# Bi1X Spring 2011
# Sequencing Project – Class Exercise

## Overview

The main objective of this exercise is to show you a simplified version of the way microbiologists identify small subunit ribosomal RNA sequences ("16S" sequences) from the environment. You will learn how to:

(1) Convert your sequencing traces to nucleotides
(2) Align the 16S sequences using the greengenes website
(3) Scan your sequences for chimeras, which you will eliminate from your database
(4) Use MEGA to produce a phylogenetic tree of your sequences and their nearest neighbors (found by greengenes).

You will then measure in MEGA the distance between your 16S sequences and the nearest neighbors and determine if they are of the same species, genus or phylum. Finally you will identify the precise taxonomic classification of the nearest neighbor for each of your 16S sequences and explain if given what's known in the literature about the physiology of these isolates it's plausible that you found these microbes in a fresh water pond (or could it be a contamination?). Please follow the instructions below and answer the questions in the bullets as you go along.

**Computer requirements:** to complete this exercise you will need a (1) PC or a Mac and (2) internet access.

**Installations before class:**
Please have the following programs installed *before you arrive in class*. This will make the analysis much faster and easier.

- WinRAR (for PCs): available at http://www.win-rar.com
- MacRAR (for Macs): available at http://www.macrar.com/
- UGENE (download the appropriate OS version): available at http://ugene.unipro.ru/download.html
- MEGA 5 (download the appropriate OS version): available at http://www.megasoftware.net/

## Part I: Learn a little more about the 16S gene

The *rrsA* gene of *E. coli* codes for a small subunit ribosomal RNA (rRNA) gene. Let's learn a little more about this gene. Go to the BioCyc microbial database (the default organism being *E. coli*)

**http://ecocyc.org/**

In the search box type *rrsA* and click **rrsA** under **genes**.

1. What is the length of this gene?
2. Since you used 'all bacterial primers' for your PCR your primers should also match this rRNA gene. Find out the length of the amplicon you generated. The primer sequences you used were

   357F (16S forward primer) 5' CTCCTACGGGAGGCAGCAG 3'
   492RL2D (16S reverse primer) 5' TACGGYTACCTTGTTACGACTT 3'

   Here is a tool to help you reverse complement a nt sequence
   http://www.geneinfinity.org/sms/sms_reversecomplement.html

   ***Hint 1.*** *to get E. coli's 16S sequence click on the* ***rrsA*** *hyperlink under* ***Genes*** *and then* press the **Nucleotide Sequence** button. *To convert this sequence into a long string of letters copy this sequence into notepad in order to remove non-text characters. Then copy this sequence into Word and replaces spaces " " with an empty string and then replace the newline character "^p" with an empty string.*

   Save *E. coli*'s sequence for later, we will be using it again for our tree.

   ***Hint 2.*** *What is special about the nucleotide sequence of one of the primers above? Look at them carefully…no, it's not a mistake…This website can perhaps help you:* ***http://www.le.ac.uk/bl/phh4/nucredun.htm***

3. What is the genomic context of this gene? Click on the **Genome Browser** button. You will see that *rrsA* is part of an operon. What other genes are present in this operon? What do they do? Is their function related to *rrsA*?
4. Does *E. coli* have more than one copy of this gene? In the EcoCyc website go to **Search → BLAST** and BLAST the *nucleotide* sequence of the rrsA gene against *E. coli's* genome (the default database). How many hits do you get? What is the name of the homologous genes that you found? Why do you think *E. coli* should have multiple copies of this gene?
5. How about other organisms? Do other organisms have multiple copies of this gene? What is range of copy numbers in nature? To find out the answer take the forward primer above and BLAST it against other organisms. You can change the organism database in BioCyc by clicking on the **change** hyperlink at the upper right corner. Try this for a few prokaryotes, including the spore forming bacterium *Bacillus subtilis* (a different strain of which causes anthrax), the highly abundant cyanobacterium *Prochlorococcus marinus*, the human pathogen *Streptococcus pyogenes* and finally the extreme thermophile *Thermococcus kodakarensis*. ***Hint:*** *don't forget to set BLAST to compare a nucleotide sequences with a nucleotide database*. ***Bonus question:*** *You should have failed to get an answer for one of these prokaryotes. Which was it? Can you give a reason why?* Now double check your results using the ribosome RNA copy number database (rrnDB) (also for the prokaryote that failed this test).

http://ribosome.mmg.msu.edu/rrndb/search.php

Just type the name of the prokaryote in the **keyword** box.

