BIOL 145 / CS 185C

Spring 2023

Lab1

200 points

Follow the instructions, and answer all red questions by editing this docx. Upload your edited file to Canvas by 11:55PM on Feb 20.



In this lab you’ll reconstruct the famous phylogenetic tree that led Jon Zehr and his research group to discover UCYN-A. There’s a nice big copy of that tree on the next page.

Part 1: Look at the tree

JZ-GroupsAB.tiff

Zehr et al, 2001.

What is the scientific name of the OTU at the top? (Top of the diagram, not root of the tree.) Format correctly!

Anabaena variabilis

The accession number (a kind of i.d.) for that OTU is the part of the label beginning with “U”. What is the accession number?

U89346

Browse to <https://www.ncbi.nlm.nih.gov/>. Select “Nucleotide” search and enter the accession number into the search field. What is the title (in bold type, almost at the top of the record)? Since this record was in GenBank at the time the Zehr team blasted it, the record should be dated on or before the date of publication of the tree (2001). What is the date of the record? The scientific name of the species to which this sequence belongs appears as the first 2 words of the “DEFINITION” field. Does the name there match the name in the published tree?

# The title is Anabaena variabilis dinitrogenase reductase (nifH) gene, complete cds

**The date of the record is 15-AUG-1997**

**The name matches the name in the published tree.**

Part 2: The fasta file

Double-click on the file zehr\_nifhs\_INCOMPLETE.fasta. If you don’t see anything coherent, right-click and select “Open with…” Choose Notepad if you’re on Windows, choose TextEdit if you’re on a Mac. You should see something like this:



The file is in a format called “fasta”. This is the standard format for storing nucleotide and amino acid sequences. The lines beginning with “>” are called “deflines”. A single sequence appears after each defline. There are no restrictions (mostly) on what the defline may contain. There *are* restrictions on what the sequence may contain: only A, C, G, T, and line breaks. The deflines in this file contain the Accession Number of a record, followed by the record title. Does the first defline in your file look like that? Compare the first and last few nucleotides of the sequence in the file to the first and last few nucleotides of the sequence at GenBank. Do they look identical?

**The defline in my file looks like that.**

**They do look identical**

Scroll to the bottom of the fasta file. The last 5 sequences have been left for you to paste in.

For the first missing sequence (L15554.1), look up the nucleotide record at NCBI (as you did with the first record). Edit the defline in the fasta file so that it contains the entire record title (it ends “partial cds”). Find the nucleotide sequence and paste it into the fasta file, below the L15554.1 defline. How should you edit the sequence to make it compatible with the green sentence above? Make those edits and paste the entire record for that sequence (defline and nucleotides) here.

**I need to remove the leftover spaces and numbers as well as the slash.**

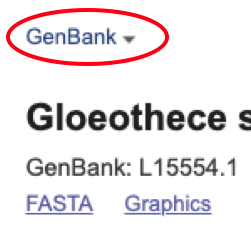
**>L15554.1 Gloeothece sp. PCC 6909 nitrogen fixation protein (nifH) gene, partial cds**

**TCTACTCGCTTGATCCTTAACTGTAAAGCGCATGTCACAGTTCTACACTTAGCCGCAGAACGGGGTTCTGTTGAAGATATAGAACTCGAAGATGTACTGCTCACAGGGTTTGAAGACATCAAATGCGTAGAATCAGGTGGTCCTGAACCTGGCGTAGGATGCGCTGGTCGTGGGATTATCACTGCCATCAACTTCCTTGAAAAGAAGGAGCTTACGAAGACATAGATTTCGTATCCTACGACGTATTAGGGGACGTTGTCTGCGGTGGTTTCGCTATGCCTATCCGTGAAGGAAAAGCACAAGAAATCTACATCGTAACCTCT**

Did you enjoy editing the sequence? Would you enjoy doing it for the remaining 4 records?

**I did not enjoy editing the sequence. I would not enjoy doing this for the remaining 4 records.**

There’s an easier way. Browse to the record for the next missing sequence. Find the tiny pull-down menu just above the record title:



In that menu, select “FASTA”. You’ll see a fasta record that is almost what you want in your fasta file. The only problem is that the defline wraps to a second line (on my machine, anyway). Paste the entire record (defline + sequence) from the browser into your fasta file. Make sure to delete the original defline in the file. Edit the new defline so that it’s all on a single line. Paste the record in your file here. Paste text, not a screenshot.

