3Software Tools (BINF*6210) – Fall 2023– Assignment #2 (15%)

DUE DATE: Friday, October 25, 2024, 5:00 PM Eastern Time

Overview: Over the course of completing Assignment #2, you will create your own objective or question, and you will adapt and build upon skills and concepts learned in class to solve new problems. The three main skills you will use and further develop during this assignment are:

- **accessing data from public databases**
- **analyzing DNA sequence data**
- **accessing and using one new R package beyond class materials.**

Below, you will find five project themes from which to choose, detailed assignment instructions, helpful links, and the grading rubric at the end. All of the below projects have a component that can involve correctly adapting class example scripts and a component of building onto that foundation by using a new package, method, or visualization type. You may choose to include BOLD or not. If you include BOLD as one of your data sources, you must include at least one other data source.

Before you select your project theme, I suggest reading over all of the project ideas carefully, and also check out the provided links. It is useful to be aware of this array of resources, even for projects you don't choose. For example, you may get ideas about new kinds of visualizations you haven't encountered before, which you may apply in Assignments #3 or #5 or in your future work beyond this course.

To ensure your success and minimize stress, start this assignment early. This project is meant to be completed in regular small increments over two weeks. Assignment #2 serves as both a learning opportunity and an assessment tool. This assignment also serves as a stepping stone towards completing larger projects (e.g. #5 and later BINF*6999 or thesis). 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 emphasis upon NCBI)
2. filter, manipulate, analyze, and visualize bioinformatics data (with emphasis here on manipulating DNA sequence data)
3. conduct reproducible analyses and use software tools for version control and collaboration (first part of this learning outcome only in Assignment #2)
4. understand and apply selected algorithms commonly used in bioinformatics, including for sequence alignment and clustering
5. adapt the above skills to learn new tools and conduct new analyses not explicitly covered in class (you should access software documentation and tutorials and incorporate at least one new function or type of figure beyond class materials)

Themes and Example Projects

Please choose one of the five following project themes:

a) Supervised Machine Learning and Classification Using Sequence Features

- Using two genes (e.g. from NCBI, and you could use BOLD too if you wish) for a single taxonomic group, build and test a classifier for separating these genes. As part of your project, you may adapt the randomForest example script from class. Be sure to consider all steps and arguments to functions used; don't just copy/paste: adapt the script as suitable for your specific problem and data set. If you do use the randomForest example script as a key resource, then you should also add a novel component. For example, you may wish to try another classification method for comparison with random forest. In the written portions and in the commenting, show that you understand what the methods are doing, not only that you can get them to run.
- This project would be a great choice for those interested in going into more depth in understanding and applying machine learning algorithms beyond the example script.
- Tip: If your classification problem ends up being easy to solve (e.g. if you easily get 100% classification success on your training and validation data with simple sequence features), then I suggest to report that outcome, but you may also wish to try out a second data set that could pose a more interesting and challenging classification problem. For example, you might choose two genes that might be expected to have similar properties (e.g. COI and cytb are both protein-coding, mitochondrial genes; for the same taxa, these might be expected to have similar properties – you could test whether these can be separated using simple features).
- Tips: I suggest using a modest-sized dataset for this theme (e.g. 1,000-10,000 sequences total; these are recommendations, not hard boundaries). Consider trying various k-mer lengths to assess impact upon your classifier.
- An alternative project idea under this same general theme would be to build a taxonomic classifier, rather than gene classifier. So, you would be assigning sequences to taxonomic groups. Another project idea that falls under this theme would be to see if you can separate categories of genes (e.g. nuclear protein-coding vs. nuclear ribosomal RNA genes).
- Some helpful links relating to this theme:
- https://www.kaggle.com/camnugent/introduction-to-machine-learning-in-r-tutorial
- https://www.datacamp.com/community/tutorials/machine-learning-in-r
- https://lgatto.github.io/IntroMachineLearningWithR/an-introduction-to-machine-learning-with-r.html
- https://www.machinelearningplus.com/machine-learning/caret-package/

b) Unsupervised Machine Learning or Clustering of Sequences

- Do different genes exhibit different clustering patterns within the same taxonomic groups? Such a project has relevance for determining if different genes are more or less suitable as

markers for future classification applications and also can reveal differences in evolutionary pressures (e.g. functional constraint) among genes.

