Lab Activity: Bioinformatics

Developed by: Karolina Fucikova, Assumption College, k.fucikova@assumption.edu
Simona Augyte, University of Connecticut, simona.augyte@uconn.edu

Learning Objectives

By the end of this activity, students should be able to:

- Glean relevant information from a primary scientific article (e.g., which genes are involved in a toxin production pathway).
- Carry out a custom BLAST search, justify selection of query sequences.
- Connect the results (presence/absence of toxin-linked genes in a genome or transcriptome) to the function of the genes' products (how the toxin is produced).
- Integrate their understanding of toxin formation with the ecological context of algal toxicity (e.g. blooms).

Assessment Method

A student will show they have mastered the learning objective when they can successfully run a bioinformatics search for algal toxin-synthesis genes and interpret the results. The best reports/presentations will connect the results (presence/absence of toxin-linked genes in a genome or transcriptome) to the function of the genes' products (how the toxin is produced), as well as discuss them in an ecological context - formation of toxic blooms, and what other factors it depends on.

Instructor Notes

Best if students work in groups of 2-3, discuss pre-lab questions and reading, formulate the hypothesis together, answer handout questions and prepare the report or presentation together.

Equipment required:
Personal laptop computer or provided desktop; internet access.

Techniques required (those which are not taught during the activity but students must already have a working knowledge):
Use of computer and internet. For graduate course, basic knowledge of command-line (unix)

Connections to major biology concepts:
Central dogma - gene (DNA, genome) vs. transcript (RNA, transcriptome) vs. toxins, enzymes, synthesis pathways (protein, proteome)
Nutrient cycling - algal blooms as a result of nutrient loading
Evolution - evolution of toxins (adaptive) in some dinoflagellates but not in others (or toxins secondarily lost)

Time required: 3 hr lab
Anticipated audience: 1) **intro majors course** 2) **upper level majors course** - using online BLAST search 4) **graduate course** - may increase difficulty by using custom-created BLAST databases from downloaded genomes or transcriptomes

Pre-lab Assignments ***ANSWER KEY***

Read the following; focus on the information related to dinoflagellates and production of toxins.
Ryan D.E., Pepper A.E. and Campbell L. 2014. De novo assembly and characterization of the transcriptome of the toxic dinoflagellate *Karenia brevis*. BMC Genomics 2014, 15:888.

Additional suggested reading:
Anderson D.M., Gilbert P.M. and Burkholder J.M. 2002. Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. Estuaries 25: 704-726.

Pre-lab concept check questions (to be completed before class)
1: Are all algal blooms toxic?
(A: no, some can be obnoxious or harmful in other ways but may not be toxic)

2: What are the common causes of toxic algal blooms?
(A: nutrient runoff from agriculture or sewage, among others)

3: How is a transcriptome different from a fully sequenced genome?
(A: transcriptomes only contain data from genes that are being expressed – transcribed – at the time when RNA is extracted from the sample; a genome contains all of the organism's genes and non-coding information)

4: Can a gene be found in a genome but not in a transcriptome? Explain your answer.
(A: yes because not all genes are transcribed at all times; therefore a transcriptome can provide information about whether the organism is making a particular product, like a toxin, at that time, not just if it has the genetic toolkit to do so)

5: Which genes are known to be involved in brevetoxin production? Are there other genes that are possibly indicative of toxin production?
(A: Polyketide synthase (PKS) genes – directly involved in brevetoxin production; voltage-gated ion channel protein genes, aquaporins and VATPases – involved in osmoregulation and putatively affected by brevetoxin presence)

6. Hypothesis:*Students should be able to formulate something like this*: If gene X is specifically linked to the synthesis of brevetoxin, it should be present in the genome/transcriptome of a known toxin-producing dinoflagellate, and absent in a non-toxic dinoflagellate genome/transcriptome.

Instructor: Potential Quiz questions

Q: What is an algal bloom and what causes it?
A: Abundant and rapid growth of algal cells in a freshwater or marine habitat, typically caused by increased nutrient availability.

Q: Match the terms on the left to the terms in the middle and phrases on the right to explain the role of sequences used as BLAST queries in your investigation.

Sequence used as query	**Function in the search**	**Used because**
Karenia brevis rubisco	Experimental query	Does not occur in algae
Karenia brevis polyketide synthase	Positive control	Expressed in all photosynthetic organisms
Homo sapiens insulin	Negative control	Directly involved in brevetoxin production
Karenia brevis aquaporin		Osmoregulation linked to brevetoxin presence

Q: While exploring a newly sequenced genome of an unidentified Dinoflagellate, you got a high-similarity BLAST hit for a gene involved in brevetoxin synthesis. Therefore, this alga will sooner or later form a toxic bloom. - is this an accurate statement? Explain your answer.
A: The alga may indeed be capable of producing brevetoxin – it has the genetic prerequisite for it. However, bloom formation depends on a number of environmental factors as well.

