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. The purpose of this project, named Spfy, was to explore new long-term data storage methods that can integrate results from different pipelines, and support downstream hypothesis testing on the pre-computed results. This would allow researchers to take a population group identified by one software module and find statistical differences in the corresponding results from another software module. 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.

What are your most significant findings from this work?

Spfy is provided as a web interface at https://lfz.corefacility.ca/superphy/spfy/ and implements a graph database for long-term data storage. The website combines a pre-computed dataset of 10,243 Escherichia coli genomes, for which in-silico serotype and Shiga-toxin subtype, as well as the presence of known virulence factors and antimicrobial resistance determinants have been computed, with the option to submit new samples. For downstream analyses, any data-type or relation in the graph is a valid option for analysis. This means that genomes can be compared on the basis of the presence or absence of pan-genome regions, serotype, subtyping data, or provided metadata such as location or host-source. All results are displayed to users in real-time, usually within 2-3 minutes.

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 [1], Cayley [2], and new graph-inspired

query languages such as GraphQL [3], in additional to older graph databases such as Neo4j [4] and Blazegraph [5]. Graph storage would provide the NML's monitoring and surveillance efforts with a common platform for integrating results from different pipelines and identifying differences between sample populations, such as different locations of origin, host species, or subtypes.

Additional References of Significance

---

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