

Project Goals

From this project's `gene_finder` and `nitrogenase_finder` code we hoped to learn about the internal functions of the tools such as BLAST and Genius commonly used for bioinformatics analysis. From the visualization we hoped to learn how to portray the information in an accessible way using `pygame`.

Project Evolution

We started with a gene finder code from earlier this semester that was a gene finding Python program that accurately determined regions of the *Salmonella* bacterium's protein coding DNA. Building off functions in that code, we were able to look through bits of DNA for a nitrogen fixing gene. From there, we were able to implement a levenshtein algorithm that could account for the genes that don't exactly match for the *nifH* nitrogenase but still fix nitrogen. Our nitrogenase finder code was then able to pass important information such as the start and end index of the gene in the open reading frame, the length of the gene, percentage match with nitrogenase, as well as a whether or not the open reading frame is a reverse complement. This information could then be dumped into a pickle file to be accessed by our data visualization program.

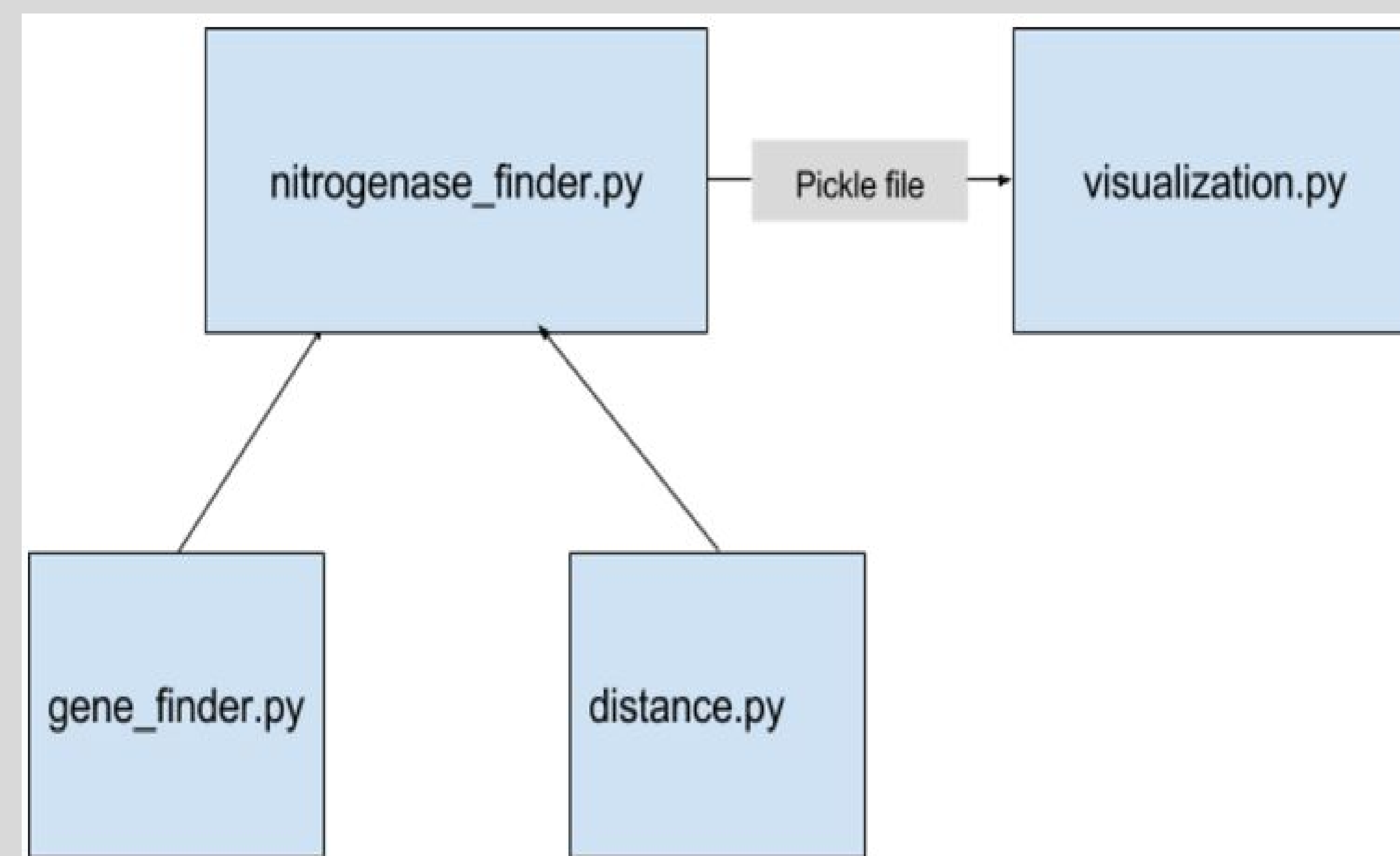
Nitrogenase Finder

Erica Lee, Liv Kelley, Rebecca Gettys

Nitrogenase Finder is a genetics tool designed to help find nitrogenase genes within a metagenome.

Implementation

This code takes a list of contigs that are gathered from a microbial community and looks for genes that can fix nitrogen. Since a gene that fixes nitrogen doesn't have to be an exact match for the nitrogenase sequence, we are using a levenshtein algorithm that finds a percent match for nitrogenase and passes it and other information of every gene that has a possibility of being nitrogenase into a pickle file that is then read by our data visualization code to produce a visualization for that gene.



The visualization code shown here depicts contigs as bars that are dark blue forward strands and light blue reverse complement strands. The orfs drawn are light gray if not matches and green if matches. Brighter green indicates a stronger match.

Description

Nitrogenase Finder is a genetics tool designed to help find nitrogenase genes within a metagenome. For this process, both the gene and metagenome should be in FASTA format. This project was originally inspired by Jean Huang, microbial biology professor at Olin College, who suggested that we make code for her research. We decided to take on the project because we wanted to understand her research project and the kinds of computing that will allow it to be successful.

Next Steps

As of right now, it can take at least 6 hours to finish running for an entire list of contigs from a microbial community. Even when we use PyPy, it takes about 15 minutes to produce results for one metagenome. Most of this time is spent searching in strings. Thus, we may implement a variation of the Boyer-Moore algorithm in the future for string matching in less time.

Resources

- Recursive Levenshtein Distance Example - <https://programmingpraxis.com/2014/09/12/levenshtein-distance/>
- Softdes professors and NINJAs
- Jean Huang
- Starter code for GeneFinder softdes project