KG-Microbe: a reference knowledge-graph and platform for harmonized microbial information

Marcin P. Joachimiak
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
MJoachimiak@lbl.gov

Harshad Hegde
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
hhegde@lbl.gov

William D. Duncan
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
wdduncan@lbl.gov

Justin T. Reese
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
JustinReese@lbl.gov

Luca Cappelletti
University of Milan, Milan, Italy
luca.cappelletti1@unimi.it

Anne E. Thessen
Oregon State University, Beaverton, Oregon, USA
thessena@oregonstate.edu

Christopher J. Mungall
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
CJMungall@lbl.gov

Abstract

Microorganisms (microbes) are incredibly diverse, spanning all major divisions of life, and represent the greatest fraction of known species. A vast amount of knowledge about microbes is available in the literature, across experimental datasets, and in established data resources. While the genomic and biochemical pathway data about microbes is well-structured and annotated using standard ontologies, broader information about microbes and their ecological traits is not. We created the KG-Microbe (github.com/Knowledge-Graph-Hub/kg-microbe) resource in order to extract and integrate diverse knowledge about microbes from a variety of structured and unstructured sources. Initially, we are harmonizing and linking prokaryotic data for phenotypic traits, taxonomy, functions, chemicals, and environment descriptors, to construct a knowledge graph with over 266,000 entities linked by 432,000 relations. The effort is supported by a knowledge graph construction platform (KG-Hub) for rapid development of knowledge graphs using available data, knowledge modeling principles, and software tools. KG-Microbe is a microbe-centric Knowledge Graph (KG) to support tasks such as querying and graph link prediction in many use cases including microbiology, biomedicine, and the environment. KG-Microbe fulfills a need for standardized and linked microbial data, allowing the broader community to contribute, query, and enrich analyses and algorithms.

Keywords: knowledge graph, microbiology, ontology, graph learning, data standardization, data science, semantic technology

Main

Not only are microbes the most abundant and diverse life forms, but they are also found in the greatest range of environments and possess the largest metabolic and functional potential which is just beginning to be harnessed for biomedicine and biomanufacturing. A vast amount of knowledge about microbes is available in the literature, across experimental datasets, and in established data resources. While the genomic and biochemical pathway data about microbes is well-structured and annotated using standard ontologies, broader information about microbes and their ecological traits is not.

We draw inspiration from the biomedical domain which has a rich set of ontologies, controlled vocabularies, and data schemas, which have been deployed in multiple biomedical knowledge resources. Examples include data schemas such as OMOP (Voss et al. 2015), MeSH (Agarwal and Searls 2009), and the Biolink Model (Mungall et al 2021a), collections of ontologies such as the OBO Foundry (Smith et al. 2007) and the NCBO Bioportal (Whetzel et al. 2011), and a growing ecosystem of graph and Natural Language Processing (NLP) tools. These resources have helped drive the standardization and interoperability of information across research domains. This set of concepts and tools provides a path to useful semantic harmonization of knowledge in other domains.

The KG-Microbe knowledge graph (KG) resource was created to support extraction and integration of diverse knowledge about microbes. Initially, our focus has been on harmonizing and linking prokaryotic data for phenotypic traits, taxonomy, functions, chemicals, and environment descriptors. Based on this information we constructed a knowledge graph, which in its current release (9/2/21) contains over 266,000 entities linked by 432,000 relations, classified into 9 and 31 Biolink Model entity and relation categories, respectively. The KG-Microbe knowledge graph effort is supported by a knowledge graph construction platform, KG-Hub (Mungall et al 2021b), designed for rapid development and deployment of knowledge graphs using available data, common knowledge modeling principles, and software tools. One key concept for KG-Hub is to connect unstructured data into broader knowledge using links to structured data such as ontologies.

KG-Microbe (github.com/Knowledge-Graph-Hub/kg-microbe) is a microbe-centric Knowledge Graph (KG) to support tasks such as querying and graph link prediction in a variety of use cases including microbiology, biomedicine, and the environment. We use Named Entity Recognition (NER) and Natural Language

Processing (NLP) tools to identify, annotate, and normalize terms found in raw data. The harmonized data contained in the KG-Microbe knowledge graph provides rich and standardized labeling for building, training, and evaluating machine learning models. The resulting KG-Microbe graph is able to answer questions like which microbes are enriched in soil environments. It can also be used to train models for various microbial trait predictions and it can report enriched features for a given set of taxa or taxa features. We demonstrate example applications of KG-Microbe with predictive models for microbial shape and metabolism using embeddings (**Figure 1**) from graph learning. Many other types of link predictions are possible based on the available KG-Microbe entity and relation categories, allowing predictions for data, which is difficult to obtain without resource intensive field and laboratory experiments such as metabolic characterization or cell imaging. KG-Microbe fulfills a need for standardized and linked microbial data, allowing the broader community to contribute as well as enrich analyses and algorithms.