- This project would be a great choice for students interested in both machine learning and genetics.
- Steps in this project: For a taxonomic group of interest (I suggest choosing a genus), obtain DNA sequence data for two different genes from NCBI (and you could use BOLD too if you wish). Using either extracted sequence features (e.g. k-mer frequencies) or alignments (if the dataset size is modest), calculate a pairwise distance matrix among sequences for each gene. Then, cluster your sequences for each gene. If you wish, you may build upon our example script performing clustering of sequences using functions available through the DECIPHER package. You would then build upon this in a novel way beyond the example scripts, for example by evaluating and comparing cluster strength using internal measures (e.g. Dunn index or Silhouette index). You could consider including a Silhouette plot as one of your visualizations.
- Tips: Your sample size of sequences should be relatively similar between genes. If one gene has a lot more data than the other, then you may consider eliminating highly similar sequences for the gene with more data. You may also consider including a relatively similar number of species between the two genes.
- Tip: I suggest using a fairly modest dataset size for the scope of this assignment (e.g. 100-1000 sequences if you wish to use alignment-based methods, up to 10,000 for a k-mer based analysis; these are approximate guidelines, not hard boundaries).
- Helpful links:
- https://lgatto.github.io/IntroMachineLearningWithR/unsupervised-learning.html
- https://cran.r-project.org/web/packages/kmer/kmer.pdf
- https://www.bioconductor.org/packages/release/bioc/html/DECIPHER.html
- https://cran.r-project.org/web/packages/cluster/cluster.pdf
- https://cran.r-project.org/web/packages/clValid/vignettes/clValid.pdf

c) Geography and Evolutionary Diversification

- How does geography relate to diversification? Does speciation within a given taxonomic group tend to occur within geographic regions, or does large-scale allopatric divergence contribute to evolutionary diversification? An example specific question under this theme would be: Do sister species for my taxon of choice live in the same geographic region or different geographic regions?
- This project would be a great choice for students interested in biogeography, phylogeny, geographic mapping, and cool visualizations.
- To address this question, you would: obtain DNA sequence data for a chosen gene from NCBI (and could add BOLD) for your taxonomic group of choice, perform sequence alignment and phylogeny generation, obtain distribution information for the species in the genetic dataset (e.g. from GBIF and/or BOLD), match up the sequence and geographic data (e.g. using species name), make a map, and compare the phylogeny and geography for the species in your data set (see links below). The phylogeny portion may be adapted from class

example scripts if you wish, while the geography/mapping portion would be a novel component. There are several helpful tutorials in the links below for the mapping part.

- Tip: use a genus as your taxonomic level of focus; include all available species in that genus.
- Tip: use a small dataset (e.g. 10-100 sequences, after all data filtering steps; that's an approximate guideline). If your taxonomic group has a lot of data for your gene, you can select sequences for inclusion based upon, for example, discarding similar sequences (e.g. those that are <1% different, etc.).
- Tip: use a gene that has support in the literature as having good signal for reconstructing phylogenetic relationships among species within genera (e.g. COI, cytb,12S, 16S, 28S). Considering the intended scope of this assignment, I suggest to choose a single gene, based upon good data availability for your target taxonomic group.
- Example helpful links:
- https://www.molecularecologist.com/2014/11/geophylogeny-plots-in-r-for-dummies/
- https://www.molecularecologist.com/2017/02/phylogenetic-trees-in-r-using-ggtree/
- https://cran.r-project.org/web/packages/phytools/phytools.pdf
- https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0587.2010.06312.x
- http://blog.phytools.org/2019/03/projecting-phylogeny-onto-geographic.html
- http://www.phytools.org/Cordoba2017/ex/15/Plotting-methods.html

