

**An Introduction to Bioinformatics and Sequence Analysis Using Tetracycline Resistance Genes**

***Pilot draft copy***

**An extension of the Prevalence of Antibiotic-Resistance in the Environment (PARE) program**

*Note: This is a first round draft copy available to PARE instructors. Please contact* [*carol.bascom\_slack@tufts.edu*](mailto:carol.bascom_slack@tufts.edu) *for feedback and suggestions for modifications to this protocol or return a “track changes” copy.*

**Jennifer Larson, Capital University**



**Author:**

**Jennifer Larson, Capital University**

**Summary**: The goal of this PARE expansion module is to expose students to online bioinformatics tools including DNA and protein sequence comparisons. Students will use the tetracycline resistance gene/protein Tet(M) as a subject for these analyses. This exercise can be performed on its own or as a separate exercise in conjunction with the basic PARE procedures without any additional materials.

**Integration with Other PARE modules**:

This module is complementary with the soil metagenomic analysis for TetR module (currently in pilot phase), but it is not necessary to complete any other PARE modules to use this module.

**Student Learning Goals/expected outcomes:**

Students should be able to:

* define bioinformatics
* perform DNA sequence comparisons using available online tools
* perform protein sequence comparisons using available online tools

**In-class time requirement:**

2 hours at one time**.**

**Out-of-class/prep time requirement:**

Familiarization with protocol: 30 min. This protocol assumes the instructor has basic familiarity with DNA analysis.

**Background**

**Bioinformatics** is the use of technology to analyze and understand biological data. This exercise will take you through some of the tools available on the internet that can be used to analyze DNA and protein sequence data.

One of the most prevalent genes known to confer resistance to tetracycline is called *tet*(M). The *tet*(M) gene belongs to a class of resistance genes that encode ribosomal protection proteins. The tetracycline drug works by inhibiting the function of ribosomes, resulting in halting production of proteins. Tet(M) is a ribosomal protection protein that binds directly to the ribosome, blocking binding of the tetracycline antibiotic. When the tetracycline can’t bind, it does not inhibit protein synthesis, so the bacterium can produce proteins and grow in the presence of the tetracycline antibiotic. Ribosomal protection is only one category of tetracycline resistance, but there are many others (*e.g*. efflux pumps, enzymatic degradation of tetracycline and rRNA mutations). In this module, we will only focus on the ribosomal protection protein, TetM.

Note: In bacterial nomenclature, gene names usually begin with a small letter and are italicized while the protein encoded by the same gene is depicted using the capitalized, non-italicized version of the same name. *tet*(M) refers to the gene encoding the Tet(M) protein.

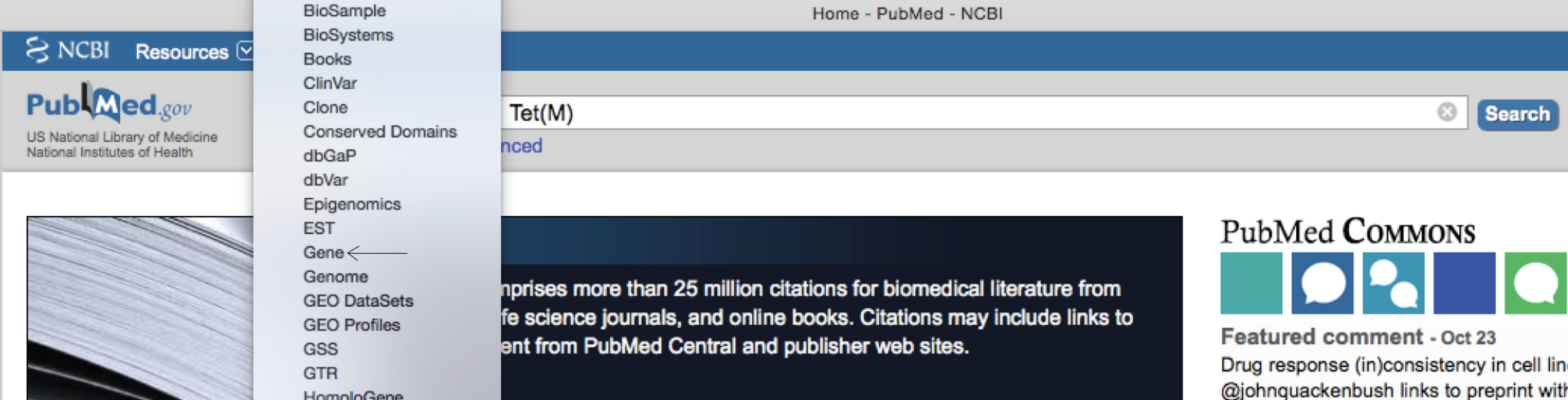
The *tet*(M) gene has been found in many different types of bacteria. Analysis of the DNA sequence has revealed small differences in the DNA sequences isolated from these different bacteria. Examining differences in the same gene from multiple organisms can give scientists insight into how a gene has evolved or might be evolving, and also provide some clue as to how closely related (evolutionarily) two different organisms are to each other. Today you will use the *tet*(M) gene to examine some of the tools available to scientists for analyzing and comparing DNA and protein sequence data.

