How to teach a computer to learn about microbes: KG-COVID-19 and microbial graph learning - Abstract

Marcin P. Joachimiak, PhD
Environmental Genomics and Systems Biology Division
Lawrence Berkeley National Laboratory

Abstract

After many decades of research, we have become fairly skilled at modeling and predicting some aspects of simpler model organisms, such as HIV, E. coli, or S. cerevisiae. However, the COVID-19 pandemic has shown that this is far from true for novel species. In fact, even for model organisms the data and tools required to build highly collaborative and integrated models are still being developed. Although we cannot make reliable predictions for most biological problems, in many cases we have large collections of data with different modalities that have not been interlinked or fully exploited. These data are often siloed by namespaces, APIs, and formats. They are also affected by data modeling and analysis choices. Based on a core set of design principles, we have developed a framework for constructing knowledge graphs which allows to harmonize biological entities and their relationships across many disparate data sources. We applied this framework to COVID-19 as well as environmental genomics projects. Using new graph learning methods we can leverage complex data to compute similarities between different biological entities, something that has been difficult thus far. We are also applying graph embeddings to perform link prediction tasks, focusing on target identification and drug repurposing for COVID-19. While large knowledge graphs and associated methods are exciting developments, their complexity requires new tools including for data search, introspection, and visualization. Advances in these areas will be critical to achieving explainability for both more traditional and new learning methods.