6. How do you think the rRNA copy number can affect the PCR reaction you carried out? Do all prokaryotes get equal representation? If all bacteria had the same rRNA copy number per genome, would then all prokaryotes get equal representation?

## Part II: Viewing your SSU rRNA sequences

1. Download the chromatograms (.ab1) of your sequencing results from the Bi 1X website. Extract files in the .zip folder.

   *In a chromatogram, each nucleotide base is represented by a different color. One can read out the sequences from the peak trace. Less distinguishable peaks indicate ambiguous or noisy data.*

2. Run UGENE. Open all of the chromatograms at the same time by highlighting all of the .ab1 files.

3. Click "Project" on the left column to see the list of files imported.

4. Double click on "[c] Chromatogram" for one of the sequencing results. Zoom in the chromatogram by pressing the magnification tool on the top right of the chromatogram window until you can see individual peaks.

5. Question: *For each sequence, describe the quality of the sequencing result by examining its chromatogram. Is the peak trace noisy or clearly distinguishable? How long is the sequence corresponding to the clearly distinguishable region?*

## Part III: Align your SSU rRNA sequences and find nearest neighbor matches

1. **Greengenes tutorial.** You will use an online tool called greengenes hosted by Lawrence Berkeley National labs to align your 16S sequences and find nearest neighbors. Get acquainted with this tool by reading the online tutorial.

   http://greengenes.lbl.gov/cgi-bin/JD_Tutorial/nph-Tutorial_2Main2.cgi

2. **Convert your abi trace files to nucleotides and generate your FASTA files.** Most sequencing reactions produce ~700-900 basepairs of high-quality sequences called "reads", which is not enough cover the entire 16S amplicon. To sequence your entire amplicon we would need to run two sequencing reactions in opposite

directions using both the forward and reverse primers and then "assemble" the two reads. However the read length that we obtain with just one sequencing reaction is long enough for us to get reasonable classification of the organism. The first step however is to obtain quality nucleotides from your traces.

Examine the chromatogram of your sequencing result and identify the region with the longest sequences that are still of high quality (that is, you can easily distinguish individual peaks) – see Fig. 1.

In UGENE, there is a quality bar on top of the chromatogram with higher bars correspond to higher qualities (see screenshot below). If you don't see the quality bar, right click (PC) or ctrl + left click (Mac) on the chromatogram and select the option "Show quality bars".