**>U22146.1 Synechococcus sp. nitrogenase MoFe protein beta-subunit (nifK), nitrogenase MoFe protein alpha-subunit (nifD) and nitrogenase reductase (nifH) genes, complete cds**

**ATGCGTCAAATTGCATTTTACGGAAAAGGCGGTATCGGTAAGTCTACCACTTCTCAAAACACCTTAGCCG**

**GAATGGCTCAAGCTGGCAACCGCATCATGATTGTTGGTTGTGACCCCAAAGCTGACTCTACCCGTTTGAT**

**CCTCAACTGTAAAGCTCAGGTAACGGTATTACACTTAGCTGCCGAACGTGGTGCTGTTGAAGATCTCGAA**

**CTCTCAGATGTATTACTCACTGGGTTTGAAAACATCAAGTGTGTTGAATCTGGTGGTCCCGAACCTGGGG**

**TTGGTTGTGCTGGACGTGGTATTATCACCTCCATCAACTTCCTTGAAGAAGAAGGTGCTTATGAAGATCT**

**AGACTTCGTATCCTATGACGTATTAGGAGACGTTGTTTGTGGTGGTTTCGCTATGCCTATCCGTGAAGGA**

**AAAGCACAAGAAATCTACATCGTTACCTCTGGGGAAATGATGGCGATGTATGCAGCTAACAACATCGCTC**

**GTGGGATCTTAAAATATGCCCACACTGGTGGTGTTCGTTTAGGTGGTTTAATTTGTAACAGCCGTAACGT**

**TAACAAAGAGATCGAATTGATCGAAGAGTTAGCTGAACGCTTAAACACCCAAATGATCCACTTCGTACCC**

**CGTTCCAAACAGGTACAAGAAGCTGAATTACGTCGTCAGACCGTTATTCAATACTCCCCTGAGCACCCCC**

**AAGCTCAAGAATACCGTGATTTAGGTGACAAGATCGTTAACAACACCAAACTCACCATCCCCACTCCTAT**

**CGACAACGACGAACTCGAAGAACTGTTGATCAACTACGGTTTACTTGGCTCAGAAGAAGAGTACAAGAAA**

**GTTATGGAAGCTGACATGGCTGCCCAAGCTTTAACCAGAGGCGCGAAGTAATGTAAGGGCTTAATTACCT**

**TAAGCCCCTACTATCGGGAAATGGGGGAGTTTAAGAAGGCAAAAGGCAAAAGGCAGAAGGCAACAGGGAC**

**AATCATTATCCCTAAGTCCTCACTCCTAACCCCTAACCCCTAACTCCCTTTACCCCCATCCTCATCAAAA**

**TTTCCTACTCGTCATTATTGACGAAGGGGGTGCGTCCGATCCTAATCGCCAATTCATCACTGAGGAACAC**

**TATGTCAACAGTAGAAGACAGAAAGCAGCTTATCCAAGACGTTCTTGATACCTATCCTGAGAAGTTAGCC**

**AAGAAACGGTCTAAACACCTCAATGTTTACGAAGAAGGCAAAGACGATTGTGGAGTAAAATCTAACATTA**

**AGTCTGCACCTGGTGTAATGACCGCTCGTGGTTGTGCTTATGCAGGATCTAAAGGGGTGGTTTGGGGTCC**

**TATCAAAGATATGATCCATATCTCCCACGGACCTGTTGGTTGCGGTTACTACTCTTGGTCTGGTCGTCGT**

**AACTATTACATCGGAACCACTGGGGTTGATACCTTTGGTACGATGAACTTTACCTCTGACTTCCAAGAAA**

**AAGACATCGTTTTTGGTGGAGACAAAAAACTCCTCAAAATCACCGAAGAAATCGAAGAATTATTCCCCCT**

**CCACAATGGGATTTCCATTCAGTCTGAATGTCCTGTTGGATTAATTGGGGATGACATCGAAGGTGTTGCC**

**AAAAAAGCGCAAAAAATTACTGGCAAACCCGTATTCCCGTTCCGTTGTGAAGGATTCCGTGGCGTTTCCC**

**AATCCTTAGGACACCACATCGCTAACGACGCAGTGCGTGACTGGGTATTTAGCCGTGATGATGCTCAAGA**

**AATCGAAACCACTCCCTATGATGTTGCCATCATTGGAGACTACAACATCGGTGGAGATGCTTGGTCTAGC**

