Genesearch: A Sequence Similarity Search Service for Genomics Workflows

Jeffrey Johnson[2*], Richard Shane Canon[1], I-Min A. Chen[1], Ken Chu[1], Paramvir Dehal[1], David Lyon[1], Torben Nielsen,[1] Hugh Salamon[1], Elisha Wood-Charlson[1]
Kjiersten Fagnan[1*]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Cohere Consulting, LLC, Seattle, WA
https://code.jgi.doe.gov/jgi-kbase/genesearch

Project Goals: Deploy a suite of composable services to support dynamic genomics data analysis for JGI, NMDC, and KBase Users (Limit to 1000 characters)

Abstract text. Please limit to 2 pages.
As genomics datasets continue to grow precipitously, it has become more difficult for researchers to incorporate new sequences into their workflows in a consistent and reproducible way. One approach to addressing this difficulty is to create a set of "primitive" software services from which one can construct reliable research workflows. Sequence similarity search is a ubiquitous element in these workflows and an obvious candidate for a reusable and composable service. In fact, many institutions already offer such services through web interfaces ([1], [2]).

JGI and KBase both aim to help DOE scientists by allowing them to measure and analyze their data in new ways. JGI and KBase share the need to express increasingly sophisticated relationships in biological data, such as phylogenetic, chemical and environmental similarities, expressed in terms of taxonomy, gene homology, chemical similarity, etc. This common goal requires significant computing resources that increase as we add more organisms, interactions, and environmental parameters to our databases. Focusing first on our users' zero order data analysis use cases, we have identified core pieces of the JGI and KBase infrastructures that can be unified and shared.

JGI and KBase have created Genesearch, a service that provides a similarity search capability with a selection of alignment search tools and databases. Genesearch is structured as a microservice[3] in the sense that it provides exactly one function and is easy to deploy and maintain. Genesearch is currently deployed and available to DOE researchers within KBase and the Joint Genome Institute, and is also available as open source software for research groups that use their own sequence databases. The software can run normally or in a Docker container, and can be accessed by web clients and through a Python interface.

Genesearch is only the first in a suite of composable services for analyzing and manipulating large datasets on DOE and non-DOE resources. We are also working on a service that maps sequence identifiers between databases used by KBase, JGI, UniProt, and NCBI. Here we hope

to solicit input from the research community and to illustrate how these reusable elements can allow scientists to overcome the data deluge problem, build confidence in their work, and focus on answering previously intractable questions.

References

1. NCBI Web BLAST, https://blast.ncbi.nlm.nih.gov/Blast.cgi (retrieved 1/25/2021)
2. Mirdita M, Steinegger M and Soeding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics, doi: 10.1093/bioinformatics/bty1057 (2019)
3. Newman S, Building Microservices. O'Reilly Media, Inc, Sebastopol, CA (2015)

Funding statement.

Notes on abstract:

• Note the placement of superscripts in the authors and affiliations.

• URL above should be specific to the project. More than one URL is permitted.

• **References** can be **Publications** instead, if needed. Use any common style for these citations.

Genesearch: A sequence similarity search service for genomics workflows

As genomics datasets continue to grow precipitously, it has become more difficult for researchers to incorporate new sequences into their workflows in a consistent and reproducible way. One approach to addressing this difficulty is to create a set of “primitive” software services from which one can construct reliable research workflows. Sequence similarity search is a ubiquitous element in these workflows and an obvious candidate for a reusable and composable service. In fact, many institutions already offer such services through web interfaces.

We have created Genesearch, a service that provides a similarity search capability with a selection of alignment search tools and databases. Genesearch is structured as a microservice

in the sense that it provides exactly one function and is easy to deploy and maintain. Genesearch is currently deployed and available to DOE researchers within KBase and the Joint Genome Institute, and is also available as open source software for research groups that use their own sequence databases. The software can run normally or in a Docker container, and can be accessed by web clients and through a Python interface.

Genesearch is only the first in a suite of composable services for analyzing and manipulating large datasets on DOE and non-DOE resources. We are also working on a service that maps sequence identifiers between databases used by KBase, JGI, UniProt, and NCBI. Here we hope to solicit input from the research community and to illustrate how these reusable elements can allow scientists to overcome the data deluge problem, build confidence in their work, and focus on answering previously intractable questions.