d) Trait Evolution

- Is a given trait phylogenetically conserved? (i.e. Do closely related species have similar trait values?) Mapping patterns in trait evolution on a phylogeny helps us to understand functional constraints, cases of rapid evolution, and convergence through evolutionary history.
- This project would be a great choice for students interested in molecular phylogenetics, evolutionary biology, and cool visualizations. This project can also be adapted to have a conservation angle. For example: Do evolutionarily similar species have similar or different conservation status (e.g. IUCN Red List status)? Are we at risk of losing major branches of the tree of life, representing unique evolutionary history?
- This project would involve: choose a specific taxonomic group with trait data available (see pre-recorded lecture on some selected sources of trait data), obtain genetic data from NCBI (and could access BOLD) for this taxon, match up the sequence and trait data (e.g. based upon the species name), and then test for phylogenetic conservatism by estimating the lambda parameter (for a quantitative variable, such as body size). If you wish, you may adapt the phylogeny part from example class scripts, while the trait/sequence matching and lambda calculation would be a novel component. The associated main visualization would be to reconstruct the evolutionary history of the trait onto a phylogeny using tools available in the R packages ape, phangorn, and phytools (see links below).
- An alternative project under this theme is to see whether two traits (or a trait and environmental variable) are correlated with one another among species, taking into account phylogenetic relationships. E.g. Are latitude and body size correlated? Are body size and litter size correlated?

- Tip: use a small-sized dataset for this project (for example 20-100 species; that's an approximate guideline, not a hard rule).
- Tip: use a gene that has support in the literature as having a good signal for reconstructing phylogenetic relationships among species in your taxonomic group. Common marker choices for reconstructing evolutionary relationships among species within genera would include: COI, cytb, 12S, 16S, and 28S (28S also used for deeper phylogenetics). (Note: multi-gene phylogenies would nowadays be more common in the literature. This is something you could mention in your discussion, but is beyond the scope of this assignment. So, I suggest using a single gene here for this theme.)
- Tip: If you are searching for a gene such as 28S through NCBI, also include the long name in your search, e.g. "28S OR large ribosomal subunit". For 18S, which is often used for reconstructing deep evolutionary relationships, this would be: "18S OR small ribosomal subunit".
- Helpful links:
- http://www.phytools.org/static.help/phylosig.html
- http://www.phytools.org/Cordoba2017/ex/2/Intro-to-phylogenies.html
- http://www.phytools.org/Cordoba2017/ex/15/Plotting-methods.html
- https://4va.github.io/biodatasci/r-ggtree.html
- https://guangchuangyu.github.io/ggtree-book/chapter-ggtree.html
- http://www.phytools.org/Cordoba2017/ex/3/PICs.html

e) Molecular Phylogenetics and Congruence

- Do two genes yield the same or different phylogenetic hypotheses? Why or why not? We may get differences in resolution if there are differences in molecular evolutionary rates between genes, for example. Discordance between nuclear gene and mitochondrial gene trees can also be an indicator of a history of reticulate evolution (hybridization and introgression).
- This project would be a great choice for students interested in genetics, molecular phylogenetics, and visualization.
- For this question, you would choose a taxonomic group (of modest size) and obtain genetic data for two different markers. After determining which species have data for both genes, you would perform a sequence alignment for each gene and build a phylogenetic tree for each gene. If you wish, you may adapt class example scripts for the phylogenetics analysis portion. Then, going beyond that, you could examine topological congruence of the two trees through visualization, such as using functions available through the phytools, phangorn, or dendextend package or other resources.
- Tip: Choose a small data set size for this project (e.g. 10-50 species for each of two genes, after data filtering steps; this is an approximate guideline).
- Tips: consider performing this project for a genus or family-level taxonomic group. Choose genes that are supported as being helpful for resolving relationships among species within genera or families. Examples of common genes used for phylogenetics include COI, cytb for vertebrates, 12S, 16S, and 28S. You would choose two genes for this project.
- Helpful links:

- http://www.phytools.org/Cordoba2017/ex/15/Plotting-methods.html
- http://blog.phytools.org/2013/09/plotting-facing-trees-using-phytools.html
- https://cran.r-project.org/web/packages/phangorn/vignettes/IntertwiningTreesAndNetworks.html
- https://cran.r-project.org/web/packages/dendextend/vignettes/dendextend.html
- https://www.r-bloggers.com/dendextend-a-package-for-visualizing-adjusting-and-comparing-dendrograms-based-on-a-paper-from-bioinformatics/

Detailed Project Requirements:

1. Introduction (short written section of 1-2 paragraphs): Create an objective or pose a question that interests you that relates to databases and DNA sequence analysis. Using one of the five themes described above, narrow into a specific taxonomic group and project objective. Your project may involve building a tool (e.g. classifier), exploring an idea using visualizations, or testing a specific hypothesis. Be clear in your introduction: what is the objective of your project? Aim to cite 1-2 suitable references in your introduction. You may use references uploaded to the CourseLink or other literature you find on your topic.

