Biol 145/CS 185C Final Project
Spring 2023

Introduction

The field of Marine Bioinformatics is growing and changing exponentially. In the class, we've aimed to give you a sense of the kinds of questions that can be asked about the ocean in a bioinformatic context. Now it is time to put your skills to the test. For this project, you can either do the "default project" described below or you can pursue your own question that you've developed in marine bioinformatics. If you chose to pursue your own question, then please let us know what you plan to do before you begin. Recall that this project is 30% of your grade, so we expect you to put a significant amount of time, care, and effort into producing a final project that you are proud of.

Due dates

- **5/17**: The written component of the project is due 5/17 by 11:59pm as a Word Doc and uploaded to Canvas
- **5/4, 5/9, or 5/17**: You will also present your project to the class on one of the three project dates. You will be able to sign up for a time to present on Canvas. Note that you are required to attend all of the presentation dates including the last one on 5/17. You will also upload a PDF of your slides to Canvas the day before your presentation.

The Default Project

Recall that you don't have to do this project, but if you do something else, you need instructor approval. We'll approve anything within reason.

To do the default project:

1) Choose a question in marine biology or ecology that requires the use of a known gene or group of genes to answer. For example, nitrogen fixation is related to *nifH*, and you need to use *nifH* to answer a question about nitrogen fixation in UCYN-A. If there's no obvious gene, use COI or ribosomal RNA (16S for prokaryotes, 18S for eukaryotes). Those are guaranteed to be present in whatever organisms you choose in step 2.
2) Choose an organism that is relevant to your question from step 1. If you have trouble with this, you can go to NCBI and retrieve the sequence of your gene for an organism you're certain of. For example, we all love dolphins so you could go to NCBI and do a nucleotide search for "dolphin COI". *Caution: the search returns any record that contains dolphin COI. Your results are probably sorted by descending length, so your first very many hits will be to the entire mitochondrial chromosome that contains COI*

and many other genes. This record length is a bit > 16,000. By page 15 the lengths are 829, which is common for COI records. The moral is to always use common sense when doing bioinformatics. Whatever data a tool gives you, think about if it's sane before using it.

3) Collect the sequence of your gene for at least 20 other organisms besides the first one. You can do this using nucleotide blast or any other approach you choose. Include some organisms that are closely related to the original, and some that are more distantly related. For example, if you started with dolphin, then choose some other marine mammals and some non-mammals (maybe fish, maybe sponges, maybe whatever). But if you started with a eukaryote, collect only eukaryotes; if you started with a prokaryote, choose only prokaryotes.
4) Build a phylogenetic tree of all your sequences. Stare at the tree. That's a big part of data-heavy science: staring at a visualization of your data until it reveals its story.
5) Optional: any other computer analysis or some or all of your data.

Requirements

BIOL 145: At least 12 pages, double-spaced, not including source code or references.
CS 185C: At least 8 pages, double-spaced, not including source code or references.

The report ***must*** be structured as described below. This structure is required by almost all peer-reviewed (i.e. legitimate) scientific journals. There are 5 sections. Everything you say must go in the proper section.

- Section 1: Background
    - The question you're investigating
    - The rationale behind the study
    - Why it's important (i.e., broader context)
    - A brief review of the background literature on your topic
    - Your plan
        - What data
        - What analysis
        - Expected results (optional)
        - Hypothesis (optional)
- Section 2: Methods
    - Data
        - Where your data came from
        - How you got your data
        - (But don't put the actual data in this section)
    - Procedure
        - What analysis technique(s) you used
        - What analysis software you used
            - Was this from a public web site?
            - Was the software publicly available downloaded?
            - Did you use your own code?
        - If you used software, describe the code

        - Language
        - Number of lines
        - What it does
        - Don't include any source code. Don't include screenshots unless they show non-text graphics generated by your code.
- Section 3: Results
    - Discuss only the results here, no interpretation
    - Tables are good
    - No figures that are screenshots of text. That's annoying and hard to read. Text results should be text, either in the body of your doc or in a table.
    - Never include a table or figure you don't write about in the text of the results section. A figure needs to be explained.
    - Each table or figure should be accompanied by a brief descriptive title that is detailed enough that the figure can be understood on its own, without having to read the written explanation in the text of the results section.
    - Figures should have both axes labeled. If appropriate, the units should be listed.
    - Tables and figures are numbered independently (you can have a Figure 1 and a Table 1).
    - Tables have a descriptive title placed above. These titles should one sentence describing the table and identifying important abbreviations, if you have any.
    - Figures have what's known as a figure legend that is placed below the figure. These figure legends describe the point of the figure in one sentence. A few more sentences can be used to describe important specific points about the figure and to define abbreviations.
- Section 4: Discussion
    - Brief broad summary of the results: I did this to these data, I got these results, which let me to these broad conclusions
    - Interpretation of the data: what conclusions do you want your readers to have?
    - How do your results relate to the information you presented in your introduction.
    - How you could improve on the methods if you got a do-over?
    - If you had the time and the money, how would you do more research into this issue?
- Section 5: References
    - We expect at least 5 primary peer-reviewed journal articles in your paper (see below)
    - If you use a Word plug-in like Zotero, this section is built for you automatically. If you write the References section yourself, follow these rules.
    - When you cite a book, article, or web site in the main text of your paper, do it like this
        - This habitat seems to generally lack UCYN-A1 and has environmental conditions that clearly differ from the tropical/subtropical oligotrophic open ocean during most times of the year (Chavez et al., 2002).
    - Or like this
        - This habitat seems to generally lack UCYN-A1 and has environmental conditions that clearly differ from the tropical/subtropical oligotrophic open ocean during most times of the year [1].