**Methods**

The National Center for Biotechnology Information (NCBI) is a database run through the U.S. National Library of Medicine. It contains a wealth of information related to biomedicine and genomics.

**1. To access this database go to the following website:** [**http://www.pubmed.com**](http://www.pubmed.com)**.**

This will direct you to the NCBI collection of databases. From here you can search for information related to a topic from a number of different databases. Begin by finding the sequences of *tet*(M) genes.



In the Search field at the top type in “tet(M)” and use the database pull-down menu (to the left of the Search field) to select “Gene”. Then click on “Search”.

You should now see a list of over 100 tetracycline resistance genes sequenced from several different organisms.

**2. Select one of the *tet*(M) gene sequences. Which organism was it sequenced from?**

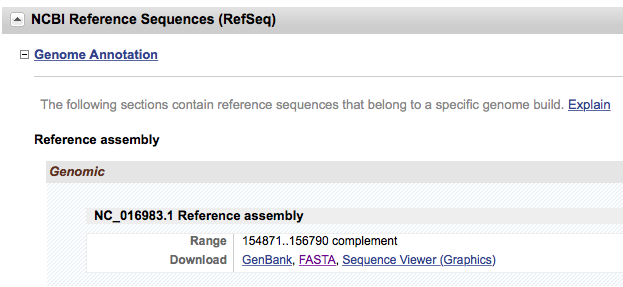
If you click on one of the *tet*(M) gene sequences, you will be shown a lot of information related to that DNA sequence.

Under “Bibliography” you can see where this sequence came from. Where did the sequence you chose come from?

**3. Cite the reference using proper formatting for a scientific citation, such as those you might use for a poster project.**

If you do not see this section or if this section does not contain a reference for the gene sequence you selected, go back and select another gene sequence.

**Gene sequence retrieval**

**4. Collect some gene sequences by scrolling down a little further until you see a heading for “NCBI Reference Sequences (RefSeq)”.**

DNA sequences can be listed in several formats. The FASTA sequence format is recognized by most sequence analysis software programs accessible on the internet so select “FASTA”. You should now see a DNA sequence. Notice that just above the DNA base sequence there is a “>” next to the name of the bacterial species. This symbol is used to tell computer programs that this is the title and the start of a new DNA sequence. This will be important information to tell the sequence analysis programs where a new sequence starts when you analyze more than one sequence at a time.

**5. Copy and paste your DNA sequence in the space below on this document (we will use this later in step #7). Include the title of the sequence. Note that you will need to shorten the sequence title so that it is one word for later analysis.**

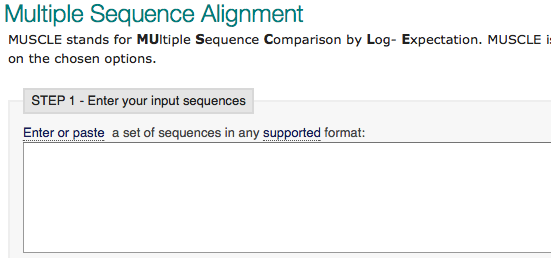
**6. Go back to your list of *tet*(M) sequences and select a *tet*(M) sequence from another organism. Copy and paste the FASTA sequence in the space below. Shorten the sequence title so that it is one word.**

**Gene sequence alignment and comparison**

Now that you have 2 DNA sequences you are prepared to perform a comparison.

