Software Tools (BINF*6210) – Assignment #1 (15%)

DUE DATE: Friday October 10, 2024 5:00 PM to CourseLink Dropbox

---

Overview: In this assignment, you will pose your own research question and adapt skills and concepts learned in class to solve new problems. You will pose a question related to the area of biodiversity research, use data from the BOLD database, and write code in R to answer your question. To ensure your success and minimize stress, I suggest starting with a simple, exploratory question, and do a good job answering it. Complete an end-to-end assignment draft, including all components. You could then add onto that, time permitting, to improve the novelty, scope, or creativity of your project. Completing this project contributes towards achieving the following five course-level learning outcomes:

1. obtain data from key databases relevant for bioinformatics and to understand the sources and limitations of these data (with focus on BOLD)
2. filter, manipulate, analyze, and visualize bioinformatics data (with emphasis upon filtering and manipulating data frames in R and preparing figures, which may be simple for Assignment #1)
3. conduct reproducible analyses and use software tools for version control and collaboration (first part of this learning outcome only in Assignment #1)
4. understand and apply selected algorithms commonly used in bioinformatics, including sequence alignment and clustering (with emphasis here on analyzing and interpreting distribution patterns in BINs, i.e. genetic clusters)
5. adapt the above skills to learn new tools and conduct new analyses not explicitly covered in class (you should access software documentation and incorporate at least one new function or type of figure)

Requirements and Project Steps for your story board:

1. Title slide: Title, name

2. Introduction (2 or 3 slides): Pose a question that interests you about the biodiversity of a particular taxonomic group of multicellular life with representation in the BOLD database. Here are some candidate project ideas, and you could choose one of the following:

a) Investigating Geographic Distributions

- If you are a beginner in programming in R, I recommend that you choose a simple, exploratory question to start and then take it from there. For example: How many species or BINs (Barcode Index Numbers) of your taxonomic group of interest have been DNA barcoded from, for example, Canada vs. Norway (or other areas of interest to you)?
- After that, as a next step in developing your project, you might then ask a further question: How similar is the BIN composition between these two regions for taxonomic groups with contrasting biological traits? (Tip: If that is your question, then check out the `vegdist()` function in the package `vegan` for a variety of options for quantifying community similarity.)

b) Exploring Sampling Completeness and Biases

- Another potential question would be: How well sampled is my taxonomic group of interest?
- Or, how many more samples are needed to achieve near comprehensive sampling for my group?
- Or, how does sampling completeness compare between two or more taxonomic groups or between two or more geographic regions?

c) Formulate and Test a Hypothesis

- You may instead choose to develop a hypothesis-based project. For example: Do the geographic ranges of BINs vary with latitude? You might hypothesize that higher environmental variability at higher latitudes, compared to the tropics, leads to species having larger geographic ranges in temperate and polar areas, as they can tolerate a range

of environmental conditions. Is this hypothesis supported for your taxonomic group? This hypothesis could be tested by evaluating whether latitude (e.g., median values among specimens within each BIN) is correlated with geographic range of BINs (e.g., latitudinal range).

You may choose any of the examples provided above or come up with your own question. You will pose your question clearly. Then, explain why your question is interesting. Clearly state whether your project is exploratory or whether you are testing a specific hypothesis. You will include written sections right inside your .R or .Rmd file.

2. Choose a taxonomic group. You may choose any taxonomic group you are interested in that has data in BOLD, so long as the group contains at least 10 species or BINs and at least 100 records (i.e. at least 100 specimens, i.e. rows in the BOLD data download file). You may choose a taxon at any taxonomic rank that is suitable for your question (e.g., genus, family, etc.). For manageability and considering this assignment scope, I suggest choosing a group such that your total dataset size doesn't exceed 10,000 records (i.e. rows in your data frame). However, this is a recommendation and not a rule; you may feel free to analyze a larger dataset if you prefer to do so, but I suggest not to analyze a huge dataset, considering the intended modest scope of this assignment. You may wish to explore the Taxonomy Browser on BOLD hierarchically to help you choose a suitably-sized taxonomic group that interests you:

https://portal.boldsystems.org/

TIP: Note that the BIN algorithm is only applied to animal COI sequences on BOLD. If you wish to choose a different group of life (e.g., plants, fungi, macroalgae), you are welcome to do so. Simply be aware that BINs would not be available, so you would need to analyze other information (e.g., the taxonomic identifications, such as genus-level or species-level identifications).