2. Code Section 1 – Data Acquisition, Exploration, Filtering, and Quality Control. You will obtain publicly available data for your project. I suggest keeping within the above guidelines regarding data set size, in order to keep your analysis to a suitable scope intended for this assignment. You may go a bit above the recommendations if you wish to do so, but this should be in service of your project idea. (*Note: there are no extra marks to be earned simply for including a larger data set.*)

After data acquisition, you will next conduct data exploration, filtering, and quality control. What is suitable in this section will depend upon your project. Here are a few examples:

For key quantitative variables (e.g. organism length, mass, etc.), calculate data summaries, such as using the summary() function. Also, prepare a suitable plot, such as a boxplot or histogram, showing the distribution of the data. If suitable for your project, you may use multiple panels if you have multiple variables of interest. Check on outliers. Are these likely real data points, or could they be data entry errors? Note that we are looking for errors, not looking to exclude data that disagree with our hypothesis. You may decide, for example, that even if there is an outlier that it is likely real data and you might perform a data transformation (e.g. log) so that the observation doesn't have extreme influence upon your results. Or, you may perform your analysis with and without extreme data points and make a comparison.

For DNA sequence data, you may consider doing a check for sequence lengths and exclude sequences having very short length compared to the majority of your data, for example. You may also consider excluding sequences with a large proportion of Ns. A suitable visualization for DNA sequence data could be a histogram of sequence lengths. If suitable for your project, you may investigate outliers based upon a pairwise distance matrix (i.e. are there sequences very

different from others in the dataset?). Another helpful quality control step is to build a visualization (such as a dendrogram) to look for sequences very different from others. You could then BLAST unusual sequences (using NCBI tool) and check the top few hits and see if there could be a misidentification in the dataset. (We will see an example of building a dendrogram to screen sequence data in an example script.)

Also, make a deliberate decision about how you will treat missing data for your project. Will you focus on genes or traits with a good sample size, for example? Do you need complete cases (e.g. can you only include species having sequence data for each of two genes)?

Your data exploration and quality control section must include 1-2 figures. One is the minimum. Two is the maximum, but two are NOT required to do well. A single, high-quality figure is also great for this section, if suitable for your project.

3. Code Section 2 – Main Analysis. You would then move on to performing your main analysis. See above example projects. What would go in this section would depend upon your project: writing code to build and test a classifier, build a visualization to explore ideas, or conduct a statistical test of a hypothesis. Your main analysis section must also include 1-2 figures. Please note that you do NOT need to include 2 figures. Rather, 1 is the minimum, and 2 is the upper limit. The grade is NOT based upon the number of figures but based upon quality and what is suitable for your project. A single, high-quality visualization (such as mirror image trees for project *e*, a geophylogeny for project *c*, or trait evolution mapped onto a phylogeny for project *d*) would be excellent for this section.

Tips on code: Throughout both code sections, comment your code well. Make it clear what you are doing and why at each step. Check carefully that your code works as intended, and use consistent formatting.

4. Quality of Visualizations. Throughout, ensure that your figures are clear and well labeled. Even for simple figures such as histograms, ensure that you have accurate, informative axis labels that include units of measurement. Also, consider readability, visual appeal, and accessibility. Use well-differentiated colours, and avoid relying upon the red-green spectrum to convey scientifically important information. Remember, you can consider using a combination of colour and symbol/pattern to convey your meaning. The grade in this section is based upon quality and novelty, not having the maximum permissible number of figures.

5. Results and Discussion (short written section of 1-2 paragraphs). At the end, write a short section describing and interpreting what you discovered through conducting your project. E.g. Did your classifier work as originally intended or not? Or, did you discover what you expected or not? Why might that be? What did you learn? Also, indicate any key caveats about your study. Here, I am looking for a few thoughtful sentences about the main points. (This is a short project; there isn't a need to go on and on about the limitations of your project!) For example, was the available dataset large enough to address your question? Were there biases in data availability

relevant for your project? What would you do next if you were going to develop this work into a larger project?

6. Acknowledgements. If you received project tips from others, include that information 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/paste 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?)

7. References. If you cite any sources from the scientific literature, cite them here. You may wish to consult 1-3 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 URLs here. Also, include the citation for key R packages you used.

8. Length – The total length for your assignment, including all components, is expected to be 5-10 pages, with 15 being the maximum permitted.