**7. To access a DNA sequence alignment program go to the following website:** [**http://www.ebi.ac.uk/Tools/msa/muscle/**](http://www.ebi.ac.uk/Tools/msa/muscle/)

This link will take you to the “MUSCLE” multiple sequence alignment software site. In the box under “STEP 1” copy and paste BOTH of your DNA sequences, including the FASTA >sequence name. Leave one empty line between the two sequences.



**8. Click on “Submit”.**

The program will align your two DNA sequences. Take a look at the alignment it gives you.

**Are your two *tet*(M) genes completely identical? What does the star symbol (\*) mean?**

**Protein sequence alignment and comparison**

**9. To compare protein sequence alignments, return to the NCBI website (pubmed.com).**

In the search field, type “tet(M)” again but this time select “Protein” from the menu. You should see a list of protein sequences related to tetracycline resistance. Find one for Tet(M) and click on it. Scroll to the bottom of the page. Under the heading “Origin” you should now see the primary structure (the amino acid sequence) of the Tet(M) protein from that particular organism.

**10. Copy and paste the sequence below.**

It is ok to include the numbers and spaces in the sequence. Add a title to the sequence at the top using the “>” symbol before your title (make sure you modify this title so that it is also one word).

**11. Go back to the list of sequences and select another Tet(M) protein sequence from a different organism and repeat the process.**

**12. Now that we have two sequences we will perform a protein sequence alignment. Go to the following website:** [**http://www.ebi.ac.uk/Tools/msa/clustalo/**](http://www.ebi.ac.uk/Tools/msa/clustalo/)

This software program can be used to compare your two protein sequences. Copy and paste your protein sequences into the empty field. Make sure you include the >Title so that the program will know that you are submitting 2 sequences and where they each start. Leave one blank line between your two sequences. Click on “Submit”. This may take a few minutes to run but you should receive an alignment of your two protein sequences.

**Are your two protein sequences identical?**

**13. Look at the symbols below the sequences. Keeping in mind that amino acids can differ in polarity, size, and charge, examine a table of amino acids to fill in the blanks below with what each of the symbols in the alignment mean.**

**Symbol Meaning**

**\* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**no symbol \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**14. Return to one of your Tet(M) protein sequences on the NCBI database website. On the right side you will see a list of options for analyzing the protein sequence. Select “Identify Conserved Domains”.**

Functional proteins are folded into a specific structure (tertiary/quaternary) within the cell. Different regions of the protein structure may have different functions. These different functional regions are called **domains**.

**15. Select the first domain listed.**

**What is the domain?**

**Describe in your own words what the function of that domain is.**

**Background information for instructors:**

Requirements for this exercise are that students have access to computers and to the internet. It is recommended that this exercise is done either individually or in groups of 2 when possible to keep everyone involved in the process.

**Things to watch out for—where students make mistakes**:

* Students frequently miss the direction to shorten the names of the sequences. For example, a sequence might be labeled:

>NZ\_CP033093.1:c34283-32364 Escherichia coli strain CP53 plasmid pCP53-38k, complete sequence

Shortening this title to: >Ecoli will help avoid possible sequence comparison errors later on.

* Students do not realize the importance of the “>” as a way of noting that this tells the software that there is a sequence following this label so they may accidentally delete it when changing/shortening their sequence label.
* Some DNA sequence or protein sequence entries will not contain all of the information described in this exercise, especially when it comes to the bibliography. When that happens, suggest to the students that they go back to the previous page and try selecting a different sequence.

**Pedagogical suggestions**

Questions to ask students are imbedded within the exercise. It is recommended that an electronic version of this lab be made accessible to students so that they may directly copy/paste into the document.

**Infrastructure/equipment requirements:** [*e.g.* fume hood, incubator, microscope, thermocycler, *etc*.]

Laptops (1 per person ideally; maximum 1 per group of 4)

Internet access

**Supply list:**

No wet lab supplies are needed

**Additional resources:**

Edvotek has a ready-to-run DNA sequencing gel kit, which could be added on to give exposure to Sanger sequencing and the historical process of sequencing DNA.

<https://www.edvotek.com/120>

**Suggested Assignment:**

**Pre-lab questions**

1. Examine the scientific literature and find a primary source that uses DNA sequencing. Describe what was sequenced in the paper and the reason it was sequenced.
2. Why are Sanger sequencing experiments divided into four reactions?