BIOL 812 Final Assignment - Report

Data source: https://www.ncbi.nlm.nih.gov/gene?term=nifH%5BGene%20Name%5D

GitHub: https://github.com/dquesnelle/812-Final-Assignment

Group 1 - Isabella Asselstine, Xinran Li, Aakanx Panchal, Dan Quesnelle

All authors contributed to conceptualization, methodology, analysis, and writing of the final report. Dan Quesnelle was responsible for data curation, Aakanx Panchal was responsible for the multiple sequence alignment code and visualizations, Xinran Li was responsible for the distance matrix code and visualizations, and Isabella Asselstine was responsible for the phylogenetic tree code and visualization.

Introduction

Nitrogen fixation is the process in which atmospheric nitrogen is converted into ammonia or other nitrogenous compounds in order to increase its reactive potential. Without the ability to fix nitrogen, the biosynthesis of nitrogen-containing macromolecule building blocks such as amino acids and nucleic acids would not be possible. This is because nitrogen in its atmospheric form is extremely stable and thus is largely unable to react with other elements to form compounds. Given the importance of nitrogen-containing compounds in the biosynthesis of plants, animals, and other organisms, the process of nitrogen fixation is essential to life.

Though nitrogen fixation can occur through inorganic means, the majority of nitrogen fixation is made possible by enzymes called nitrogenases. These enzymes are found in certain bacterial species and work by combining free nitrogen with other elements to form ammonia, nitrates, and nitrites, which may then be used in further reactions. Nitrogen-fixing bacteria often associate with plant roots where they are able to provide the plant with metabolically active nitrogen. These nitrogen-fixing microorganisms play an important role in nitrogen cycling as they are responsible for over 90% of all nitrogen fixation on earth (1).

Nitrogenases are encoded by a variety of *nif* genes. One such gene is *nifH*, which codes for one of the three nitrogenase subunits called dinitrogenase reductase. *nifH* is the choice genetic marker to search for when identifying nitrogen-fixing microorganisms. In our project, we will be looking at the alignment of *nifH* genes between a variety of bacterial species. Learning more about the conservation of *nifH* sequences is important when studying how *nifH* genes differ between bacterial species.

In our project, we wanted to explore two questions. Firstly, What interesting insights can we glean from an alignment, distance matrix, and phylogeny of 100 *nifH* gene sequences? Secondly, are there non-rhizobium *nifH* sequences that are closer to the *nifH* sequence of the model rhizobium *Sinorhizobium meliloti* 2011 than the *nifH* sequences of other rhizobia? To address these questions, we will look at the *nifH* gene from 100 bacterial species and compare their relatedness via a multiple sequence alignment, distance matrix, and a phylogenetic tree.

Methods

nifH gene sequences from 100 different species of bacteria and archaea were collected from the NCBI Gene database. The gene IDs of the 100 gene sequences were compiled into a table along with other identifiers like gene and species name. To isolate the gene IDs from the other components of the table, we ran the file through a script that returns only the gene IDs in numeric form. It's in this form that we can use these gene IDs to retrieve the actual sequences. We took advantage of a command-line tool called Entrez Direct to retrieve our sequences. Entrez Direct is part of a larger tool called Entrez that is used for searching and fetching information from NCBI, and is one of the preferred methods for retrieval of genomic, coding and protein sequences [5]. We ran the numeric array of gene IDs through an Entrez script that searched NCBI for each gene sequence, converted them to FASTA format and then outputted them to a file that would end up containing all 100 sequences.

The gene sequences were retrieved, converted into fasta format, and then aligned using the msa package for multiple sequence alignment in R. The output of the msa function used to align our sequences was of the class MsaDNAMultipleAlignment, which is it why it was then converted to the fasta format before fed into further analyses. MView by Nigel Brown (2) was used to visualize our alignment, producing figure1.html and figure2.html; it added annotations including sequence sources (i.e., "Genus species strain"), percent coverage, percent identity, and colouring by properties such as mismatches relative to the reference sequence.

Before the visualization, we would like to remove sequences with large gaps to avoid numerical errors in later calculations. This was done through a Python script using a customized function named ratio_test. This script removed any sequences consisting of more than 60% gaps and generated a fasta (DMinput.fasta) file containing the remaining aligned sequences.

The preprocessed DMinput.fasta was then loaded as a DNAbin class through readDNAMultipleAlignment and as.DNAbin functions in R. This was followed by a transformation into a distance matrix. The distance

matrix can be calculated using different models targeting different aspects of the data with different assumptions. The models we explored in this project were TS, TV, indel, JC69, K80, K81, and TN93. The tile plots for distance matrices were grouped by models and saved as "DistMat(TS&TV).pdf", "DistMat(indel).pdf", and "DistMat(JC69&K80&K81&TN93).pdf".