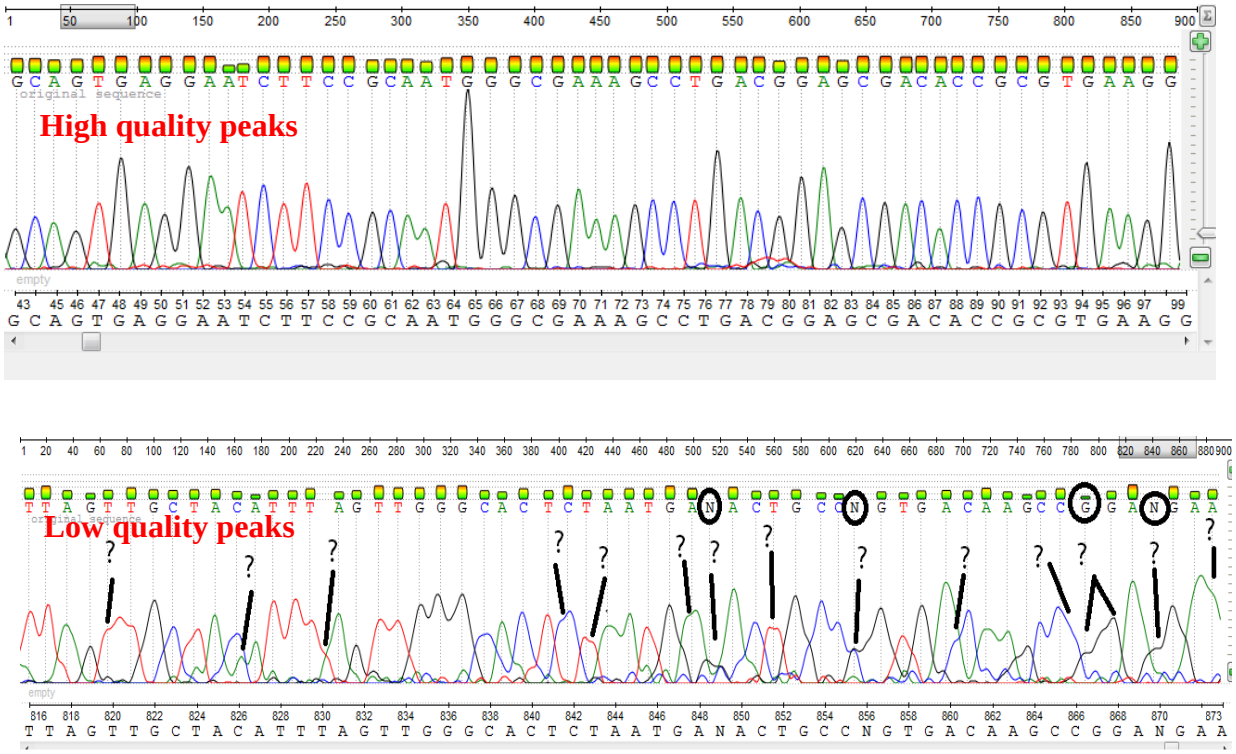


**Figure 1.** Example of high quality bases and low quality bases.

Select the identified nucleotide sequence by right clicking (PC) or ctrl + left clicking (Mac) on the chromatogram and choosing "Select → Sequence region" option. Enter the starting and ending positions of the sequence and click the "OK" button.

Copy and paste the nucleotide sequence of each trace in a text editor. Save each sequence as a separate FASTA file. A FASTA file is a text file (.txt) with the following simple format:

```
>SEQ_NAME
SEQUENCE...
```

The format is not case sensitive, but use only alpha-numeric characters and avoid spaces.

3. **Chimeras.** When running your PCR reaction there is the chance that two different PCR products will combine to form a chimera amplicon, which of course is an unwanted artifact that needs to be identified and discarded. There are several online tools to help you identify these artifactual sequences. We will be using a tool called Bellerophon. Read more about chimeras and how Bellerophon works in the following link

   http://comp-bio.anu.edu.au/Bellerophon/doc/doc.html

   Submit a FASTA file containing *all* of your clone library sequences to the Bellerophon server to check for chimeras (combine all the FASTA files that you have to one big FASTA file):

   http://comp-bio.anu.edu.au/bellerophon/bellerophon.pl

   Make sure to check "align sequences". If your sequences are shorter than 600bp change the window size to 200bp.

   Do you have any chimeras? Eliminate any chimeras from analysis.

4. **Align sequences.** Next you will use greengenes to align your 16S sequences. Alignment of 16S sequences is different than alignment of protein coding nucleotide sequences because the 16S aligners take into account the secondary structure of the 16S gene. For example, sequences on both sides of a stem loop need to match up in order to form the stem loop. Greengenes also attempts to align your sequences against a very large database of pre-aligned 16S sequences and will include in the results the nearest neighbors it could find (i.e. other 16S sequences with the highest similarly to your sequence). You can also specify if you want greengenes to return any nearest neighbors (which include uncultured organisms) or nearest neighbors which are isolates (and thus better characterized). We will choose both. The nearest neighbor sequences will be returned in separate FASTA files (see below). To proceed, load your FASTA files one at a time to the greengenes aligner **(Align** icon**)** using the settings below:

   http://greengenes.lbl.gov/cgi-bin/nph-NAST_align.cgi

   - **My Taxonomy** (lower right)– activate **ncbi** taxonomy (choosing all options) → **make changes to my interest list**→ click on the **Align** icon again

- **Minimum length** = 500bp (this should be lower than the length of your submitted sequence)
- Check options 1,2, 4 and 5 in "Files you desire"
  - xyz_NAST.xls (greengene alignment report)
  - xyz_NAST.fasta (your aligned sequences)
  - xyz_nn_NAST.fasta (nearest neighbors)
  - xyz_nni_NAST.fasta (nearest isolates)
- Set the number of sequences to be returned to 5
- Leave the rest of the parameters to their default values
- Enter your email address and hit **proceed.** Progress is updated on your web browser.
- Use WinRAR (PC) or MacRAR (Mac) to open the files on your computer if they do not open directly.
- After you obtain the results, based on the tutorial, inspect the excel file to see which sequences were aligned and identified.