**CGTATTCTTCTCGAAGAAATGGGTCTGCGCGTCGTTGCTCAATGGTCTGGAGACGGAACCATCAACGAAA**

**TGATGCAAACCCCCAAAGTGAAACTCAACCTGATTCACTGTTACCGTTCCATGAACTACATCAGTCGTCA**

**CATGGAAGAAAAATACGGTATTCCCTGGTTTGAGTACAACTTCTTTGGTCCTACCAAGATTGCTGAATCC**

**TTACGCGCGATCGCTGCTCTGTTTGATGACACCATCAAAGAAAATGCAGAGAAAGTAATTGCTAAGTACG**

**AACAACAAACCGCAGAAGTCTTAGCCAAATACCGTCCTCGTTTGGAAAACAAAACCGTCATGATGATGGT**

**GGGTGGACTACGTCCTCGTCACGTTGTTCCTGCTTTCACAGACTTAGGCATGAAAATGATCGGAACCGGA**

**TATGAGTTCGCTCACGGTGACGACTATAAACGTACCACTGAGTATGTTGATGATGCAACCCTCATCTATG**

**ATGACGTAACTGCCTACGAGTTCGAGAAATTCGTTCAAGAACTGAAACCCGACTTAGTTGCTTCTGGCGT**

**TAAAGAGAAGTATGTCTTCCAGAAAATGGGACTACCTTTCCGTCAAATGCACTCTTGGGATTACTCTGGT**

**CCTTACCACGGTTATGATGGGTTCGCTATCTTTGCACGGGATATGGACTTAGCTCTCAATAACCCGACCT**

**GGGGATTAATCAAATCTCCTTGGAATAAGTAAGAGGGAATTAGGGAATAGGGAACAGGGAACAGGGAATA**

**GTGAATTGGGAAGTCCTTCACTAACTCTCTAGTTACGAAGCCACCCCCTCACCCCCTCCCCCCCCTCATC**

**TCCTCACCTCCTCCCCCTTTCTCGACAACTGCATCATTAATCCTTGACCGAACAACGGAGTATCACGAAT**

**GTCTCAGAAAATTGATAAAATCCAAGACCACGTTGAGTTATTCCACCAACCAGAGTACCAAGAGCTATTT**

**GAAAACAAGAAAGCTCTCCAAGGAATGGCTTCTGATGAGAAAGTCGCTGAAATAGCCGAATGGACCAAAA**

**CCTGGGAATATCGGGAAAAGAACTTCGCTCGTGAAGCTCTGACCATCAACCCCGCTAAAGCTTGTCAACC**

**TTTGGGTGCTATCTTAGCTGCGGTTGGTTTTGAAGGAACCCTCCCCTTTGTGCATGGATCACAAGGTTGT**

**GTGGCTTACTTCCGTACCCACTTTACCCGTCACTTCAAAGAGCCTTTCAGTGGTGTTTCTTCTTCCATGA**

**CTGAAGATGCAGCCGTCTTCGGTGGACTGAAAAACATGATCGAAGGGTTACAGAATGCTTATAGTCTCTA**

**TCAACCCAAAATGATTGCTGTCTGTACAACTTGTATGGCAGAAGTTATTGGGGATGACTTAGGTTCCTTC**

**ATTGGCAATGCTAAGGCTGACGGTTCTGTTCCTAAAGATTTCCCCGTTCCCTTTGCTCACACTCCTTCTT**

**TCGTGGGTTCTCATATCACGGGATATGACAACATGATGAAAGCCATCCTGTTGAACTTAACCGACGGCAA**

**GAAACCTACCACCAGCAACGGTAAAGTTAACTTTATTCCTGGGTTTGAAACCTATGTTGGTAACCTACGC**

**GAACTGAAGCATTTAACCAGTGCTATGGGGGTTGATGCTACCATTTTAGGAGACAACGAACTCTATTTAG**

**ATTCTCCTAACGATGGCGAGTTCAAAATGTACCAAGGTGGTACTACTCTAGAAGAAGGTGCTGATGCTAT**

**CAATGCAACCAAAACCATCGCACTGCAAACCTATCCCACCGTTAAAACCCTCGAATACATCGAGAAAGAA**

**TGGCAGCAACCCACCGCTACCTATCGTCCTTGGGGTATTAAAGGAACGGATGAGTTCGTCATGGCTTTAT**

**CTGAACTCACTGGGAATCCTGTTCCTCCCGAATTGGAACTAGAACGGGGACGCGCAGTGGATGCTATGAC**