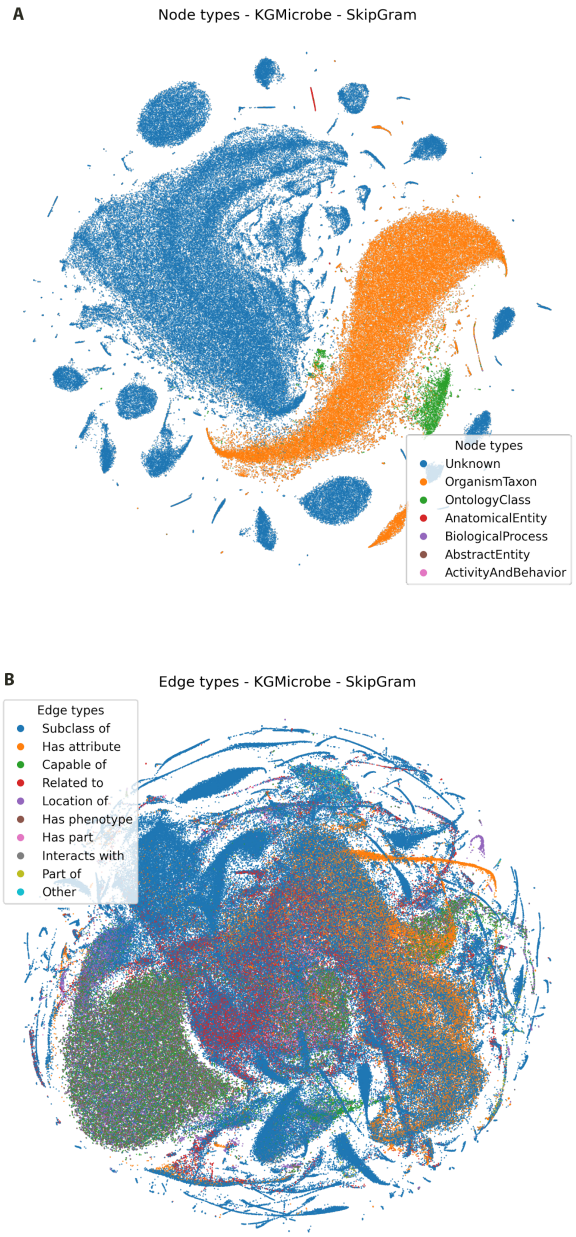

Figure 1: Visualization of tSNE (Maaten and Hinton 2008) dimensionality reduction of KG-Microbe graph node (**A**) and edge (**B**) embeddings, respectively. Each point corresponds to a node or edge and is colored by Biolink Model categories for edge and node types. Graph embeddings were generated with the embiggen package (Cappelletti et al. 2021), using the SkipGram method.

2.2.5 Acknowledgments

This work was supported by a grant from the Laboratory Directed Research and Development (LDRD) Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231.

References

Agarwal, Pankaj, and David B. Searls. 2009. “Can Literature Analysis Identify Innovation Drivers in Drug Discovery?” *Nature Reviews. Drug Discovery* 8 (11): 865–78.

Cappelletti, L., T. Fontana, E. Casiraghi, V. Ravanmehr, T. J. Callahan, M. P. Joachimiak, C. J. Mungall, P. N. Robinson, J. Reese, and Valentini Giorgio. 2021. *Embiggen: Embedding Generator - Embiggen*. Github. https://github.com/monarch-initiative/embiggen.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9 (Nov): 2579–2605.

Mungall, C. J., et al. 2021a. *Biolink Model*. https://github.com/biolink/biolink-model.

———. 2021b. *KG-Hub*. LBNL. https://knowledge-graph-hub.github.io/.

Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, et al. 2007. “The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration.” *Nature Biotechnology* 25 (11): 1251–55.

Voss, Erica A., Rupa Makadia, Amy Matcho, Qianli Ma, Chris Knoll, Martijn Schuemie, Frank J. DeFalco, Ajit Londhe, Vivienne Zhu, and Patrick B. Ryan. 2015. “Feasibility and Utility of Applications of the Common Data Model to Multiple, Disparate Observational Health Databases.” *Journal of the American Medical Informatics Association: JAMIA* 22 (3): 553–64.

Whetzel, Patricia L., Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. 2011. “BioPortal: Enhanced Functionality via New Web Services from the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications.” *Nucleic Acids Research* 39 (Web Server issue): W541–45.