***Hints:*** *If you receive empty nn* ***and*** *nni files check the following:*

- *Did you activate the ncbi taxonomy?*
- *Is your sequence longer than the* ***Minimum length*** *cutoff specified in the greengenes options?*
- *Is the text file you submitted in FASTA format (see above)?*
- *Did you select only high quality base pairs for your sequence?*
- *Try reducing the* ***Minimum Percent Identity*** *to 50%, 30% and 10%.*
- *If you still get empty nn and nni files, try to classify your organism using the rdb classifier:*

  *http://rdp.cme.msu.edu/classifier/classifier.jsp*

  *If your sequence is unclassified at the phylum level then just submit this classification and move to the next 16S sequence. Otherwise, you may have overlooked one of the previous issues. If you are still stuck, ask one of your TAs for help.*

5. **Create your MEGA FASTA file.** Combine your xyz_NAST.fasta file with your xyz_nni_NAST.fasta file and your xyz_nn_NAST.fasta files (do this for each of your 16S pond sequence separately since these sequences can be quite divergent and the treeing algorithm assumes the sequences are similar). You may want to add a header to the sequence name to distinguish nni from nn sequences. This will be your new FASTA file. We will next use MEGA, which is a little fussy about the input format of the FASTA file. To convert the greengene FASTA file to a format MEGA can do the following:

   - Replace spaces (" ") and any "=" symbols with underscores "_".
   - Replace '.' symbols with dashes '-'.

- Make sure the extension of the FASTA file is 'fasta'. To change an extension of text file on a PC use Notepad and in "save as" change the "save as type" from "Text Documents" to "All Files" and use the file extension 'fasta'.

## Part IV: Phylogenetic analysis

1. **Phylogenetic analysis.** We would like to generate a Neighbor Joining tree for each of your 16S sequences and its nearest neighbors. This will allow you to visualize the evolutionary distance between each of your sequences and their nearest neighbors. We will also calculate precisely the similarity between these 16S sequences. Based on this percent similarity you can determine the degree of certainty in your taxonomic classification. Since the 16S gene is highly conserved, mutations correlate with time, and so the amount of sequence divergence correlates with evolutionary separation. For example, two 16S sequences having ≥97% identity are generally considered to belong to the same species, ≥95% identity the same genus and ≥80% identity the same phylum. Based on these empirical rules you can determine if your nearest neighbor is the same species, genus or phylum as the 16S sequence you found.

2. **Constructing a phylogenetic tree in MEGA.**
   Open MEGA and drag your corrected FASTA file into MEGA's main window (turquoise window).
   - Select the "analyze" option
   - Select "nucleotide sequences"
   - When asked "Protein coding data?" select No
   - Click on the "TA" box in the MEGA work environment

   Next, build a NJ tree with bootstrapping using a Jukes-Cantor nucleotide substitution model. Go to MEGA's main window and then choose:

   - **Phylogeny → Construct/Test Neighbor Joining**
     - **Phylogeny Test**
       - Test of Phylogeny → Bootstrap method
       - No. of boostrap replications: 1000
     - **Substitution Model**:
       - **Substitution Type → Nucleotide**
       - **Model/Method → Jukes-Cantor**
     - Hit **compute**
     - Press the **"caption"** button on the tree you just created to see the precise definition of this tree. How many nucleotide positions were used in the tree? You should get at least 500bp, otherwise something maybe wrong – ask one of your TAs.

   You should obtain a tree that looks something like Fig. 2. Here are some assignments and questions for you regarding your trees (answer for each tree you generate)

a. Draw your NJ tree using two topologies: rectangular and radiation. What is the difference between these two topologies? (submit only the rectangular tree)
b. Why is the Jukes-Cantor model more accurate than simply using the p-distance to calculate evolutionary distances? What does the Jukes-Cantor model attempt to correct for?
c. Do the bootstrap values support your classification?
d. The tree that you submitted is an unrooted tree. What kind of organism would you need to root this tree? What would you gain by rooting the tree?
e. Use MEGA to find the percent identity between your sequence and its nearest neighbors. To calculate distances in MEGA, from MEGA's main window choose

- **Distances → compute pairwise**
  - **Substitution Model**
    - **Model/Method → p-distance**

See Fig. 3. for an example how this output looks like. The results are 1 - percent identity/100.

***Hint:*** *you can drag up and down the sequences in the resulting distance matrix window to help you located your nearest neighbors*.

f. Who is the nearest 16S sequence to each of your pond sequences? What is the percent identity to that sequence? Is the nearest neighbor the same species, genus or phylum as your 16S? Can you use the nearest neighbor to classify your organism reliably?