- Then in your references section:
  - Chavez FP, Pennington JT, Castro CG, *et al*. (2002). Biological and chemical consequences of the 1997-98 El Nino in central California waters. Progr Oceanogr 54: 205–232.
    - References must appear in alphabetical order by first author. If there are 4 or more authors, just name the first 3, and then write "*et al.*" in italics, with a period after al. It's Latin for "and others".
- Or like this:
  - (1) Chavez FP, Pennington JT, Castro CG, *et al*. (2002). Biological and chemical consequences of the 1997-98 El Nino in central California waters. Progr Oceanogr 54: 205–232.
    - Citations in the text should be in increasing numerical order.
- No plagiarism (see below)
- No cite-cites (see below)

More rules for the written report:

- Figures may not contain screenshots of text.
- Everything you write must be in your own words. If you paste something that someone else wrote, *even if it's in quotes and cited*, that's plagiarism. See the syllabus for consequences. We use sophisticated software to check this, so please don't burn yourself down. You won't be graded for the beauty of your English or perfection of your grammar.

The Oral Presentation:

Presentations will happen on the last two days of lecture, and during the final exam time slot. Presentations will be 10 mins with 5 mins for questions. The presentation schedule will be randomly generated. Your presentation deck may be PowerPoint or PDF; it is to be uploaded to Canvas one day before your presentation at the latest. You can't change it after the submission date. Those who present on the first day get a 5% score bonus. During your presentation, your deck will be screenshared from your instructor's computer.

If your project includes code, your slides should describe the code but should not list the code.

Screenshots of text are not allowed.

Plagiarism

is defined here: http://www.sjsu.edu/cs100w/policies/plagiarism.html. All students at SJSU are expected to understand these rules. If you rewrite a substantial amount someone else's work, replacing words with synonyms, that's plagiarism. Example:

Original work:
"But soft, what light through yonder window breaks?
It is the east, and Juliet is the sun.
Arise, fair sun, and kill the envious moon." -Shakespeare, Romeo and Juliet

Definitely plagiarism:
“Ssssh! What’s that light coming through the window over there?
That’s where the sun comes up, and my new girlfriend is the sun.
Get up, beautiful sun, and murder the jealous lunar ball.”

If you copy someone else’s words, *even if you cite correctly*, that’s still plagiarism. The general rule is that *everything* has to be in your own words. We know how to find out if you did that, and we know how to find out if you used ChatGPT. Please don’t.

A “cite-cite” is when you cite a source to support a claim, but your source doesn’t support the claim either, your source just cites *someone else* who (hopefully) supports the claim. The source that you cite must provide the evidence directly: it must be a *primary* source.

Example: In a 2018 article, I wrote the following: “In 2002, Hebert proposed cytochrome c oxidase I (COI) as a standard for molecular barcoding of animals[4].” My reference #4 was the article where Paul Hebert made that proposal. If you want to say that Hebert made that proposal in 2002, cite his article, not mine. This is because if someone wants to find the original work, they shouldn’t have to find my article, figure out where I make my claim about Hebert, look up my reference #4, and so on.

“Cite, don’t cite-cite.” - Jon Zehr, to Phil
“Cite, don’t cite-cite, and definitely never cite-cite-cite.” – Phil, to you

One caveat is if you cite a peer-reviewed review paper that summarizes the literature on your topic of interest. It is ok to cite a review paper if you cite it in the following way: (reviewed in deVries et al., 2021). Note that these review papers are NOT primary sources and will therefore not count towards your reference count of 5 primary sources.

Original software: If you’re taking this course for CS credit, you have to write some code. If you’re taking it for BIOL credit you are welcome to write some, and that will certainly help improve your grade, but no pressure … you can get an A without writing any code. The software can be in whatever language you like, except Haskell or Fortran. With a project of this size, it’s hard to think of a software component that you can write in reasonable time and will substantially advance your research. So don’t worry about advancing your research, just get programming experience writing something that’s roughly related to bioinformatics. Maybe a pairwise aligner, or a protein alignment scorer, or something like that. Demonstrate that your code works by inputting some of the sequences you collected. Talk to Dr. Heller if you don’t know what code to write. Your code should be at least 300 lines, including comments.

Grading

Your project is worth 30% of your grade. You will be evaluated on this 250-point scale:

Written (**180**)	Category **(points)**	Emerging (**min points**)	Developing (**medium points**)	Mastered (**max points**)
	Clarity **(15)**	Many unclear passages that prevent understanding.	A few unclear passages that prevent understanding.	No unclear passages that prevent understanding.
	Organization **(15)**	Writing is haphazard and disjointed with weak organization.	Organization is for the most part clear and coherent.	Organization is consistently clear and coherent.
	Scientific accuracy **(40)**	Much of the information presented is inaccurate.	Most information presented is accurate.	All information presented is accurate, demonstrating a good understanding of the subject.
	Use of Bioinformatics approaches **(60)**	Inappropriate bioinformatics approaches were used or bioinformatics approaches were used incorrectly.	For the most part, appropriate bioinformatics approaches were used and they were used correctly.	Appropriate bioinformatic approaches were used and they were used correctly in all cases.
	Interpretation of data **(50)**	Data not interpreted correctly or student showed a lack of understanding of the correct interpretation.	Data interpreted correctly for the most part, or interpreted the majority of the data correctly, and displayed an understanding of the correct interpretation of the data.	Data are interpreted correctly and a deep understanding of the correct interpretation of the data is demonstrated
Presentation (**70**)	Oral presentation **(70)**	Presentation was unclear, did not flow well, and did not observe time limitations.	Presentation was unclear or did not flow well, or did not observe time limitations.	Presentation was clear (the presenter gave a clear introduction to the question and its relevance, described the methods and results well, and interpreted the data well), flowed well, and observed time limitations.