Potential Practical Exam options:

After being given a query protein sequence, determine whether a genome or transcriptome of another species contains a gene coding for this protein.

Example: student gets a FASTA file of superoxide dismutase of *Chlamydomonas reinhardtii* (AAB04944) including the gene/protein and organism name. Is this gene/protein also found in *Chlorella variabilis*? (pick a species, for which a whole genome or transcriptome is available) What does this protein do and what can we tell from its presence/absence in a genome or transcriptome?

Northeast Algal Society **Phycology Lab Manual**
Lab Activity: Bioinformatics
Developed by: Karolina Fucikova, Assumption College, k.fucikova@assumption.edu
Simona Augyte, University of Connecticut, simona.augyte@uconn.edu

Pre-lab Assignments Name: _____

Read the following; focus on the information related to dinoflagellates and production of toxins.
Ryan D.E., Pepper A.E. and Campbell L. 2014. De novo assembly and characterization of the transcriptome of the toxic dinoflagellate *Karenia brevis*. BMC Genomics 2014, 15:888.

Additional suggested reading:
Anderson D.M., Gilbert P.M. and Burkholder J.M. 2002. Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. Estuaries 25: 704-726.

Pre-lab concept check questions (to be completed before class)
1. Are all algal blooms toxic?

2: What are the common causes of toxic algal blooms?

3: How is a transcriptome different from a fully sequenced genome?

4: Can a gene be found in a genome but not in a transcriptome? Explain your answer.

5: Which genes are known to be involved in brevetoxin production? Are there other genes that are possibly indicative of toxin production?

6.Generate a hypothesis for the potential relationship between genes, genome/transcriptome and the synthesis of brevetoxin.

Lab Procedure - in groups of 2-3:

1. In the first 15min, discuss and refine your answers to the pre-lab questions with your teammates. As a group type up the revised answers.

2. Formulate a hypothesis connecting the genes identified in pre-lab questions to algal toxicity.

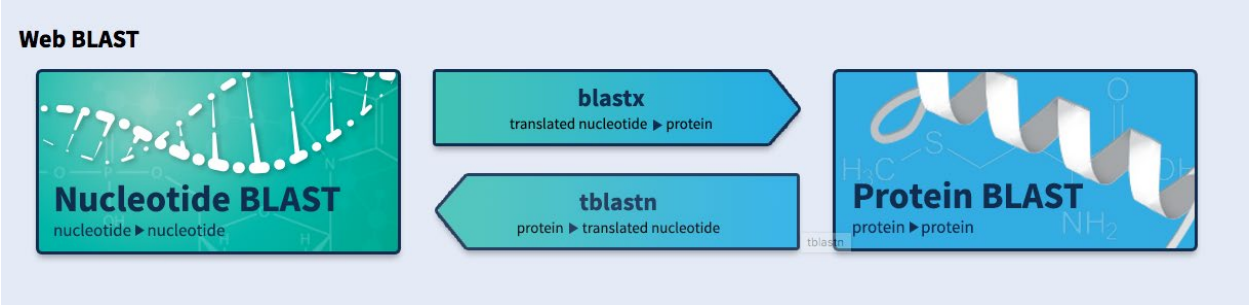
3. Select a species from a list provided by your instructor (Appendix 1). Some of the species on the list are thought to be toxic and some non-toxic. Spend 15 minutes on background research for your species. Start with a simple Google search but be sure to evaluate the websites you get information from – are they articles published in peer-reviewed journals? If not, do they cite primary articles? Do they cite any sources at all?).

4. Type up 5 key background facts you have learned about your species on the organismal level (for example, is your species freshwater or marine? Bloom former or not? Toxin-producer or not?). List your references, remember to only use appropriately vetted web sources.