**CGATAGTCATGCTTGGTTACATGGTAAAAAAGCGGCTATCTATGGCGACCCTGACTTAGTCATGGGAATG**

**CTGCAATTCATGTTAGAGATGGGTGTTGAACCTGTTCACGTTTTGGTTCACAACTCTACCACTGAATTTG**

**AAGAAGAAGCCAAAGCTCTCTTAGCTTCTAGTCCTTATGGTCAAAAAGCCACCGTTTGGGGCGGTAAATA**

**CCTCTGGCACCTCCGTTCCTTACTGTTTACTGAACCTGTTGACTTCTTAATCGGGAATTCCTACGGTAAA**

**TACCTCTGGCGTGATACCAAGATTCCTTTAATCCGCATCGGGTATCCTATCTTTGATCGCCACCACTTAC**

**ATCGCTATTCTACCATTGGTTACAATGGCGCGATTAACCTGCTCAATTGGATTGTTAATGGTCTGTTTGA**

**AGAAATCGACCGCAACACCAATATCCCCTCGAAGACCGACATTTCCTTCGATTTAGTTCGTTAA**

Repeat what you just did for the remaining 3 missing sequences.

**>AF300829.1 Synechocystis sp. WH8501 dinitrogenase reductase (nifH) gene, partial cds**

**TCTACCCGTTTAATCCTCAACTGTAAAGCTCAGGTAACTGTATTACACTTAGCTGCTGAAATGGGTTCTG**

**TTGGAGACTTAGAACTCGAAGACGTAATGCTCGAAGGGTTTGAAGGCATCAAGTGTGTAGAATCTGGTGG**

**TCCTGAGCCTGGAGTTGGTTGTGCTGGTCGTGGTATTATCACCTCCATCAACTTCCTAGAAGAAGAAGGA**

**GCTTACGAAGACTTAGAATTCGTATCCTACGACGTATTAGGGGACGTTGTATGTGGTGGTTTCGCTATGC**

**CTATCCGTGAAGGAAAAGCACAAGAAATCTACATCGTTACCTCT**

**>AF299418.1 Unidentified marine bacterial clone HT1904 dinitrogenase reductase (nifH) mRNA, partial cds**

**TCTACCCGTTTAATCCTCAACTGTAAAGCACAGGTAACAGTTCTTCACTTAGCAGCAGAACAGGGTTCCG**

**TTGAAGACTTAGAACTCGAAGATGTATTGCTCGAAGGATTTGAAAACATCAAGTGTGTAGAATCTGGTGG**

**TCCTGAGCCTGGAGTTGGTTGTGCTGGTCGTGGTATTATCACCTCCATCAACTTCCTAGAAGAAGAAGGA**

**GCTTACGAAGACTTAGATTTCGTATCCTATGACGTATTAGGAGACGTTGTTTGTGGTGGTTTCGCAGTGC**

**CTATCCGTGAAGGAAAAGCACAAGAAATCTACATCGTTACCTCT**

**>AF016616.1 Unidentified bacterium clone AO11 dinitrogenase reductase (nifH) gene, partial cds**

**TCTACCCGTTTAATCCTTAACTGTAAAGCACAGGTAACTGTGTTACACTTAGCAGCAGAACAGGGTTCCG**

**TTGAAGACTTAGAACTCGAAGATGTATTGCTCGAAGGATTTGAAAACATCAAGTGTGTAGAATCTGGTGG**

**TCCTGAGCCTGGAGTTGGTTGTGCTGGTCGTGGTATTATCACTTCCATCAACTTCCTAGAAGAAGAAGGA**

**GCTTACGAAGATTTAGACTTCGTATCCTATGACGTATTTGGAGACGTTGTATGTGGTGGTTTCGCAAAGC**

**CTATCCGTGAAGGAAAAGCACAAGAAATCTACATCGTTACCTCT**

Now the name of your fasta file is no longer accurate. Save your work, close your editor, and rename your file so that it isn't misleading. When you rename, keep the “.fasta” extension. What file name did you choose?

