Nitrogenase Finder

Erica Lee, Liv Kelley, Rebecca Gettys

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

Project Goals

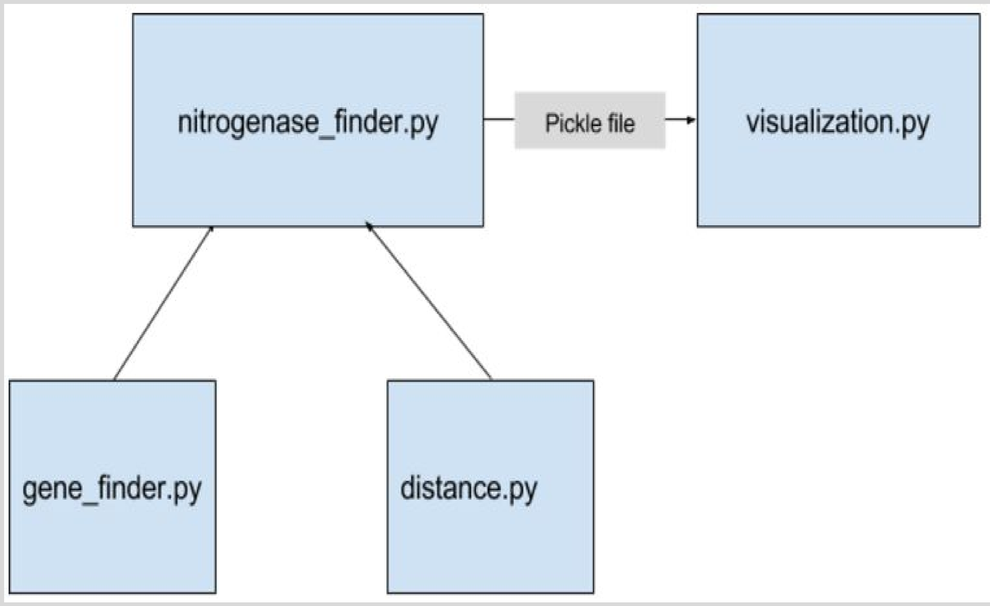
This project was originally born out of a suggestion by Jean Huang, a microbial biology professor at Olin College. We pursued this so that we could learn about the internal functions of the tools such as BLAST and Genius that we used for bioinformatics analysis.

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 DNA that code for proteins. 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.

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.



Description

Nitrogenase Finder is a genetics tool designed to help find nitrogenase genes within a metagenome; both the gene and metagenome should be in FASTA format. This project was originally born out of a suggestion by Jean Huang, a microbial biology professor at Olin College. We pursued this so that we could learn about the internal functions of the tools such as BLAST and Genius that we used for bioinformatics analysis.

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 contig. 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