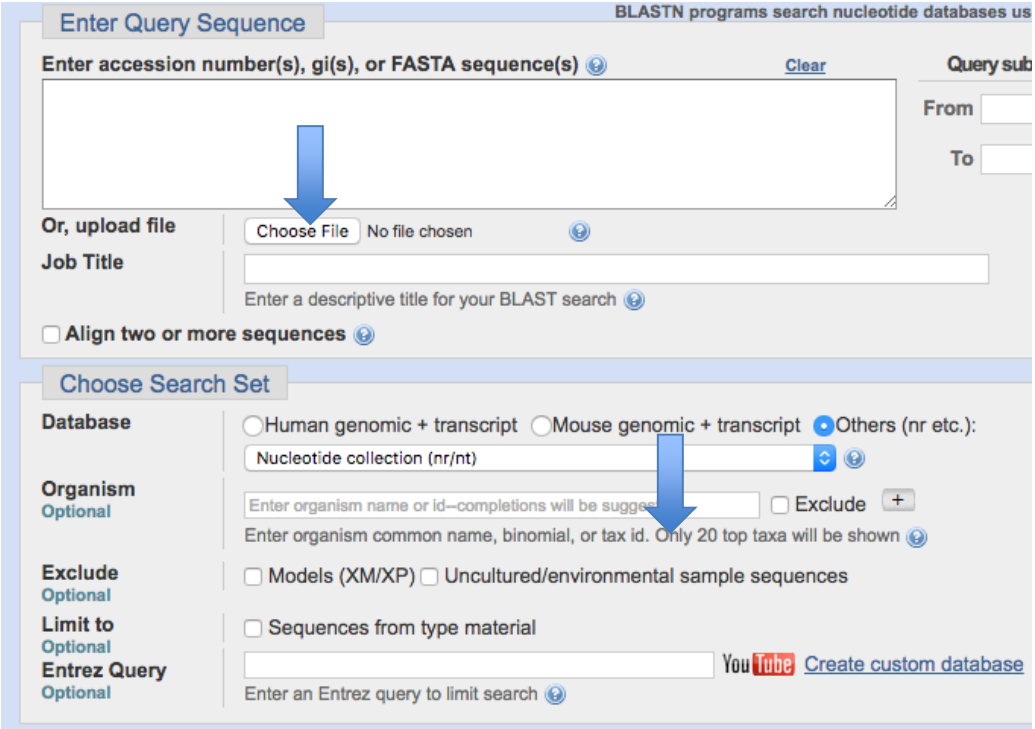
5. Use NCBI website (https://www.ncbi.nlm.nih.gov/) to obtain query sequences of *Karenia brevis* genes thought to be involved in brevetoxin production, which you identified in your pre-lab questions. Save all your query sequences into a plain text file in fasta format. You may use nucleotide or protein data, but not both as mixed data may produce errors. Name the file brevetoxin_query.fasta.

6. Use the BLAST web page (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to search the genome or transcriptome of your species:

First, select which BLAST program is appropriate for your search.



Discuss with your group what this diagrammatic overview of BLAST functions means. Where does your query file fit into this scheme? Can you guess why it wouldn't work to have protein and DNA data in the same query file? Write down which program you chose and why. Second, upload your brevetoxin_query.fasta file or copy and paste its entire content into the BLAST search window. Before hitting the BLAST button, restrict search to the species you are focusing on.



7. Use BLAST the same way as above to search the *Karenia brevis* genome and transcriptome.
8. Query both species with ALP13672 (*Karenia brevis* rubisco large subunit) and AAA59172 (human insulin). Discuss with your team the importance of this part of the experiment (further online research may be needed).

· Note: graduate courses may involve the use of command line, such as downloading query sequences using efetch, creating custom blast databases and searching them; see simplified example in the Appendix 2.

9. Summarize the results of your BLAST analysis in a table and a short descriptive paragraph.
10. Summarize the meaning of your results in a Discussion paragraph. What do the results tell you about your species? Does it produce toxins, or is it at least able to? Be sure to check if your matches come from a genome or transcriptome. Are these data consistent with what you have learned in your background research?

Post-lab Activities

Self-evaluation (10min at the end of lab):

Write a brief reflection on today's lab activity. Be sure to outline the content and identify the main take-home message(s). What was your contribution to your team's work? What did other teammates contribute? What have you learned from this exercise?

Appendix 1: List of species to use:
Query - *Karenia brevis* strain Wilson transcriptome - GI numbers provided in paper

Transcriptomes to be searched - can be used either as "unknown dinoflagellate" or using their actual names:
Gyrodiniellum shiwhaense (transcriptome: GFHE01000001-GFHE01009122)
Ansanella granifera (transcriptome: GFBE01000001-GFBE01083652)
Crypthecodinium cohnii (transcriptome: GFIV01000001-GFIV01087881)
Gambierdiscus polynesiensis (transcriptome: GETK01000001-GETK01115726)
Pyrodinium bahamense (transcriptome: GBXF01000001-GBXF01000002)
Noctiluca scintillans (transcriptome: GELK01000001-GELK01063552)
Karenia brevis strain SP1 (transcriptome at DOI 10.6070/H4VX0DH6)
Karenia brevis strain SP3 (transcriptome at DOI 10.6070/H4VX0DH6)

Appendix 2: Example exercise for an upper-level course with more extensive use of command line.

- Install blast tools (if you have Homebrew, it's a simple matter of 'brew install blast')
- Verify that your installation worked
    - which blastn
    - blastn -version
- Prepare the query file, either by manual copy and paste from NCBI or if you have efetch installed, get the query sequences using efetch. Example:
  efetch -db nuccore -id ALP13672 -format fasta >> query.fa
- Better yet, write a unix loop to get all sequences from a list of accessions into the file!
- Obtain the transcriptome you are about to search in fasta format. Put it in a new directory.
- Build a custom nucleotide and protein database from the transcriptome
    - makeblastdb -in yourtranscriptome.fa -dbtype prot -parse_seqids
    - repeat for nucleotide (-dbtype nucl)
- Use ls to see the new files created by makeblastdb
- Search the database using the appropriate blast program, for example:
    - blastp -db yourtranscriptome -query query.fa
    - modify the above line to save the output into a file named out.txt