3. Obtaining data. Use the website to download data, and then import it into R. Tip: Depending upon your question, you might wish to use a combination of taxonomic and geographic terms

for your API call. That would help to narrow the scope of your analysis if you choose a very species-rich or highly barcoded taxonomic group.

4. Write code to answer your question in a .R or .rmd file. Follow the "Best practices in R" resource on how to structure, format, and comment your code. If your project involves testing a hypothesis, then your code should involve a statistical test (such as linear regression if you are testing for an association between two continuous quantitative variables). Test that your code works as intended and is also well formatted before submitting!

5. Methods slide. Visualize your pipeline.

6. Figures. Your storyboard must include 2-3 figures, one on each slide, with as header what information the audience should extract from the figure. This could include an exploratory plot to investigate the variability of your data and to check for outliers or potential data errors. An example could be building a histogram to visualize the distribution of latitude values in your dataset. You would also have at least one figure to show your main result. For example, you could display a bar plot to show the number of BINs or species among several geographic regions. You could also consider exploring functions and packages for creating map figures. Again, I suggest starting simpler, thus ensuring you can generate relevant figures before the assignment due date. Then, time permitting, you could work on improving your report by generating more sophisticated or creative figures. All figures should have informative axis labels. Also, consider your colour, symbol, and pattern choices for clarity in conveying your message.

Please do not include extra figures in your storyboard. The maximum is three. If extra figures are encountered, the first three only will be graded. Your code can include code for more figures, but you will decide which of those are important for your storyboard.

7. Results and Discussion (2 or 3 slides). In these slides, describe and interpret what you discovered through conducting your project. What did you find? Did you find what you expected or not? Why might that be? What did you learn? Also, indicate any caveats about your study. For example, was the available dataset large enough to address your question? What would you do next if you were going to develop this work into a larger project?

8. Novelty and Creativity. For this assignment #1, it is expected that you will often adapt lines of code from example scripts from the course, showing that you understand what the code does and how to alter it for your own purposes. Beyond that, try to include at a minimum one project component that is novel in comparison with course materials provided to date. Examples of novel components could include: adding a statistical test, adding a multivariate analysis and visualization (e.g., PCA or other technique to show BIN community composition data among countries or geographic regions), inclusion of a map figure or other type of new figure, etc. Please note that it is not expected that anyone will do all of these things! Pick something of interest to you. Then, seek tutorials online, read the code documentation, and try out executing something new.

9. Acknowledgements slide. If you received project tips from others or help solving problems, include that information in this section. We will have a class activity to discuss your project ideas; so, if you receive advice that helped you to develop your project, indicate so in this section. Briefly, indicate who you talked to, the nature of the advice, and how this impacted your project. You may speak to other class members. It is NOT permitted to complete the assignment for someone else or to copy blocks of code from others. If someone helped you to get unstuck when you were facing an error message, then indicate who and what you learned from this. (To clarify: You ARE allowed to talk to others, but you are NOT permitted to let them fix the problem for you without you actually understanding what is going on. Write about: What was causing the error message, and what did you learn during the process of obtaining help to fix it?)

10. Reference slide. If you cite any sources from the scientific literature, cite them here. You may wish to consult 2-3 papers from the literature to help you to develop your idea and/or interpret your results. You would include such references as an in-text citation in the relevant sentence of your assignment, i.e. introduction or discussion (e.g., Xu et al. 2020). Also, list the full reference here at the end. Be consistent in your formatting. If you used any specific online tutorials or a specific StackOverflow posting, for example, also include those here.