We then wanted to create a phylogenetic tree to show the relatedness of our chosen bacterial species based on their *nifH* genes. To do this we took the distance matrix K80 that was previously calculated to create the tree. We used the ggtree package and created a circular tree with the bacterial genus and species names as labels. This was saved as "Phylogeny.pdf".

Results

Based on the comparison between number of transitions and transversions (DistMat(TS&TV).pdf), among sequences 60 to 79, transversions were more common than transitions, while for the rest of the sequences, it was quite the opposite. The number of indels (DistMat(indel).pdf) indicated that sequence 81, of *Methanococcus maripaludis*, and sequence 83, of *Methanosarcina mazei*, had significantly more insertion/deletion gaps, and therefore, they were expected to be associated with high values in distance matrices. In terms of distance matrices, we used four different models for calculation, resulting in four different tile plots. In the order of JC69, K80, K81, and TN93, as assumptions were gradually eliminated and the model became more generalized [4, 6, 7, 9], the distance among sequences were generally getting closer and closer in the DistMat(JC69&K80&K81&TN93).pdf. It was also worth mentioning that sequences 80 to 83 remained with relatively high values across all four matrices, which validated our assumptions based on the number of indels.

***Azospirillum brasilense* Sp7** Our alignment revealed some interesting results in terms of relatedness of bacterial species. As expected, the majority of the sequences with highest percent identities to our reference *nifH* sequence from *Sinorhizobium meliloti* 2011 were sequences that belonged to rhizobia bacteria. Interestingly, it was found that *Azospirillum brasilense* Sp7 displayed the closest non-rhizobium *nifH* sequence to our reference sequence, at a percent identity of 80.3%. In fact, *A. brasilense* showed higher percent identity to *S. meliloti* than other rhizobia such as *Rhizobium leguminosarum* (79.2%) and *Cupriavidus taiwanensis* (67.8%). This relation can be seen in the phylogenetic tree too, in which *A. brasilense* groups more closely to *S. meliloti* than other rhizobia which are branched further away. This result may suggest potential horizontal gene transfer of *nifH* from wild *S. meliloti* to *A. brasilense*. Because legumes (commonly associated with rhizobia like *S. meliloti*) and cereals (commonly found with plant growth promoting bacteria like *A. brasilense* in their soils) are often intercropped, there is overlap in the environments that both bacteria occupy. This could have catalyzed the transfer of genes including *nifH* from *S. meliloti* to *A. brasilense*.

***Vibrio natriegens* NBRC 15636** We also found that the *nifH* sequence of *Vibrio natriegens*, a Gram-negative marine bacterium, shared a 62.7% identity with that of *S. meliloti*. There is already interest in replacing *Escherichia coli* strains, used widely for molecular cloning and synthetic biology applications, with *V. natriegens*, as it has a growth rate twice as fast as that of *E. coli*. In addition to its fast growth rate, as nitrogen fixation ability in *V. natriegens* has been observed [3], it would also be worth investigating whether *V. natriegens* could be engineered to produce ammonia at rates comparable to those of implementations of the Haber-Bosch process, a process that requires high fossil fuel inputs for sufficient production. *V. natriegens* could bring us a more sustainable alternative to ammonia production, helping fertilize soils and assure global food security.

References

1. Britannica, The Editors of Encyclopaedia. "nitrogen fixation". Encyclopedia Britannica, 9 Jul. 2021, https://www.britannica.com/science/nitrogen-fixation. Accessed 17 April 2022.
2. Brown NP, Leroy C, Sander C. 1998. MView: A Web compatible database search or multiple alignment viewer. Bioinformatics 14 (4):380–381.
3. Coyer JA, Cabello-Pasini A, Swift H, Alberte RS. 1996. N_2 fixation in marine heterotrophic bacteria: dynamics of environmental and molecular regulation. Proc Natl Acad Sci U S A 93(8):3575-80.
4. Jukes TH and Cantor CR. 1969. Evolution of protein molecules. In Mammalian Protein Metabolism, ed. Munro, H. N., pp. 21–132, New York: Academic Press.
5. Kans J. 2013. Entrez Direct: E-utilities on the Unix Command Line. In: Entrez Programming Utilities Help, Bethesda (MD): National Center for Biotechnology Information.
6. Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111–120.
7. Kimura M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci U S A 78:454–458.
8. Lee HH, Ostrov N, Wong BG, Gold MA, Khalil AS, Church GM. 2019. Functional genomics of the rapidly replicating bacterium *Vibrio natriegens* by CRISPRi. Nat Microbiol 4:1105–1113.
9. Tamura K and Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10:512–526.