**3. Classify your 16S sequences.** Obtain further NCBI taxonomic information about the nearest neighbor to your 16S sequence (preferably an isolate) by searching its NCBI ascension number in the NCBI nucleotide database

http://www.ncbi.nlm.nih.gov/nuccore/

***Hint:*** *the NCBI ascension number appears in the organism name given by greengenes and is the second field after the first underscore). See an example in Fig. 4.*

a. What is the precise taxonomic classification of your nearest neighbors?
b. Read up on the microbes you identified (e.g. by looking up the references in the NCBI record, searching Google scholar for the microbe's name, searching WikiSpecies, etc.). Given this organism's physiology (respiration method, energy and carbon sources, typical osmolarity/pH/temperature ranges) is it

plausible you found it in a fresh water pond or could it be a contaminating sequence in your PCR?

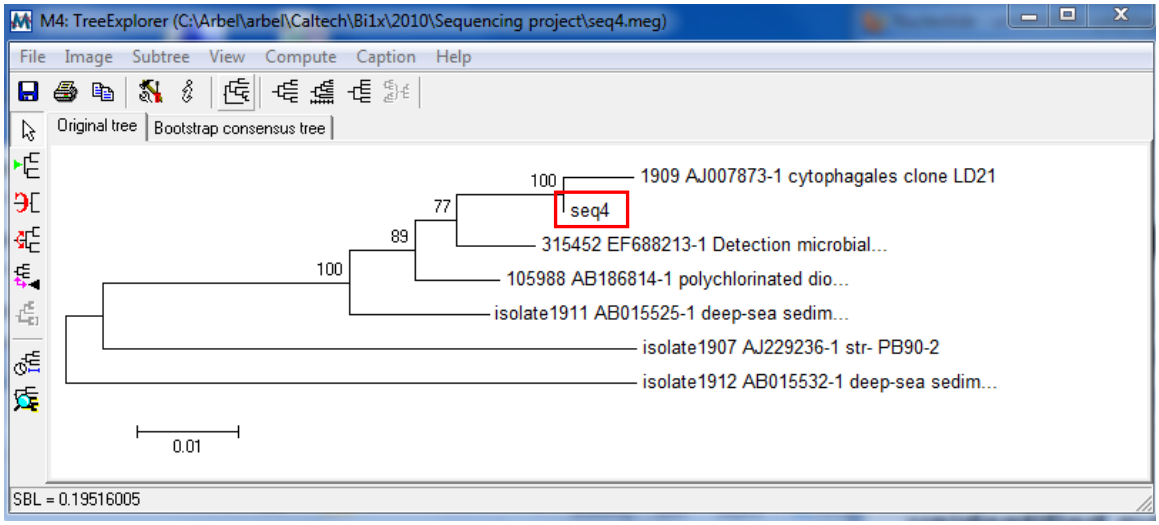


**Figure 2.** Example of a NJ tree with 1000 bootstrap replicates generated by MEGA to identify seq4. There were a total of 726 positions in the final dataset.

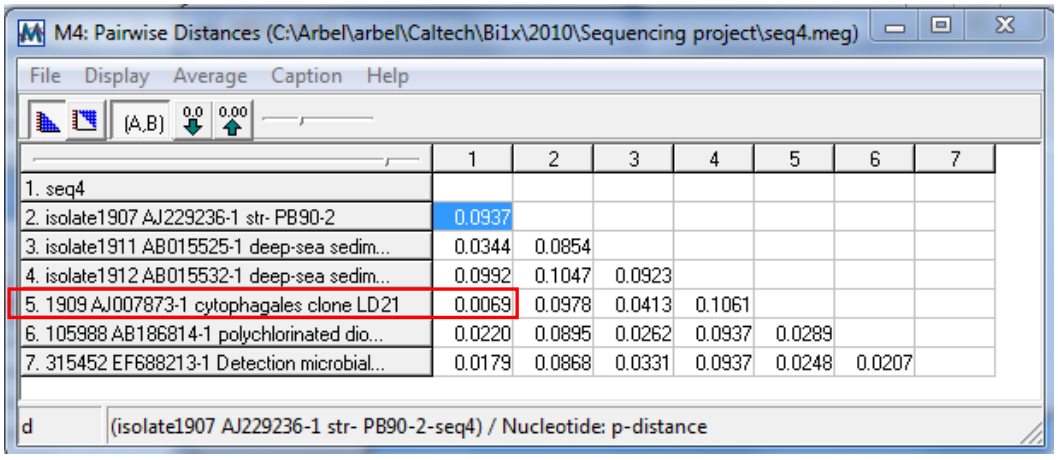


**Figure 3.** p-distance between seq4 and its nearest neighbors. The minimum distance is with respect to the cytophagales clone (99.37% identity).

