genotype-Phenotype Relationships

Supervised Machine Learning

1. Mention the availablility 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 Interations

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

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

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

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
PMID: 21121028

6. **PhenoGeneRanker: A Tool for Gene Prioritization Using Complete Multiplex Heterogeneous Networks**
Cagatay Dursun, Naoki Shimoyama, Mary Shimoyama, Michael Schläppi, Serdar Bozdag
Cold Spring Harbor Laboratory (2019-05-27) https://doi.org/gf6rtr
DOI: 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

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 · PMID: 30423077 · PMCID: PMC6129260

9. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, 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 · PMID: 28936969 · PMCID: 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 · PMID: 22859915 · PMCID: 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 Tsafou, Janos X. Binder, Lars Juhl Jensen
Methods (2015-03) https://doi.org/f3mn6s
DOI: 10.1016/j.ymeth.2014.11.020 · PMID: 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 · PMID: 29126136 · PMCID: 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 · PMID: 28924171 · PMCID: 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 · PMID: 28155639 · PMCID: 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 · PMID: 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 · PMID: 25971816 · PMCID: 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

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

20. **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 · PMID: 23203871 · PMCID: PMC3531103

21. **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 · PMID: 29447159 · PMCID: PMC5831415

22. **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 · PMID: 24293645 · PMCID: PMC3964996

23. **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 · PMID: 29490008 · PMCID: PMC6061699

24. **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 · PMID: 31199464

25. **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

26. **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 · PMID: 20885778 · PMCID: PMC2944780

27. **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 · PMID: 25352553 · PMCID: PMC4383874

28. **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 · PMID: 27902695 · PMCID: PMC5130168