Biol/CS 123B Term Project

Fall 2023

The Default Project:

You don’t have to do this, but if you do something else you need my approval. I’ll approve anything within reason. Whatever you do, your topic may not be the same as your 123A project.

To do the default project:

1. Choose some biological, medical, or environmental issue that is related to a known gene or group of genes. For example, nitrogen fixation is related to *nifH/K/D/N*. COVID-19 vaccines are related to viral spike proteins.
2. Collect at least 25 sequences (but more is better) of the gene, from various closely related species. Call this your “basic data set”.
3. Collect at least 25 sequences (but more is better) of the gene, from a clade that is related to the basic set. Call this your “related data set”. For example, if your basic data set is from primates, then your related data set might be rodents.
4. Apply 123A techniques to your data to get a sense of the diversity of your chosen gene(s). Build multiple-sequence alignment(s) and phylogenetic tree(s).
5. Use 80% of the sequences in your basic data set to build a HMM. Use 80% of the sequences in your related data set to build a HMM.
6. For each of the 20% of sequences that weren’t used to train an HMM, compute the log-Viterbi score on each HMM.
7. If you’re taking CS 123B (not BIOL 123B), you have to write at least 200 lines of code in any language, to help collect and/or analyze your data. If you’re taking BIOL 123B, code is optional (strongly recommended for MSBI students and Bioinformatics minors). The code should be listed in an appendix, not in the main body of your report

The Written Project Report: Due 11:55 PM on Monday November 20 by file upload to Canvas. Word only, no PDF.

BIOL 123B: At least 15 pages, double-spaced, not including source code or references.

CS 123B: At least 10 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 and nowhere else.

* Section 1: Background
  + The problem/issue you’re investigating
  + Why it’s important
  + Your plan
    - What data
    - What analysis
    - Expected results
    - Hypothesis
* Section 2: Methods
  + Data
    - Where it came from
    - How you got it
    - (But don’t put the data itself in this section)
  + Procedure
    - What analysis technique(s) you used
    - What analysis software you used
      * Public web site?
      * Publicly available downloaded software?
      * 123B HMM?
      * Your own code?
    - If there’s software, describe the code
      * Language
      * Number of lines
      * What it does
      * Source code goes in an appendix
* Section 3: Results
  + Raw results, 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.
* Section 4: Discussion
  + Broad summary: I did this to this 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?
  + Was your hypothesis supported?
  + How you could improve on the methods if you got a do-over
  + If you had the time, how would you do more research into this issue
* Section 5: References
  + If you use a Word plug-in like Zotero, this section is built for you automatically.
  + When you cite a book, article, or web site in sections 1-4, 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, Ryan JP, Michisaki RM, Schlining B et al. (2002). Biological and chemical consequences of the 1997-98 El Nino in central California waters. Progr Oceanogr 54: 205–232.
  + Or like this:
    - (1) Chavez FP, Pennington JT, Castro CG, Ryan JP, Michisaki RM, Schlining B et al. (2002). Biological and chemical consequences of the 1997-98 El Nino in central California waters. Progr Oceanogr 54: 205–232.
  + 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.

The Oral Presentation:

Presentations will happen on the last 4 days of lecture, and during the final exam time slot. Presentations will be 12” - 15”, with a few minutes of questions at the end.

The presentation schedule will be randomly generated. Your presentation deck may be PowerPoint (without animations) or PDF; it is due at noon on the day before your presentation. Anyone who presents on Nov 21 (the first presentation session, Thanksgiving week) gets extra credit. You can volunteer for Nov 21.

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.

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 animals4.” 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 me

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

Grading

Your project is worth 20% of your grade. You will be evaluated on the following 200-point scale:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Written | Category **(points)** | Low | Medium | High |
| Clarity **(30)** | Many unclear passages that prevent understanding. | A few unclear passages that prevent understanding. | No unclear passages that prevent understanding. |
| Organization **(20)** | 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 **(30)** | 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 **(40)** | 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 **(40)** | Student did not  interpret data correctly  or students showed a  lack of understanding  of the correct  interpretation. | Student interpreted  data correctly for the  most part, or  interpreted the  majority of the data  correctly, and  displayed an  understanding of the  correct interpretation of the data. | The student  consistently  interpreted the data  correctly and  displayed a deep  understanding of the  correct interpretation  of the data. |
| Presentation | Oral presentation **(40)** | 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, flowed well, and observed time limitations. |

More on grading

* Attendance at all presentation sessions is mandatory, unless you have a medical excuse or a genuine emergency. “I had to study for my other class” is not an emergency.
* A simple pass/fail quiz may be given in any presentation session. Quiz grades affect project grades.

Optional reading for you: Below is a fake article that demonstrates all the principles of good writing. If you *hate* writing reports, these principles will reduce your suffering. If you love writing them, these principles will increase your enjoyment.

Grape jelly as a novel enhancement to peanut butter sandwiches

Philip Heller 1

1Department of Computer Science, San Jose State University

Introduction

Peanut butter is commonly described as the best food ever [1][2][3][4]. Fossils indicate that dinosaurs made peanut butter sandwiches as early as 60 million years ago [5]. Ancient images of peanut butter can be found in European cave paintings [6], tomb hieroglyphs in Egypt [7], the Dholavira complex in India [8], and recently discovered stone huts on Mars [9].