10. Have fun!

Overall Scope and Length: Your final storyboard is expected to be approximately 10-13 slides in length. Please note that longer is not considered better! (As we progress, e.g., by learning iteration techniques, our code actually gets shorter!) Convert your slides to a pdf file, and save it in the "doc" folder of your project folder structure. Your folder structure for this assignment should at least have a R, doc, and figs folder, but you can include the other suggested folders if your project requires this. Submit your folder structure as a zipped file to CourseLink, to ensure that everything is contained in one file that has all the components to run, evaluate, and interpret your assignment.

Tips for Success - Checklist

- ✔ Start your project early to avoid stress at the end.
- ✔ If you get stuck, first try to resolve the issue yourself: read the error messages carefully, Google your error message, read the function documentation, see online tutorials, etc. Then, after all those things, you may reach out to others for help to get past your issue. Please note that if you would like to request help from either the instructors or your peers, we recommend to do so at least 48 hours before the due date to enable reasonable time to respond. We recommend to complete a draft early and use the last days for proofreading and refining your assignment. We recommend to submit early to avoid technical difficulties at the last minute.
- ✔ Before submission, select the entire script (e.g., you can use the 'Control-A' keyboard shortcut) and run the entire script in one go. It should run without error messages. If you get an error message, fix that issue and then repeat until the full script runs. We also encourage you to swap code with a peer (reciprocal arrangement) and you can run one another's scripts and ensure they run. If you do so, this should be mentioned in the acknowledgements section.

- ✔ A storyboard is slightly different from a presentation. It uses the same format as a presentation (i.e., your choice of presentation software), but you have the flexibility of using more traditional bullet points to explain your decisions.
- ✔ If you are struggling to decide what question to work on, then I suggest you consider the sampling topic. While we are analyzing specifically BOLD data here, topics such as sampling completeness and bias are very general scientific issues. Thinking about this topic would be relevant for many of your future projects and many types of datasets.
- ✔ If you are struggling to decide what taxonomic group to choose, why not choose something you know nothing about and learn something completely new to you? Check out the taxonomic hierarchy through the Taxonomy browser and pick something unfamiliar.

Resources

Here are a few resources you may find helpful:

https://www.r-graph-gallery.com/all-graphs.html

https://rdrr.io/rforge/vegan/f/vignettes/FAQ-vegan.Rmd

https://peat-clark.github.io/BIO381/veganTutorial.html

https://cran.r-project.org/web/packages/iNEXT/vignettes/Introduction.pdf

https://kembellab.ca/r-workshop/biodivR/SK_Biodiversity_R.html

Submission Instructions:

Submit your assignment as one zipped file to the designated Dropbox folder on CourseLink.

Software Tools (BINF*6210) – Assignment #1 Grading Rubric

Context and Tips for Success:

- This grading rubric is intended to help you to be as successful as possible on Assignment #1. Instructor judgement will also be applied during grading. For example, high-quality or novel project components not explicitly mentioned below would also be considered during grading.
- Note that each level is cumulative. That means you should meet the criteria of the lower level to move on to receiving a grade at the next level. Aim to complete all components at the first level (70%) first. Completing an end-to-end assignment first will mean you can submit a passing assignment on time. You don't want to be in the situation where perhaps you spent all your time developing a great scientific question but then end up facing a challenging error message with your code the night before the due date!
- Then, after you check off everything in the 70% category, work on your assignment further to achieve a higher grade by improving the quality, scope, novelty, or creativity of your project. This table is populated with examples of components that would earn a particular grade.
- I have started the rubric at 70%, because I am assuming that everyone wants not only to pass this specific course (65% is the passing grade for an individual course at the graduate level at the University of Guelph). I assume you also want to do well overall, pass your graduate degree (at least 70% overall GPA required), and achieve a level of competency that will help you in your further studies and future career. You can do it! Your grade may fall in between two levels. This rubric is intended as a guide only to help you to be successful.
- I hope that you enjoy completing this assignment!

