Final Project Guidelines for BST 281

Final project initial proposal due: **4/15/2022**
Final project write-up and data/code packet due: **5/13/2022 (end of semester)**

The BST 281 final project requires working as a group to perform and document a series of genomic data manipulation experiments. Students will work in groups of 3-5 with each group member performing and describing one or more steps in the analysis (including relating their inputs and outputs to those of the other group members). Each group member's analyses must represent novel work and be distinct from the analyses of the other group members.

As an example, a (very common) structure for a 4-member final project group involves student *A* processing raw data to generate a feature × sample table, student *B* performing statistical analyses to identify features that covary with sample metadata, student *C* performing feature-set enrichment analyses on the resulting differentially abundant features, and student *D* analyzing differentially abundant features in a network context. (*This is also a very common structure for genomic data projects in the "real world."*)

That said, many other possible project specifications are possible (and encouraged!). Consider, for example, the menagerie of potential analyses one can perform starting from a single human gene of interest: 1) building a non-redundant catalog of gene homologs; 2) determining if the gene functions as a reliable phylogenetic marker; 3) determining if conserved vs. variable regions correlate with known disease-associated mutations; 4) identifying contexts in which the gene is (differentially) expressed; 5) building a 3D structural model for the gene's protein product; 5) characterizing the gene's neighborhood in one or more molecular networks (physical, genetic, co-expression); and so forth. Think about the many data types we've discussed in the course, the analyses that interrelate them (or otherwise take them as input), and try to come up with something creative!

Project specifications to avoid / requiring special proposal permission:

1. Projects that involve reproducing an existing analysis (e.g. paper) without using substantially different analysis methods (see "novelty" discussion below).

2. Projects that involve each group member performing the same analysis steps on different input data.

3. Projects that involve group members performing individual analyses that are difficult to relate to one another.

4. Projects that involve (re)processing very large amounts of raw data (e.g. 100s of GBs). *Reprocessing a limited amount of raw data (e.g. 5-10 samples) from a paper using new methods (to compare and contrast) would be OK as one step of a project (with other steps then relying on the full intermediate data from the paper itself).*

The group will first submit and iterate on a proposal for the final project with the instructors to ensure its sufficiency and feasibility (details below). At the end of the semester, the group will submit a single project write-up with shared (short) introduction and discussion sections flanking a series of member-specific methods/results subsections describing that person's individual contributions (alongside a data figure; details below). Each member will also submit a data/code packet in support of their individual analyses. Each student will receive an individualized final project grade based on the quality of their individual writing, completeness of their individual data/code packet, and a "project cohesion" score shared among all group members (based on shared written components and the overall flow of the project and write-up).

Submitting your final project initial proposal (DUE by email 4/15/2022)

The initial project proposal describes who is working in the group, the general theme of the project, the input (not necessarily "raw") data that the group will be starting from, and the set of analyses that each group member will perform. The goal of the initial proposal is to make sure that the project involves a sufficient and feasible amount of work for the number of students involved and the amount of time in which to complete it. The project will also be assessed for novelty and to make sure it avoids the problematic designs introduced above. Note that "novelty" here means "we haven't done this before" and (when focusing on input data from a specific paper) "we're not just reproducing a previous paper." It is not critical for the proposed work to be of great societal importance or for it to promise to produce newsworthy results. The work should, however, be scientifically sound and have the *potential* to produce interesting results (e.g. statistically significant trends of at least modest effect size).

Be as specific as possible when outlining the proposed work: citing who is performing which steps, using which data, and which methods. If one analysis step picks up where another leaves off, make sure to indicate that. It is not critical to cite specific analysis parameters at this stage (e.g. a BLAST *E*-value cutoff). Specify the motivation for analysis steps if they are unusual or potentially not clear from context. Here is an example pair of analysis steps:

> "*In part 1, Eric will gather CDK1 homologs by using BLAST with the human CDK1 protein sequence as a query and Swiss-Prot as a reference database. He will cluster the resulting hits to produce a non-redundant sequence catalog using CD-HIT (the goal of this step is to avoid forwarding large numbers of near-identical CDK1 homologs, e.g. human disease variants, to downstream steps).*"
>
> *"In part 2, Kelsey will multiply align Eric's sequence catalog using MUSCLE, QC the MSA to remove unreliable regions using Gblocks, and then build a phylogenetic tree over the sequences using FastTree. This is one of the endpoints for our analysis.*"

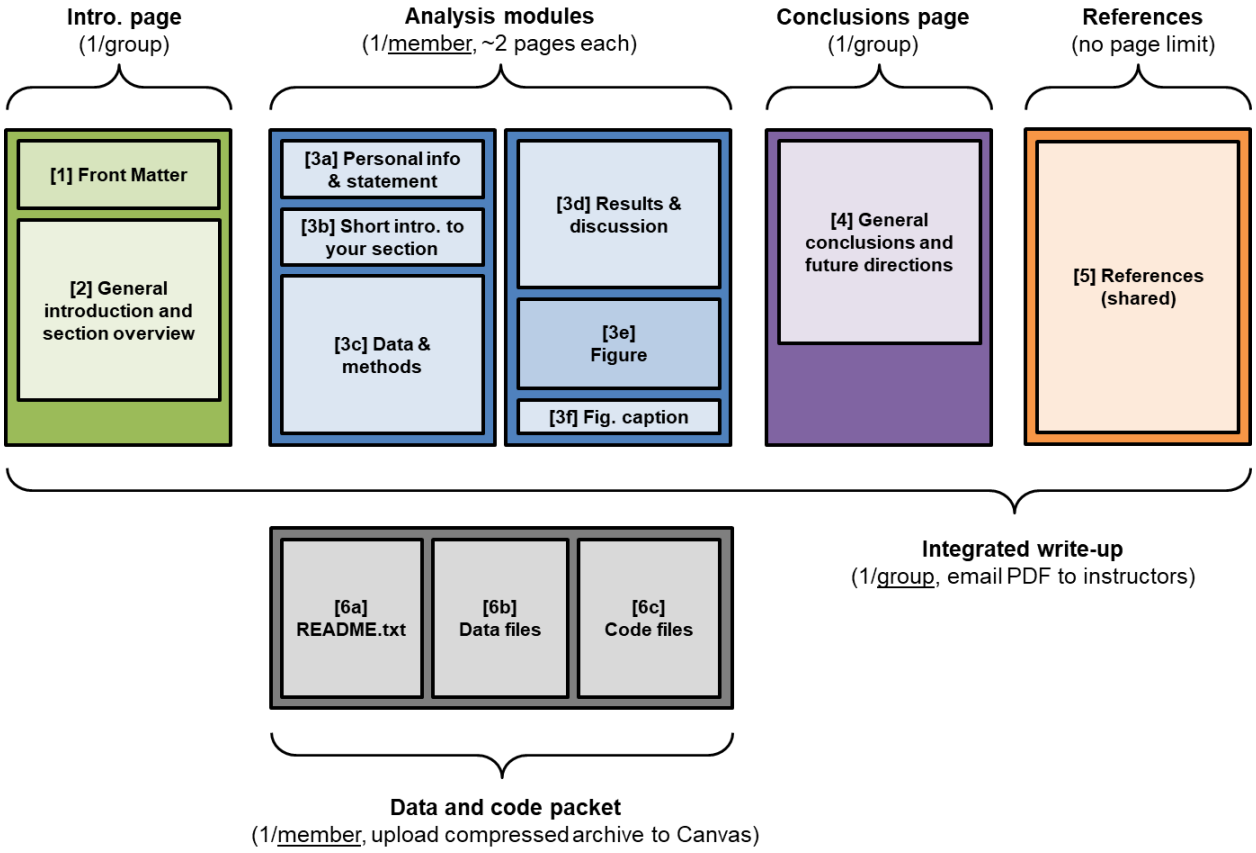
The final project proposal should also contain an attestation from all members indicating that the proposed work is a unique creation for BST 281. If you are using data from one or more group members' individual research, note specifically that the analyses you are proposing have not already been completed. If you are adapting analyses you applied previously (potentially on distinct data in the context of other courses/research) specify that as well so we can confirm that the proposal is sufficiently distinct. We have had issues in the past where the non-uniqueness of final-project work came to light late in the semester, resulting in either 1) a scramble to revise the project or (worse) 2) substantial loss of credit. These outcomes are easy to avoid by being clear and honest during the proposal stage.