Unfortunately, peanut butter sandwiches are sticky and dry. In a 1922 study, 98% of American elementary school children described peanut butter sandwiches as “somewhat sticky” or “very sticky” [10]; similar results were reported in 1990 [11]. Attempts at improvement have not been rewarding. Beginning in the 1940s, a series of experiments with various inorganic enhancements produced consistently low satisfaction; enhancement materials included gravel in 1946 [12], crayons in 1955 [13], and salt in 1976 [14]. The search for an inorganic enhancement was discontinued in 1980 for humanitarian reasons, after the infamous peanut-butter-and-styrofoam study at Stanford [15]. Beginning in the 1960s a separate line of inquiry, involving organic enhancements, produced moderately better results and fewer casualties. Enhancements included chutney in 1962 [9], mustard in 1977 [11], chewing gum in 2004 [12], and feathers in 2010 [13]. The most promising advancement came in 2012 with the jam experiments of Professor J. Ellie Bean [14]. Bean reported that 61% of American elementary school children described peanut butter sandwiches with strawberry jam as “somewhat sticky” or “very sticky”.

These results indicate that more research is required. I hypothesize that if jelly were to be added to traditional peanut butter sandwiches, satisfaction could be further improved. This appears to be a completely novel solution, with no previous mention in the literature. Here I report promising results of enhancing with grape jelly. The recipe for this enhancement, which I call “Peanut Butter and Jelly” or simply “PBJ”, is publicly available at <https://github.com/PBJ>.

Methods

Peanut butter was spread on one slice of bread and grape jelly was spread on another slice. The two slices were then pressed together and cut diagonally.

100 children from the Sesame Street Elementary School in San Jose were fed a PBJ. A control group of 100 children from the same school were fed peanut butter sandwiches without jelly. All participants were then asked to describe their sandwich as “Not Sticky”, “Somewhat Sticky”, or “Very Sticky”.

Results

Survey results are shown in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Not Sticky | Somewhat Sticky | Very Sticky |
| Contol Goup | 1 | 73 | 26 |
| PBJ Group | 90 | 10 | 0 |

Table 1 – Survey Results

Discussion

99% of control group members and 10% of PBJ group members described their sandwiches as “Somewhat Sticky” or “Very Sticky”. These results demonstrate that compared to consumers of the proposed PBJ sandwich, consumers of a traditional peanut butter sandwich are 9.9 times more likely to experience their sandwich as sticky.

I conclude that jelly is an effective enhancement to the traditional peanut butter sandwich, and is a significant improvement over all other reported enhancements.

Further research may identify other jelly flavors that produce comparable or superior results.

References

. . .

Things to Notice

*Peanut butter is commonly described as the best food ever [1][2][3][4].*

I can't say “peanut butter is the best food ever”. That’s a personal/subjective statement. Scientific writing requires facts, not opinions. But if the opinion is broadly held, and the popularity of the opinion is documented (as in references 1-4), then *that’s* a fact.

In Season 1 of “Game of Thrones”, a servant tells Daenerys, “The last dragon died a thousand years ago, Khaleesi. It is known.” This *it is known*, with the implication that it is not to be questioned, runs opposite to the scientific spirit where anything may be questioned using the scientific method.

*Fossils indicate that dinosaurs made peanut butter sandwiches as early as 60 million years ago [5]. Ancient images of peanut butter can be found in European cave paintings [6], tomb hieroglyphs in Egypt [7], the Dholavira complex in India [8], and recently discovered stone huts on Mars [9].*

The rest of the paragraph supports the opening sentence. The supporting evidence is presented in historical order; this part of the article tells the story of peanut butter.

*Unfortunately, peanut butter sandwiches are sticky and dry.*

The first paragraph showed that peanut butter by itself is great. This paragraph introduces a context (sandwiches) where the great thing isn't so great. The job of the paragraph is to make readers care about the solution represented by my research. But it’s not time to talk about the solution yet. First I make readers want there to be a solution.

*In a 1922 study, 98% of American elementary school children described peanut butter sandwiches as “somewhat sticky” or “very sticky” [10]; similar results were reported in 1990 [11]. Attempts at improvement have not been rewarding. Beginning in the 1940s, a series of experiments with various inorganic enhancements produced consistently low satisfaction; enhancement materials included gravel in 1946 [12], crayons in 1955 [13], and salt in 1976 [14]. The search for an inorganic enhancement was discontinued in 1980 for humanitarian reasons, after the infamous peanut-butter-and-cardboard study at Stanford [15]. Beginning in the 1960s a separate line of inquiry, involving organic enhancements, produced moderately better results and fewer casualties. Enhancements included chutney in 1962 [9], mustard in 1977 [11], chewing gum in 2004 [12], and feathers in 2010 [13].*

Here I present the trouble. Earlier I told the story of the topic (peanut butter). Here I tell the story of the problem. Notice that there are 2 lines of investigation (inorganics and organics), and their timelines overlap. If I present the studies in chronological order, I’ll have to describe a couple of inorganic experiments, then some organic ones, then another inorganic, back and forth from 1946 to 2010. My story will be historically accurate, but confusing. 8 experiments is a lot, and I should try to impose some simplifying organization if I can. The distinction between inorganic and organic enhancements is natural.

