Spfy: bringing real-time, big data batch analyses of E. coli to SuperPhy

Kevin K Le [1,*], Matthew D Whiteside [1], James Hopkins [1], Victor PJ Gannon [1], and Chad R Laing [1]

[1]National Microbiology Laboratory at Lethbridge, Public Health Agency of Canada, Twp Rd 9-1, Lethbridge, AB, T1J 3Z4, Canada

Tel: +1 403-382-5516; Fax: +1 403-381-1202; Email: chad.laing@canada.ca

This website is free and open to all users and there is no login requirement. The code for this webserver, and all third party software used, are available under the open-source Apache 2.0, BSD 3-clause, or similar licenses.

The website is available at `https://lfz.corefacility.ca/superphy/spfy/`. Spfys code is provided at `https://github.com/superphy/backend` and documentation at `https://superphy.readthedocs.io/en/latest/`.

Our proposal covers an update to Superphy (1), an online predictive genomics platform targeting Escherichia coli. The update, called Spfy, uses graph data structures to store and retrieve results for computational workflows, facilitating the management and querying of tens of thousands of whole-genome E. coli sequences, and efficient downstream processing. Current comparative computational workflows chain different analysis software, but lack storage and retrieval methods for generated results. By making the storage and retrieval of results part of the platform, with data effectively linked to the organisms of interest through a standardized ontology, we can mitigate the recomputing of analyses. Within Spfy, the output from all analyses is stored, and linked together in the context of a genome graph. This graph also stores metadata for each genome, facilitating inquiries ranging from population genomics to epidemiological investigations. Integrated data storage will be necessary as publicly available whole genome sequencing data for bacterial pathogens currently numbers in the tens of thousands, with hundreds of thousands set to be available within the next few years.

Spfy was tested with 4,622 public E. coli assembled genomes from Enterobase, storing every sequence and results for all included analysis modules. Spfy provides real-time subtyping, and the results are immediately displayed to the user following their completion. Subtyping options include O-antigen, H- antigen, Shiga-toxin 1, Shiga-toxin 2, and Intimin typing. Reference-lab tests include virulence factor and anti-microbial resistance annotation. All genomes are analyzed within the pan-genome framework of E. coli, and results from all analyses are automatically associated with the source genome. The resulting database had 1,333 nodes and 683,666 leaves, with 374,836,872 object properties.

Existing scientific workflow technologies such as Galaxy (2), and pipelines such as the Bacterium Analysis Pipeline (BAP) (3) and the Integrated Rapid Infectious Disease Analysis (IRIDA) platform `http://www.irida.ca/` help automate the use of WGS data for public-health surveillance. Like IRIDA and BAP, Spfy automates workflows for users, and like Galaxy, Spfy uses task queues to distribute selected analysis. File uploads begin through the ReactJS- based website, where user-defined analyses options are selected. To these concepts we add Docker containerization for task queue workers, allowing analysis software to safely run in parallel. For result storage, existing workflow technologies use relational tables (2), or store resulting files to disk (3). Because output from these programs is user-specific or transitory, results from identical comparisons are often recomputed. Additionally, output from different analyses are structured using distinct terminology and formats, which must be converted before they can be compared. Without a unified structure, these conversions quickly become impractical for broad usage. Graph-based storage of all results solves these problems. To avoid proliferating ontologies, and to allow Spfy to integrate with existing ones, annotations from the GenEpiO (4), FALDO (5), and TypOn (6) ontologies are used to describe biological data. The entire platform is packaged using Docker-Compose, and can be recreated with a simple command.

The Spfy update has been up since May 2017 and Superphy has been up since early 2016. The server accepts assembled E. coli genomes with the .fasta or .fna extensions. Submissions are checked against a reference set of E. coli gene sequences before running analyses. Outputs are displayed on the website in tables and can be downloaded as .csv files.

Keywords: Comparative genomics analysis, Epidemiology, Microbial genomics, Graph database

REFERENCES

1. Whiteside, M.D., Laing, C.R., Manji, A., Kruczkiewicz, P., Taboada, E.N., and Gannon, V.P.J. (2016) SuperPhy: predictive genomics for the bacterial pathogen Escherichia coli. *BMC Microbiol*, **16**, 65.

2. Goecks, J., Nekrutenko, A., Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, **11**, R86.

3. Thomsen, M.C.F., Ahrenfeldt, J., Cisneros, J.L.B., Jurtz, V., Larsen, M.V., Hasman, H., Aarestrup, F.M., and Lund, O. (2016) A bacterial analysis platform: an integrated system for analysing bacterial whole genome sequencing data for clinical diagnostics and surveillance. *PloS One*, **11**, e0157718.

4. Griffiths, E., Dooley, D., Graham, M., Van Domselaar, G., Brinkman, F.S.L., and Hsiao, W.W.L. (2017) Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance *Front Microbiol*, **8**, 1068.

5. Bolleman, J.T., Mungall, C.J., Strozzi, F., Baran, J., Dumontier, M., Bonnal, R.J.P., Buels, R., Hoehndorf, R., Fujisawa, T., Katayama, T., et al. (2016) FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. *BMC J Biomed Sem*, **7**, 1068.

6. Vaz, C., Francisco, A.P., Silva, M., Jolley, K.A., Bray, J.E., Puseele, H., Rothganger, J., Ramirez, M., Carriço, J.A. (2014) TypOn: the microbial typing ontology *BMC J Biomed Sem*, **5**, 43.