While there is no strict length requirement for the final project proposal, we would expect to see at least a few sentences of general introduction to the problem that you want to study followed by a few sentences per group member describing the data and analysis methods they will use (see examples above). The novelty attestation should be as long as it needs to be to address the handful of specific issues noted above (or to raise any specific concerns you have).

Send your proposal by email to Eric and Kelsey with all group members in the CC list. Use a unique and informative subject line such as “BST 281 final project proposal - CDK1 phylogeny” (to match the analysis examples above). Doing so makes it easy for us to find final project proposal threads and use them for communication with individual groups. You can include your proposal in the body of the email or as an attachment. Be prepared to iterate with the instructors 1-2 times to add/remove/clarify analysis components. It is OK to include a small number of open questions in your proposal (e.g. “which data/method would you recommend for step X?”); try to solidify as much other detail as possible before sending, however.

Writing and submitting your Final Project (DUE by email + Canvas 5/13/2022)

Your final project submission will consist of two elements: i) an integrated write-up describing the project that contains group and individualized components (due by email to the instructors in PDF format) and ii) an individual data/code packet (due by Canvas upload in ZIP or TAR.GZ format). These items are due by 11:59pm Friday 5/13/2022 (there are no options for late hand-ins other than taking an INC for the course). These two deliverables are diagrammed graphically below and then discussed in detail in the following sections.



The integrated write-up

We recommend working on your integrated write-up in a collaborative online environment such as Google Docs, Microsoft Office 365, or HackMD. However you assemble the document, the final version must be delivered to both instructors by email as a single PDF containing your group and individual components. Include all group members in the email's CC list and confirm that all members have signed off on the submitted document. The individual sections of the document are detailed below.

Part 1: Front Matter

- Include your names, emails, and Harvard IDs.
- Include a brief title for your project.

Part 2: General introduction and section overview

- Describe the scientific background for the project (why it's interesting/important and why you chose to work on it). This can be somewhat longer than the introduction provided in the final project proposal and is expected to contain citations.
- Make sure to emphasize the *biological* importance of the project: what gap (however small) in current biological understanding is it aiming to fill?
- Introduce any important concepts and vocabulary that will be important for understanding the analysis sections.
- Provide a brief overview of the analysis modules to follow (this can be in LESS detail than the final project proposal - e.g. excluding specific methods).
- This section will probably be in the neighborhood of 300-400 words.
- This section should be developed collaboratively and will count toward the "project cohesion" component of your individual grades.

Note: We'll provide suggested word counts throughout this document. You won't be penalized if you go over this count by a bit, but there will be penalties for writing substantially more than is requested. Conversely, if you fall well below the suggested word count, then you are probably omitting important details.

Part 3: Individual analysis module

Each group member will write up their individual analysis in about two pages (including a figure) as described in the following sections (Parts 3a-f) and outlined graphically above. The total text for an individual analysis section should be in the neighborhood of 750 words. Each analysis module (i.e. group member's section) should start on a new page.

Note: The outline below is a guide and not a strict rubric. For example, if it's easier for you to combine Methods and Results followed by a separate Discussion section, that is OK. Similarly, your individual sections may not hit every bullet point suggested below. However, your final text ought to address most of the ideas described below in some way. We recognize that analysis modules focused on data preparation and QC will lean more heavily on methods/technical detail vs. new biology (and will take this into account when grading). All analysis modules should be written in sentence/paragraph style with smooth transitions (not an outline).

Part 3a: Personal info and statement

- Repeat your name, email, and Harvard ID from the title page.
- If you'd like to add any meta-commentary about your work (e.g. indicating that you tried something that didn't work out and therefore isn't described in the main text), you can do so here. This text is optional and doesn't count toward your word limit.

