A Knowledge Base for Running Distributed Science Applications

Joseph Phillips[1,2], Claire Ruttencutter[1], Laura Burdick[1], Steve Whitney[2], Windsor Aguirre[1]
[1]DePaul University, [2]Applied Philosophy of Science

1 Introduction

Scientists have a wealth of analysis tools. This is because small Python programs can attack interesting problems, and because GitHub serves as the universal location where scientist-developers distribute their code.

However, not all scientists have savvy to use these tools. Potential users must install the necessary software. And, if operating systems differ, they must administer virtual machines.

Our knowledge base web application addresses this. Beyond serving as a web interface between users and command-line driven software, the system organizes and helps interpret output. The system is:

- Distributed. Requests are sent to machines (physical or virtual) without users knowing or caring how.
- Multi-functional. The system is designed as a general interface to many specialized applications. The output of one program may be used as input to another.
- Customizable. We handle disagreements about the organization of ontologies by allowing users to pick one or customize their own.
- Uniform. Users see one interface for all the software. This leverages and extends the user-base of existing software and APIs.

The system will be used by undergraduate biologists at DePaul University to assay the distribution of aquatic organisms around Chicago. Environmental DNA (eDNA) is used to track the spread of invasive species, to track changes in the distribution of rare and endangered species, and to better grasp the structure of aquatic communities.

2 Approach

Since at least 2001, biologists have suggested collecting eDNA samples to assess ecosystems[1]. Mitochondrial gene "Cytochrome c oxidase subunit 1" (CO1) is the most commonly used gene to identify species through DNA barcoding[2].

Barcoding is facilitated by inexpensive DNA analysis tools and techniques. Among them is Oxford Nanopore, which sells modestly priced sequencers. Nanopore freely distributes its software[3], and third-party programs are available on GitHub. This, however, is often difficult to use for bench scientists lacking bioinformatic skills.

After processing freshwater eDNA samples, students sequence and analyze them. This typically results in tens to hundreds of files, each of which is a little more than 1MB and includes one thousand sequences. Sequences of the CO1 are then found, filtered for quality, aligned, clustered, and identified using existing software and APIs such as BLAST hosted by the NIH[4]. Our tool is a knowledge-aware interface for these programs and APIs, and allows bench scientists and undergraduates to analyze thousands to millions of sequences.

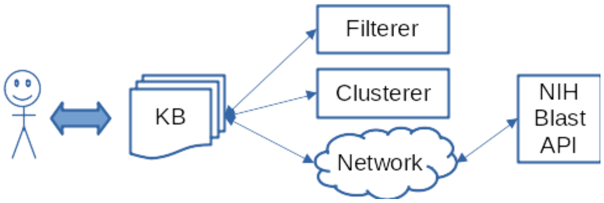

Figure 1: The kb facilitates uniform access to programs and APIs

3 Results and Conclusion

Our system will be used by biology students in the 2024-2025 academic year to analyze their eDNA samples. Several pipelines exist on GitHub to analyze DNA. However, the students are beginner biologists, and very few have any familiarity with running Python programs from the Unix command line.

4 Collaboration

Our team includes a biologist, computer scientists, and one computer scientist with a background in chemistry. Our system uses modern software technologies but is designed for biologists. And, although designed for novices, it is architecturally general enough to be extended for use to full-time, practicing scientists.

Bibliography

[1] Andreasen, J.K; O'Neill, R.V.; Noss, R; Slosser, N.C. "Considerations for the development of a terrestrial index of ecological integrity" *Ecol. Indic.*, 1 (2001).

[2] Hebert, Paul D. N.; Cywinska, Alina; Ball, Shelley L.; deWaard, Jeremy R. (2003-02-07). "Biological identifications through DNA barcodes". *Proc. of the Royal Society B: Biological Sciences.* 270 (1512): 313–321.

[3] Nanopore, https://nanoporetech.com.

[4] Altschul, Stephen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David J. "Basic local alignment search tool". *J. Molecular Biology*. 215(3): 403-410 (1990).