Assignment Component	**Good (Grade of 70%)**	**Very Good (Grade of 80%)**	**Excellent to Outstanding (Grade of 90-100%)**
Storyboard (/10)	7/10 *all the components of the storyboard are included *a clearly phrased question is posed (a simple, clear question is fine) *a statement is made about why this question is interesting *reader can follow what the project is about and what is the objective *writing is understandable, with few grammatical errors	8/10 *novel question is posed (may use example question from instructions, but project would go beyond what is covered in example scripts) *reference is made to the literature (e.g., 1-2 relevant references cited)	9-10/10 *novel question is posed *reference is made to the literature (e.g., 1-3 highly relevant references cited) *critical thinking displayed (e.g.,: outlining expectations under two contrasting hypotheses; that's just one example)
Code – Part 1 (Data Exploration) /30	21/30 *code is contained in one zipped file with all components * you followed all relevant best practices *summarizes key variables of interest (e.g., mean, range of numerical variables) *check for serious data entry errors in any variables you are analyzing (e.g., check range of latitude data) *most or nearly all lines of code may be adapted (correctly) from example scripts	24/30 *may include summary statistics and plots to show the distribution of relevant variables *includes thoughtful, deliberate approach as to how you will treat NAs in your variables of interest for analysis *may include some code that is newly written (more than minor adaptation of provided example scripts)	27-30/30 *strong critical thinking demonstrated to explore potential errors or biases in the source data *may include novel types of data summaries or graphics (beyond those covered in example scripts) to show the distribution of data *includes some novel code

Code – Part 2 (Analysis to Address Question) /30	21/30 *includes an analysis to answer your main question (approach could include summary statistics, graphics, statistical test, or a combination)	24/30 *includes an original component, such as a statistical test or a graphical approach to answering your question (an example could be a map figure) *includes some lines of code that are newly written (more than minor adaptation of provided example scripts)	27-30/30 *includes a component with strong novelty or creativity *uses one or more packages or functions not included in example scripts *may include a sophisticated graphical or statistical approach (e.g., multivariate analysis)
Quality of Figures /20	14/20 *storyboard includes 2-3 figures *figures may be simple (e.g., histogram, boxplot, barplot), but should be suitable for type of data being presented *axes should be clearly and correctly labeled *symbol size/style should be easy to understand	16/20 *includes at least one type of novel figure (could be a map, Venn diagram showing overlap in BIN composition among geographic regions, etc.) *colour, symbol shape, and/or pattern may be used in one or more figures, with clear meaning	18-20/20 *all figures are polished, visually appealing, and clearly communicate the data and results *creativity or novelty expressed in one or more figures

Discussion & Conclusion /10	7/10 *clearly state what you found, answer your original question *state whether your results were aligned with your expectations or not, why might that be?	8/10 *comment thoughtfully on any caveats of your study (e.g., limited sample size, sampling bias, etc.) *comment on what would you do next if you were to expand this study	9-10/10 *excellent critical interpretation of your findings and caveats *outlines next steps for this research (e.g., did the findings generate a hypothesis that could be tested in the future?) *refers to the literature (e.g., 1-2 relevant references cited)
*Acknowledgements section expected to be present for academic integrity, if you talked to anyone else about your assignment. *Reference list must be included if you cite any references. You would also include citations to any vignettes that you consulted, online tutorials, specific StackOverflow postings, etc.			
Total Graded out of 100% and Valued at 15% of Course Grade			

Best Practices	Complete?
Visualize pipeline	
Ensure all steps are included and commented appropriately	
Include all information on dataset, packages used, and choices made	
Cite all packages used and any other pipelines consulted	
Consistent code formatting	
If necessary, start with a small dataset to test code before scaling up	
Include steps to explore and filter dataset	
If necessary, include steps to optimize parameters or do sensitivity analysis	
Create figures as checks within the code	
Include performance checks to ensure code is working properly (e.g., check if NAs are removed after the code)	
Include reproducibility checkpoints (e.g., sample size after filtering for a standard data set)	
Open pipeline up for review	
Keep track of changes and versions	
Make pipeline publicly available	
Ensure pipeline is general enough to be used on multiple datasets	
Make the pipeline as efficient as possible (loops, down-sampling, multiple cores, …)	