**zehr\_nifhs\_COMPLETE.fasta**

Part 3: Align the fasta

In class you studied pairwise alignment (PWA); that’s what BLAST uses. Bioinformatics also uses multiple sequence alignment (MSA). In a way, a phylogenetic tree is a way of visualizing a multiple sequence alignment.

You have seen that BLAST is very fast. (If it seemed slow, you were probably waiting your turn on the computer at NCBI; that isn't the algorithm’s fault.) Since BLAST relies on PWA, we know that PWA must also be very fast. *MSA isn't fast!* In fact, for any reasonable number of sequences of any reasonable length, MSA would take longer than the life of the universe so far. So to do PWA we use “heuristics”. These are algorithms that execute in reasonable time and give answers that might not be exactly right but are usually very close.

A common heuristic algorithm for MSA is Clustal-Omega. Browse to its web site at <https://www.ebi.ac.uk/Tools/msa/clustalo/>. Paste a screenshot of the top 2 boxes (“STEP 1” and “STEP 2”).

Graphical user interface, text, application, email

Description automatically generated

If the top field says “PROTEIN”, change it to “DNA”. Click “Choose File” and select your fasta file. In STEP 2, for Output Format, select “Pearson/FASTA”. Then click the “Submit” button.

After (hopefully) a few seconds, you’ll see a results page that looks like your fasta file, but there will be indels in the sequences.

What is the accession number of the first sequence that only has indels before the first nucleotide and after the last nucleotide?

U73132.1

What is the accession number of the first sequence that has indels among the nucleotides?

AF003336.1

Can you tell, by looking at the MSA, which sequences are most closely related/similar to one another?

**If you are talking about the individual alignments that are matched, you can based on the quantity of indels between the first nucleotide and the last which as it increases means they are less related.**

**Otherwise, if comparing all the matched sequences to each other, you can't. There isn't much that an MSA can tell you. MSAs are mostly used as inputs to other algorithms, especially tree building algorithms.**

Click on the “Phylogenetic Tree” button. Paste a screenshot of the tree.

Diagram, schematic

Description automatically generated

What are the accession numbers of the first 3 sequences in the MSA? Are they the same as the first 3 OTUs in the Zehr tree?

**The accession numbers of the first 3 sequences in the MSA are U73132.1, L15550.1, U73137.1. They are NOT the same as the first 3 OTUs in the Zehr tree.**

Yeah, that’s inconvenient. Clustal (as well as other MSA algorithms) doesn’t order its output to match its input. Recall that there are many equivalent ways to draw a tree while correctly portraying the evolutionary relationships among the OTUs. Any particular rendering of a tree is like a mobile: if the breeze blows, your view of the sculpture changes, but it’s the same sculpture.

Mark Leary Designs

Fortunately, we’re really only interested in Group A. What are the accession numbers of the 7 Group A sequences? (They begin with “AF”, not “HT”.) Do they form a monophyletic group in the Clustal tree? Paste a screenshot that demonstrates your answer.

**The accession numbers of the 7 Group A sequences are:**

1. **AF059626.1**
2. **AF059634.1**
3. **AF059637.1**
4. **AF299425.1**
5. **AF059642.1**
6. **AF059640.1**
7. **AF299420.1**

**They form a monophyletic group in the Clustal tree because they can be separated from the tree with one cut.**

**Diagram

Description automatically generated with low confidence**

Part 3: Expand the tree

The original Zehr tree was constructed from 48 *nifH* OTUs, and the Group A clade consisted of only 7 OTUs. Those are small numbers, compared to the current number of known *nifH* sequences. A study published in 2014 data-mined GenBank and found thousands of *nifH* sequences. Read the abstract of the publication (you downloaded it with this assignment) and answer these questions:

* How many *nifH* sequences did that study find?

