NML Science Story

Spfy: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses. Le KK, Whiteside MD, Hopkins JE, Gannon VPJ, Laing CR. 2018. Database. Available at: 10.1093/database/bay086

What was known about this area prior to your work, and why was the research done?

The NML develops whole-genome sequence (WGS) based analyses to rapidly predict standard reference laboratory results, and offers integrated bioinformatics pipelines for researchers. A logical extension is to integrate results from different analysis software, in a meaningful manner, for population-scale studies. Current projects, such as the Genomic Epidemiology Ontology (GenEpiO) [1] project out of Canada, standardizes genomics terminology into a knowledge graph relating results from different analysis software results and metadata provided by researchers. Our project, named Spfy, implements a graph database to model bioinformatics results, and to perform statistical comparisons between population groups at-scale.

What are your most significant findings from this work?

Spfy allows users to upload WGS samples for different subtyping analyses, and then integrates results into a graph database for downstream comparisons. For example, Spfy can determine if a statistically significant difference exists, using Fisher's exact test, among any identified AMR genes, between all stored E.coli genomes of serotype O157, and genomes of serotype O26; where, AMR and serotype results are generated from different software modules. Our most significant findings are the design of graph structures for future extension, and enabling statistical comparisons between any and all datasets stored in the graph database.

A few of the technical findings were that:

1. Generic, directed edges allow results from new analyses modules to be integrated and all developed statistical comparisons to work without modification.
2. Finding the largest or smallest value for nodes of a given type is expensive and should be stored in a different datastore. For example, we index an incremental key ("spfyid1", "spfyid2", ...) outside of the graph database.
3. Performance-wise, we see a linear relation between the number of nodes retrieved and the analysis time. For example, 1.5 million nodes/attributes can be compared in approximately 90 seconds, and 2 million nodes/attributes in approximately 110 seconds.

What are the implications or impact of the research?

Our project demonstrates how graph databases may be used as a future storage step for bioinformatics pipelines developed by the NML. In the past few years, the field of graph databases has seen rapid expansion with new players such as Dgraph [2], Cayley [3], and new graph-inspired query languages such as GraphQL [4], in additional to older graph databases such as Neo4j [5] and Blazegraph [6]. Graph storage would provide the NML's monitoring and surveillance efforts with a common platform for identifying differences between sample populations, such as different locations of origin, host species, or subtypes.

Additional References of Significance

---

1. https://github.com/genepio/genepio
2. https://github.com/dgraph-io/dgraph
3. https://github.com/cayleygraph/cayley
4. https://github.com/facebook/graphql
5. https://github.com/neo4j/neo4j
6. https://github.com/blazegraph/database