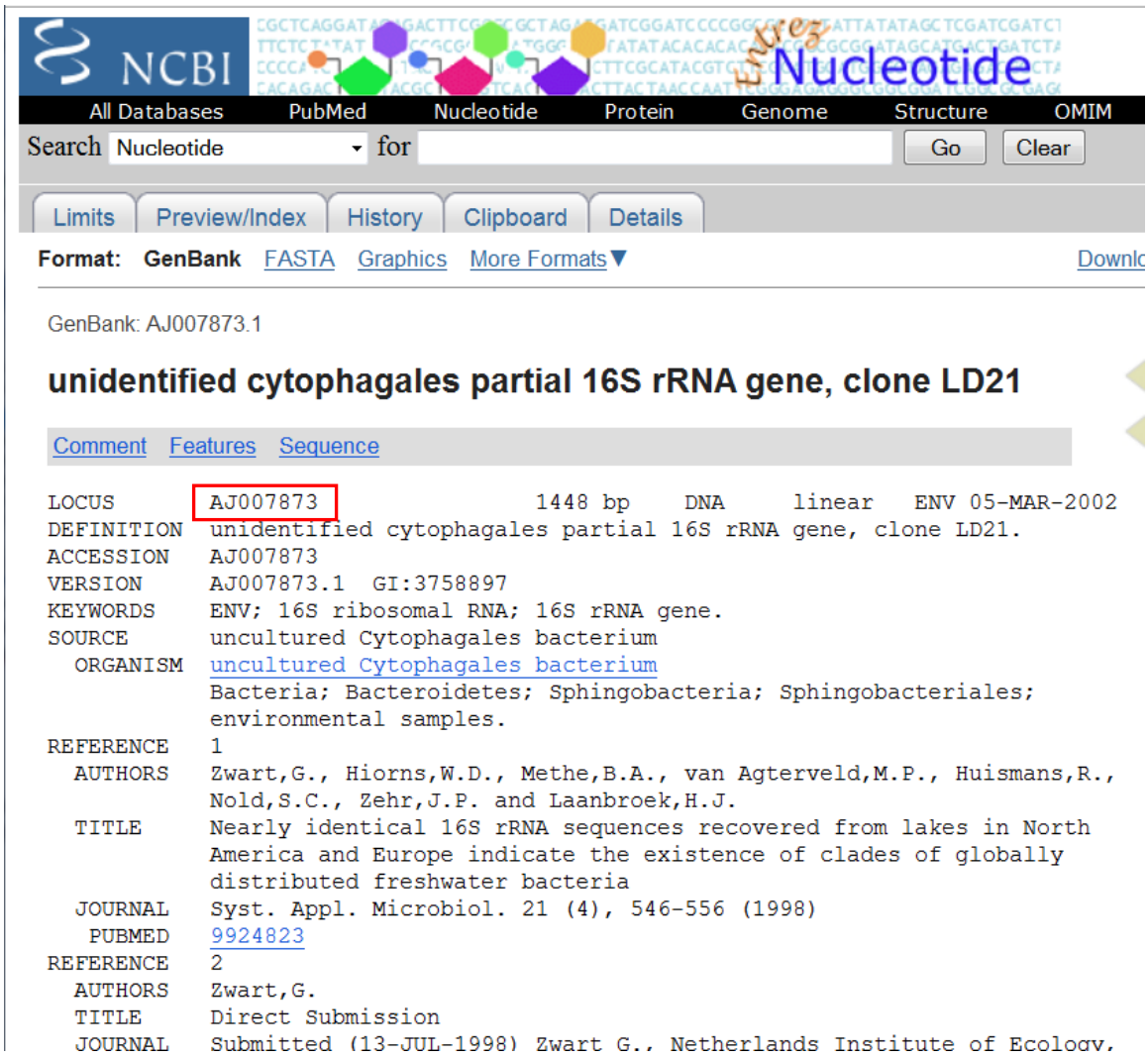


**Figure 4.** NCBI recored for cytophagales clone (AJ007873-1)