Think about the outline of this text as it is:

* Enhancements
  + Inorganic
    - Gravel
    - Crayons
    - Salt
    - Cardboard
  + Organic
    - Chutney
    - Mustard
    - Chewing gum
    - Feathers

Compare to the outline of a purely historical order:

* Enhancements
  + - Gravel
    - Crayons
    - Chutney
    - Salt
    - Mustard
    - Cardboard
    - Chewing gum
    - Feathers

The first outline shows how its story is better organized. Sometimes historical order isn't the best order – you have to think about the organization that best instructs your readers.

*The most promising advancement came in 2020 with the jam experiments of Professor J. A. Emerson [14]. Emerson reported that only 61% of American elementary school children described peanut butter sandwiches with strawberry jam as “somewhat sticky” or “very sticky”.*

The Emerson experiment is the state-of-the-art. My work has to be significantly better. But the shortcomings of Emerson’s work need to be presented respectfully.

Here ends the literature review. Hopefully my readers are interested in the sub-sub-field of peanut butter, and care about the important problem of sandwiches. They have spent their valuable time reading up to this point. Now I owe them a significant solution to the problem.

*These results indicate that more research is required.*

After the literature review I put all that prior work in the context of the research I am about to present. This sentence obeys Rule 2c: it logically follows from what came just before.

*I hypothesize that if jelly were to be added to traditional peanut butter sandwiches, satisfaction could be further improved.*

If the research is hypothesis-based, this is a good place to put the hypothesis. The general form of this sentence is “I hypothesize that a certain solution (to the problem I just presented and made you care about) will be effective.

In a real article, you would give more extensive information about your idea. But remember that this isn't the place to go into full detail. Too much detail will lose your audience. Give them the big idea; they know that details will follow in the Methods section.

*This appears to be a completely novel solution, with no previous mention in the literature.*

Research has to be novel and significant. Here I claim that it’s completely novel. In this case by definition I can't cite anything to support the claim. If my solution were partly novel (maybe an extension of someone else’s idea), that’s fine. I just have to mention whose work I am extending; that work would already have been presented in the literature review.

*Here I report promising results of enhancing with grape jelly. The recipe for this enhancement, which I call “Peanut Butter and Jelly” or simply “PBJ”, is publicly available at* [*https://github.com/PBJ*](https://github.com/PBJ)*.*

The remainder of the paragraph sets up the remainder of the article. Traditionally this is also where you state where your open-source material can be found.

Notice what isn't said. You might have been told that all sections should end with connecting material: something like, “This section has described peanut butter in the context of sandwiches, and has listed various attempts to make peanut butter sandwiches less sticky. In the following sections I present my methodology, results, and discussion”. This wastes your readers’ time. They just read the Introduction, and they can see what sections are to follow. Telling them the obvious serves no purpose and violates Rule 4b. (“My writing instructor told me I always need connecting material” is not an explanation. It’s like “It is known.”)

Connecting material should be about ideas, not about the order of your sections. The big idea of the Introduction is, “Peanut butter sandwiches need something more.” The big idea of the rest of the article is, “Jelly!”. In terms of ideas, all the necessary connecting material is in the sentence “*Here I report promising results of enhancing with grape jelly.”*

*Methods*

The point of science is to report reproducible results. The Methods section is where you tell readers how to reproduce your results. It’s a recipe. Anything not directly about what you did belongs somewhere else.

*Results*

Now your readers know what you did. Here you tell them what you observed when you did that. Anything other than your direct observations belongs somewhere else.

*Discussion*

Here you tell your readers how you interpret the data.

*99% of control group members and 10% of PBJ group members described their sandwiches as “Somewhat Sticky” or “Very Sticky”.*

Notice that I don’t just write sentences that present the raw numbers in the table. Those numbers have already been presented. The sentence above organizes the raw numbers in a way that leads to an interpretation. It obeys Rule 2c, as it is a logical outcome of the table.

*These results demonstrate that compared to consumers of the proposed PBJ sandwich, consumers of a traditional peanut butter sandwich are 9.9 times more likely to experience their sandwich as sticky.*

This sentence follows logically from the previous one.

*I conclude that jelly is an effective enhancement to the traditional peanut butter sandwich, and is a significant improvement over all other reported enhancements.*

Again, this sentence follows logically from the previous one. From the table (containing only numbers) to this sentence, there is a progression from raw data to progressively higher level interpretation. The sentence above has no numbers at all.

*Further research may identify other jelly flavors that produce comparable or superior results.*

This sentence is connecting material, not to subsequent parts of this article but to future work. As with all connecting material, don’t put it in just because it is known that you should. The Discussion section ends cleanly if the last sentence is omitted.