**The study found 34,420 nifH sequences.**

* What is a common use for the *nifH* gene?

**The nifH gene is commonly used for identifying organisms that can fix nitrogen.**

Given the large numbers of currently known *nifH* sequences, it’s possible that the “Group A” clade is an illusion stemming from the small sample size. Maybe there are UCYN-A sequences (discovered since 2001) that wouldn’t cluster with the original Group A OTUs. Conversely, maybe there are non-UCYN-A sequences that would cluster with the original Group A OTUs. Either way, UCYN-A would no longer be a monophyletic group. So … is UCYN-A monophyletic? A rigorous answer to that question would be a major science project, but you already have all the bioinformatic tools you need to gain a little insight. The steps are given on the next page. Before you peek, think about how you might gain a little more confidence that UCYN-A is monophyletic. Write your thoughts here.

**To start determining whether UCYN-A is monophyletic, I would do a Multiple Sequence alignment between the already determined UCYN-A group A and the new non-UCYN-A group. This would allow me to create a phylogenetic tree and see if any non-UCYN-A cluster with any from the UCYN-A group.**

Here’s how you could do it:

1. Collect some more UCYN-A *nifH* sequences and add them to your fasta file.
2. Collect some more *nifH* sequences that aren't UCYN-A, and add them to your fasta file.
3. Use Clustal-Omega to align the sequences.
4. Look at the phylogenetic tree.
5. See if UCYN-A is still monophyletic.

In detail:

Collect some more UCYN-A *nifH* sequences

Choose one of the *nifH* accession numbers from the “Group A” portion of the original tree. Remember, those accession numbers begin with “AF”. Which did you choose? Go to the blastn page at NCBI (<https://blast.ncbi.nlm.nih.gov/Blast.cgi> and select the “blastn” tab). Enter your accession number in the Query Sequence field. Since you only want subjects (hits) that are from UCYN-A, find the “Organism” field under “Choose Search Set”. Begin typing “Candidatus Atelocyanobacterium thalassa”. At some point the web page will complete the name for you, but make sure it does it right … it should end with “(taxid:713887)”. Now click the blue “BLAST” button and wait for the results. Paste a screenshot of the resulting table of hits.

Accession #: **AF059626.1**

Graphical user interface, text, application, email

Description automatically generated

Accession numbers for your subjects are at the right of the table. Choose any 5 accession numbers where:

* The description begins “Candidatus Atelocyanobacterium Thalassa”
* And the E-value is less than 1e-100
* And the subject length (shown in the “Acc len” field) is roughly 300-600

What accession numbers did you choose?

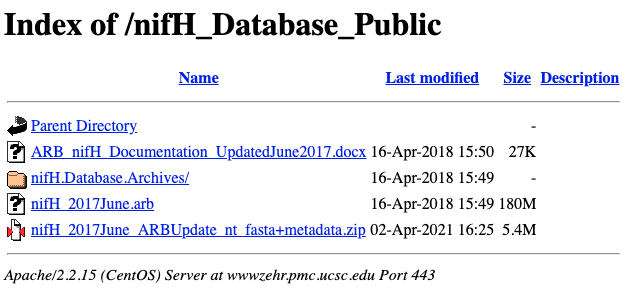
1. KF806612.1
2. KF806610.1
3. KF806607.1
4. KF806604.1
5. KF806611.1

Make a copy of your fasta file and rename it “extended\_zehr\_nifhs.fasta”.

Click on each accession number to see its GenBank page in a separate tab. As before, use the tiny pulldown near the top to show the sequence in fasta format (choose “FASTA”, not “FASTA text”). Paste your 5 new sequences, including their deflines, at the bottom of your extended\_zehr\_nifhs.fasta.