Part 3b: Short introduction to your section

- Remind the reader where your analysis module fits within the overall project.
- Remind the reader what the goals of your analysis module are. This will be similar to the short overview of your module in your final project proposal.

Part 3c: Data and methods

- Describe the input data you collected in detail. Questions you might answer here:
    - Why was this particular data chosen?
    - Where did you get the data?
    - What format did it arrive in?
    - How many samples/features were present?
    - What percentage of features were annotated?
    - What had been done to the data before you started working on it (e.g. outlier removal or normalization)?
    - Include citations to data from published works or databases
- Describe the transformation you applied to the data in detail. For example:
    - Names and versions of software.
    - Descriptions of custom code/algorithms.
    - Parameter choices (including defense of those choices, e.g. why you didn't stick to a method's default value for a particular analysis).
    - Descriptions and justifications for statistical tests (including parameter choices, e.g. number of permutations or effect-size thresholds).
    - Format of the output data from transformation (see above questions).
- Describe computational controls for each step.
    - What made you trust the data you were starting from?
    - What made you trust that your analysis results were correct?
- This section can and should include references to your data and code packet.
- This section might include references to your figure.
- Feel free to bold/underline important ideas/transitions (in place of subsection headers that one would typically use in a longer manuscript).

Note: The goal of this section should be to walk a colleague through the process of reproducing your work in a linear fashion. It is also a place to show us the technical details of 'omics-scale data analysis you've picked up over the course of the semester.

Part 3d: Results and discussion

- Discuss your global findings. Example directions include:
    - What were the broad 'omics-level trends in your data?
    - What was the magnitude and statistical significance of these trends?
    - How do these trends relate to your initial hypotheses/goals for the project?
    - How do these trends relate to known phenomena from the literature?
        - Include citations where applicable.
    - How do these trends relate to phenomena we discussed in class?
- Discuss specific examples.
    - Include discussion of at least one (and preferably a few) highly specific examples that you selected from your 'omics-scale analyses.
    - Indicate why you selected these examples / why they are biologically interesting.

    - Discuss if your examples match or differ from your global trends or other global trends known from the literature.
- Indicate how your analysis connects to preceding analyses or contributes to analyses that follow (in the linear sequence of the document).
- This section might include references to your data and code packet.
- This section can and should include references to your figure.
- Feel free to bold/underline important ideas/transitions (in place of subsection headers that one would typically use in a longer manuscript).

Note: While the Data and Methods section leans technical, Results and Discussion should lean biological. Use this section to illustrate your understanding of biological ideas in relation to the analyses you performed. "Biological ideas" includes both emergent/systems-level phenomena (e.g. the idea that co-expressed proteins tend to be functionally related) as well as detailed molecular phenomena (e.g. the idea that Cyclin-dependent kinases are protein kinases that require binding of a separate cyclin subunit for full enzymatic activity). Err on the side of conservatism in your biological conclusions: e.g. make sure to consider simple explanations for your observations, including technical explanations, before reaching for wild conclusions.

<u>Part 3e: Figure</u>

- Each analysis module MUST include a figure.
- The figure can (and likely should) have multiple panels.
- You should label your panels by letter and refer to them individually in your text.
    - E.g. "I observed that X was correlated with Y (**Fig. 3b**)."
- Figure panels can draw from any of the following:
    - Schematic/cartoon diagrams of your analysis workflow.
    - Sanity checks of data quality.
    - High-level overviews of 'omics-scale data (e.g. ordinations, heatmaps).
    - Plots of individual trends (e.g. scatterplots, barplots, box plots).
    - Schematic/cartoon diagrams of ideas from your discussion.
- Show raw data where possible, especially when highlighting specific trends (e.g. showing an X vs. Y scatter plot is better than just showing a trendline).
- Follow principles of effective scientific data visualization as discussed in class.
- Try to have the figure "speak for itself" (label panels/axes/complicated mappings).

