Building decentralized indexes for public genomic data

Lab for Data Intensive Biology

Luiz Carlos Irber Júnior

lcirberjr@ucdavis.edu

Department of Population Health and Reproduction, University of California, Davis, USA

@luizirber

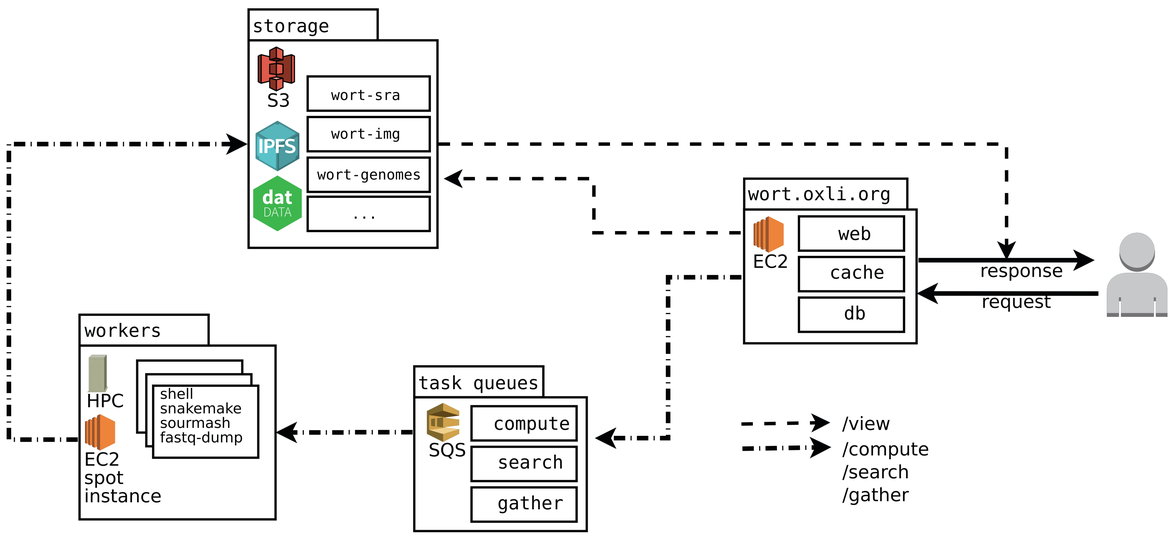
luizirber/2018-biods

Introduction

Traditional **algorithms** for **searching** genomic sequences **don't scale well** for the **large public genomic databases** currently available. With data archives like the SRA reaching **multiple petabases of data** and with **decreasing costs for sequencing** the rate of data generation is only increasing, and new computational methods are necessary. **Sequence Bloom Trees** [Solomon, 2016] allow searching for specific sequences in large databases, opening a path to solve other problems with similar data structures.

On top of the data discoverability problem, data access is also another issue. Downloading and sharing large collections don't take advantage of subsets of the data available outside the central servers hosting the full data, and traditional solutions involve explicit caches that need to be maintained and updated frequently.

Current architecture



Future Work

The current architecture is a proof of concept, with a **concrete, then abstract** approach: have something working first with a public API, then **refactor** and **generalize**. While it is deployed on AWS it can also be **run in other cloud providers**, and the next goal is to replace most of the **task queue and communication** with **P2P technologies**, using more of IPFS and dat (and not only as file storage).

The **WebAssembly** support in **sourmash** also allows doing more data processing in the browser, instead of transferring large datasets to the server. Currently wort is more focused on the API and command line usage, but more functionality will be added to the web frontend.

NCBI provides an **alpha feature** based on **STAT** to report the **taxonomic composition of reads** within a sequencing run. This analysis can also be done with **sourmash gather**, and a **browser extension** can overlay these results in the SRA Run Browser. This extension idea is similar to what the **BioJupies** project [Torres, 2018] does for RNA-Seq datasets.

Any **public database** can be store and queried using **wort**, and we intend to add more over time.

sourmash signatures

A **sourmash signature** is a **collection of MinHash sketches** from the **same original dataset** but with **distinct parameters**. A **MinHash sketch** [Broder, 1997] is a data structure that allows **fast similarity and containment estimates** for datasets, first used by **Mash** [Ondov, 2016] for genomic data. In **sourmash** we **extended MinHash sketches** to take the **complexity of the dataset** into account and make it possible to **compare** from **microbial genomes** to **metagenomes**. These signatures are most useful when they are easily available, especially for public genomic databases like RefSeq or the SRA.

sourmash also includes a **Sequence Bloom Tree** implementation for **indexing MinHashes**, allowing both **searching for similarity** and **taxonomic classification** of datasets. But there is no functionality to make it easier to **share** these indexes, and it is up to the user to send these sketches around or **build their own indexes**.

wort

wort is a **database** for **sourmash signatures**, providing APIs to allow **searching** public genomic databases for **dataset similarity**, performing **taxonomic classification** of samples and **submission** of signatures for **inclusion** in **search indexes**.

wort also aims to explore **data locality** on networks to overcome the **bandwidth limitations** in **centralized databases**, and move **data preprocessing** into **web browsers** using **WebAssembly** to allow novel solutions where the **user data doesn't need to leave their computers** to be able to participate in the ecosystem.

The main goal for wort is to create a resource that allows **permissive improvements** without **depending on central coordination**, using **decentralized web technologies** like **IPFS** [Benet, 2014] and **dat** [Ogden, 2017] as building blocks.

Currently there are **686k microbial SRA runs** and **65k JGI-IMG datasets** available for download in wort. They can be downloaded using the **public HTTP API**, via **IPFS** or **dat**. We are also moving the **sourmash prepared indexes** to wort, providing **RefSeq** and **GenBank** signatures too.

References

Benet, Juan. 2014. "**IPFS - Content Addressed, Versioned, P2P File System**." arXiv:1407.3561 [Cs], July. http://arxiv.org/abs/1407.3561.

Broder, Andrei Z. 1997. "**On the Resemblance and Containment of Documents.**" In Compression and Complexity of Sequences 1997. Proceedings, 21–29. IEEE. http://ieeexplore.ieee.org/abstract/document/666900/.

Ogden, Maxwell. n.d. "**Dat - Distributed Dataset Synchronization And Versioning.**". doi:10.31219/osf.io/nsv2c.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. "**Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.**" Genome Biology 17: 132. doi:10.1186/s13059-016-0997-x.

Solomon, Brad, and Carl Kingsford. 2016. "**Fast Search of Thousands of Short-Read Sequencing Experiments**." Nature Biotechnology 34 (3): 300–302. doi:10.1038/nbt.3442.

Torre, Denis, Alexander Lachmann, and Avi Ma'ayan. 2018. "**BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud**." BioRxiv, June, 352476. doi:10.1101/352476.

Titus Brown, C., and Luiz Irber. 2016. "**sourmash: A Library for MinHash Sketching of DNA**." The Journal of Open Source Software 1 (5). doi:10.21105/joss.00027.