Collect some more *nifH* sequences that aren't UCYN-A

Browse to the latest rev of the *nifH* database at the Zehr lab’s web site: Go to <https://www.jzehrlab.com/>, click on the “nifH” tab, and then click the “nifH database” button. You’ll see this list of links:



Click on “[nifH\_2017June\_ARBUpdate\_nt\_fasta+metadata.zip](https://wwwzehr.pmc.ucsc.edu/nifH_Database_Public/nifH_2017June_ARBUpdate_nt_fasta+metadata.zip)” to download a zip file. Unzip it. You’ll see 2 file, one of which is a .fasta. Double-click on it. If you don’t see anything coherent, right-click and select “Open with…” Choose Notepad if you’re on Windows, choose TextEdit if you’re on a Mac. The top of the file looks like this:



Choose any 5 sequences where

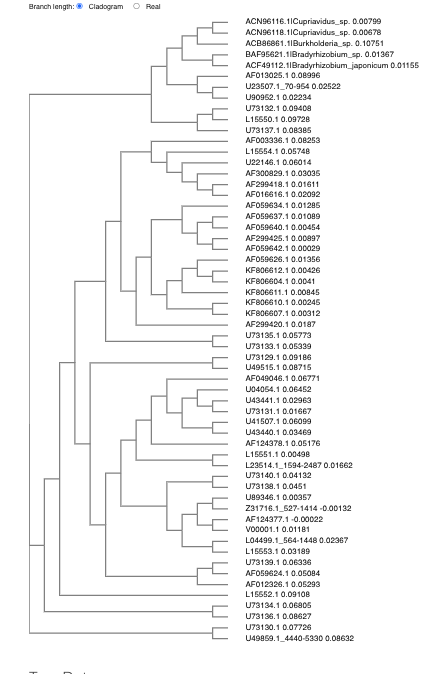
* The length is 300-700 bp
* The accession number isn't already in your fasta file
* The scientific name is given (so don’t choose something like “uncultured archaeon”), and the name isn't *Candidatus Atelocyanobacterium thalassa*.

Paste the 5 defline+sequence records onto the end of your fasta file. What accession numbers and organisms did you choose? (Accession numbers appear on the deflines, between “>” and “|”). Fill in the table below.