Note: Figures are a critically important aspect of scientific communication and will be assigned a lot of weight during evaluation. Don't make this part an afterthought!

<u>Part 3f: Figure caption</u>

- Your figure caption should include:
    - The number of your figure within the overall document.
    - A short title for the figure.
    - Brief descriptions of the subpanels.
- Avoid including non-trivial results/methods text in the caption: these details are better embedded in your main text.

- It's OK to include small amounts of detail in the caption that are helpful for understanding the figure but which aren't appropriate for the main text or the figure itself.

Part 4: General conclusions and future directions

- Like the general introduction, this section should be developed collaboratively and should wind up in the neighborhood of 300-400 words.
- Use this section to make connections between the various analysis modules, indicate what you learned as a group completing the project, and speculate about some next steps in the research (if you were to pursue it further).
- This section should start (and be contained) on a new page.

Part 5: References

- Include standard bibliographic information for any works cited across your combined document (this single section should cover both the joint introductions and conclusions and all individual analysis modules).
- You may use any citation and bibliography format you deem convenient.

The individual data/code packet

Each group member will submit an individual compressed archive file (ZIP or TAR.GZ format) containing data files, code files, and a descriptive README.txt file in support of their individual analyses. You can think of this packet like the "Supplementary Information" of a journal article: i.e. your text and figure should make sense without having access to this information, but you can refer to it in your writing where useful, for example "After quality control, my co-expression network contained 1,701 edges (see **packet/edges.tsv**)." Try to keep the compressed packet to <100 MB to avoid upload problems (see the tips under Part 6b below for suggestions).

Part 6a: README.txt

- This is the most important file in your data packet.
- Your README.txt should include a manifest of the included data and code files along with short descriptions of each.
    - E.g. "edges.tsv: This file, which is in edge-list format, includes the final set of co-expression edges I analyzed for guilt-by-association effects."
    - E.g. "gba.py: This Python script takes an edge-list format network and a mapping of nodes to node sets as command-line arguments and then computes guilt-by-association statistics."
- You can use this file to describe large datasets that aren't included in the packet itself.
    - For example, you can provide a URL or database + accession number for raw data that you gathered from the internet.
    - Similarly, you can describe large intermediate or final data you computed that are too cumbersome to include in the packet itself.
- You can use this file to describe data that you aren't able to share (e.g. raw data from your individual research).

Part 6b: Data files

- Raw, intermediate, and final data you used for your analyses.
- See the README.txt description above for data exclusion criteria.

Part 6c: Code files

- Include any scripts, modules, or notebooks you created to help with your analyses.
- We do NOT expect every individual to write substantial new analysis code / algorithms as part of their final project analyses.
- However, we DO expect (with rare exceptions) that all individuals will need to write SOME code to reformat data, generate plots, or perform statistical tests. (This could be as simple as enumerating commands you ran from a terminal as a shell script.)
- You do not need to include publicly available code that you used in the packet (though you might want to describe it by comment in your own code or in your README.txt file).
- Given the short turnaround for final project grading, it is unlikely that we will be running your code. However, the code you provide (coupled with your data and methods descriptions) should be sufficient to reproduce your analysis if we tried.

Assessment

Each group member will be graded individually based on the following criteria (only the first criterion, project cohesion, is a shared score among the group members):

- **Project cohesion (15%)**
    - Covers the clarity and insightfulness of the general introduction and discussion sections (Parts 1 and 4 above).
    - Covers how the distinct analysis modules relate to one another.
    - Covers the shared references section (Part 5).
- **Motivation and descriptiveness of Data and Methods (25%)**
    - Covers the module introduction + data/methods criteria (Parts 3b-c).
- **Clarity and insightfulness of Results and Discussion (25%)**
    - Covers the results/discussion criteria (Parts 3d).
- **Informativeness and quality of the scientific figure (25%)**
    - Covers the figure/caption criteria (Parts 3e-f).
- **Completeness of the data and code packet (10%)**
    - Covers the packet criteria (Parts 6a-c).
    - Strong emphasis on the README.txt file.