9. I hope that you enjoy completing your project and learn a lot along the way. Have fun!

<u>Tips for Success - Checklist</u>

- ✔ Start your project early to avoid stress at the end!
- ✔ Before you start coding, sketch out the steps in your analysis. This helps you to think about the suitable ordering of steps. Also, this creates a checklist for you of coding tasks, each of which is more manageable. “Do Software Tools project” is too large an item to put on your daily task list and makes it hard to get started. “Download cytb sequence data for genus Rana from NCBI and check sample size” is a much more achievable task for a single sitting! Break your project down into small, achievable steps.
- ✔ After each example script from class, run through it again, adapting it for your candidate dataset to solidify your understanding of the various functions and as a stepping stone for your project.
- ✔ 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, 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 peers, please do so in advance. The day before the due date should be for proofreading and polishing your assignment only. I recommend submitting early to avoid technical difficulties at the last minute.

- ✔ Before submission, start a new RStudio session. Then, 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 correctly.
- ✔ Don't ignore the written sections. Even though they are intended to be concise, those are an important component of the assignment. What are you doing and why is it interesting? And, what did you find out, and what are your thoughts on next steps for this line of enquiry? It is important for scripts to have a context.
- ✔ If there is something you were hoping to achieve for Assignment #2 but don't manage within the available time for this assignment, still do take notes about your idea! You and your group members could consider that as an idea for a code/project improvement for Assignment #3. Assignment #3 will involve swapping code from your Assignment #1 or #2 and working together to make improvements to the code. You can bring your ideas to the table regarding potential improvements to your own script.

Submission Instructions:

Submit your assignment in two formats, each to the designated Dropbox folder on CourseLink:

- **PDF.** This should include the full assignment (introduction, two commented code sections, 2-4 figures total, results/discussion, acknowledgements, references).
- **.R or .Rmd file (additional files may also be submitted if suitable, e.g. data file).**

Software Tools (BINF*6210) – Assignment #2 Grading Rubric

Context and Tips for Success:

This grading rubric is intended to help you to be as successful as possible on Assignment #2. 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. 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, and/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%. While 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%)**
Introduction (1-2 paragraphs) (/10)	7/10 *objective of project is clearly stated *a statement is made about why this project is interesting *writing is readily understandable	8/10 *compelling statement is made about why this project is relevant in bioinformatics or biology *reference is made to the literature (e.g. 1-2 relevant references cited) *logical flow to the writing *very good quality of writing, with few grammatical or punctuation errors	9-10/10 *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) *excellent quality of writing, with very few to no grammatical or punctuation errors
Code – Part 1 (Data Exploration and Quality Control) /30	21/30 *code must work as intended *summarizes key variables of interest (e.g. mean, range of numerical variables, sequence lengths) *check for evidence of serious errors in any variables you are analyzing *most or nearly all lines of code may be adapted (correctly) from course example scripts *use commenting to explain what you are doing at each step *code well formatted	24/30 *clear explanations are included about data exploration and reasons for specific data filtering steps *includes thoughtful, deliberate approach as to how you will treat NAs in your variables of interest for analysis *considers class balance (if suitable for project, e.g. machine learning) *thoughtful explanations are included about why you are doing what you are doing; considers settings for arguments used (if you use default, that's fine, but explain why) *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 present in the source data that may influence your project *may include novel types of data summaries or graphics to show the distribution of data and to reveal potential errors or biases *includes some novel code

Code – Part 2 (Analysis to Address Question) /30	21/30 *code must work as intended *includes an analysis to address your main objective (see example projects above) *most or nearly all lines of code may be adapted (correctly) from example scripts *use commenting to explain what you are doing at each step *code well formatted	24/30 *includes an original component beyond example scripts, such as using a new method or a graphical approach (multiple examples are included above) *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
Quality of Figures /20	14/20 *project must include 2-4 total 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 (multiple examples are described above) *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 (1-2 paragraphs) /10	7/10 *clearly state what you found, address your original objective *state whether your results were aligned with your expectations or not, why might that be? *writing is readily understandable	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 *writing flows well, with few grammatical or punctuation errors	9-10/10 *excellent critical interpretation of your findings and caveats *outlines next step 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)

			*excellent quality writing, with very few to no grammatical or punctuation errors
*Acknowledgements section is 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. (Credit included in intro and discussion grades.)			
Total Graded out of 100% and Valued at 15% of Course Grade			