|  |  |
| --- | --- |
| Accession # | Organism (formatted correctly!) |
| BAF95621.1 | >BAF95621.1|Bradyrhizobium\_sp. 357 bp dna  ATCCTGATTGTAGGGTGCGATCCGAAAGCGGACTCGACCCGCCTAATTCTGCATGCCAAGGCTCAAGACACGATTTTGAGTCTTGCCGCGAGCGCTGGCAGCGTGGAGGACCTAGAGCTCGAGGACGTAATGAAGGTCGGCTACCAGGACATTCGCTGCGTCGAGTCCGGTGGTCCTGAGCCAGGTGTCGGTTGCGCCGGCCGCGGTGTCATCACCTCGATCAATTTTCTTGAAGAGAACGGAGCCTATGAGAACATTGACTATGTCTCTTACGATGTGCTTGGCGACGTTGTTTGCGGTGGCTTTGCGATGCCAATCCGCGAGAACAAGGCGCAGGAGATCTACATCGTGATGTCT |
| ACN96116.1 | >ACN96116.1|Cupriavidus\_sp. 474 bp dna  CTGGGCGAGATTCTCATCGTTGGGTGTGATCCAAAGGCGGACTCCACACGGCTTATTTTGCATGCTAAAGCACAGGATACAATCCTCTCGCTTGCGGCTGAAGCGGGATCCGTCGAGGATCTTGAACTGGATGACGTAATGAAGATTGGTTACAAGGACATCCGATGTGTCGAGTCGGGGGGGCCTGAACCGGGGGTTGGCTGCGCAGGCAGAGGTGTTATAACGTCAATCAATTTCCTGGAAGAGAACGGCGCGTATGACGGTGTTGATTACGTTTCCTATGATGTGCTCGGCGATGTCGTCTGTGGCGGTTTCGCGATGCCAATACGTGAAAACAAGGCCCAAGAGATCTACATTGTTATGTCTGGCGAGATGATGGCGATGTACGCCGCAAACAACATTTCTAAAGGAATTTTGAAGTACGCAAATAGCGGCGGCGTGCGCTTGGGTGGGTTGATCTGTACGAAGGCGAAG |
| ACN96118.1 | >ACN96118.1|Cupriavidus\_sp. 474 bp dna  CTGGGCGAGATTCTCATCGTTGGGTGTGATCCAAAGGCGGACTCCACACGGCTTATTTTGCATGCTAAAGCACAGGATACAATCCTCTCGCTTGCGGCTGAAGCGGGATCCGTCGAGGATCTTGAACTGGATGACGTAATGAAGATTGGTTACAAGGACATCCGATGTGTCGAGTCGGGGGGGCCTGAACCGGGGGTTGGCTGCGCAGGCAGAGGTGTTATAACGTCAATCAATTTCCTGGAAGAGAACGGCGCGTATGACGGTGTTGATTACGTTTCCTATGATGTGCTCGGCGATGTCGTCTGTGGCGGTTTCGCGATGCCAATACGTGAAAACAAGGCCCAAGAGATCTACATTGTTATGTCTGGCGAGATGATGGCGATGTACGCCGCAAACAACATTTCTAAAGGAATTTTGAAGTACGCAAATAGCGGCGGCGTGCGCTTGGGTGGGTTGATCTGTAACGAGCGAAGG |
| ACB86861.1 | >ACB86861.1|Burkholderia\_sp. 567 bp dna  ATACTCATCGTCGGCTGCGATCCGAAGGCTGACTCGACACGGCTGATATTGCATGCGAAGGCACAGGACACAATTCTCTCTCTCGCGGCGGAAGCCGGCTCCGTGGAGGACCTCGAACTCGAGGATGTCATGAAGATTGGCTACAAAGACATACGCTGTGTGGAGTCGGGTGGTCCAGAGCCGGGTGTCGGCTGCGCGGGTCGCGGTGTCATTACATCGATTAATTTCCTGGAGGANAATGGCGCCTACGACGGGGTCGACTACGTTTCATACGATGTGCTGGGCGACGTGGTGTGCGGAGGATTTGCAATGCCAATTCGCGAGAACAAGGCACAGGAGATTTATATCGTGATGTCAGGTGAGATGATGGCTATGTACGCTGCGAACAATATCTCAAAAGGGATTCTCAAGTATGCCAACAGCGGTGGCGTGCGTTTGGGCGGCCTGATTTGCAACGAGCGAAAGACCGACAAGGAACTCGAGCTTGCGGAATCGCTCGCAGCGATGCTCGGCACACGTCTGATCCACTTTGTACCGCGCGACAACATCGTTCAGCATGCGGAGCTA |
| ACF49112.1 | >ACF49112.1|Bradyrhizobium\_japonicum 687 bp dna  TCAGAAATCCTGATTGTAGGGTGCGATCCGAAAGCGGACTCGACCCGCCTTATTCTGCACGCCAAGGCTCAAGACACGATTTTGAGTCTTGCCGCGAGCGCCGGCAGCGTGGAGGATCTAGAGCTCGAGGACGTAATGAAGGTCGGCTACCAAGACATTCGCTGCGTCGAGTCCGGTGGTCCTGAGCCAGGTGTTGGCTGCGCCGGCCGCGGTGTCATCACCTCGATCAATTTTCTTGAAGAGAACGGAGCCTACGAGAACATTGACTATGTGTCTTACGATGTGCTTGGCGACGTTGTTTGCGGTGGCTTTGCGATGCCAATCCGCGAGAACAAGGCGCAGGAGATCTACATCGTGATGTCTGGTGAAATGATGGCAATGTATGCCGCAAACAATATTTCCAAGGGGATCCTGAAATACGCGAACTCAGGTGGGGTGCGGTTGGGCGGTCTGATCTGCAACGAGCGGCAGACTGACAAGGAATTGGAACTGGCGGACGCGTTGGCCAAGAAGCTTGGCACTCAACTGATCTACTTCGTGCCGCGTGACAATGTGGTGCAGCATGCAGAGCTGCGTCGCATGACGGTGCTTGAATATGCACCCGATTCCAAGCAGGCTGATCACTATCGGAAACTAGCGGCCAAGGTTCACAATAATGGCGGCAAAGGCATCATCCCGACCCCGATC |

Build a new tree

Browse to Clustal-Omega, align your sequences, and click the “Phylogenetic Tree” button. Paste a screenshot of the new tree. Is UCYN-A still monophyletic? Justify your answer.



**The new tree is not monophyletic. There appeared to be a clade earlier before more sequences were added and as they are added, the more they are closely related to other non-UCNY-A. The number went from 7 to 5 in the supposed clade before. It appears to be polyphyletic.**