THOMPSON
RIVERS
UNIVERSITY

805 TRU Way
Kamloops, BC
V2C 0C8, Canada

DASC 6310-
Winter 2026

DASC 6310: Data Analysis in Biology and Life Science

Course Project Description

This course requires completion of an group project which has several marked deliverables throughout the the term. The goal is to provide students with practical experience in researching, designing, implementing, and evaluating research topics in biological data analysis. It involves research skills, writing skills, and presentation skills. Each group must consist of 2–3 members.

The project topic must be relevant to the general field of biological data analysis, chosen by the student(s) and approved by the instructor.

Grading Scheme and Schedule

There will be several deliverables of this project. Requirements and marking details of each part will be discussed in details in the next subsections. The weights and deadlines of the three parts are as follows:

Deliverable	Due date	Weight
Project Topic with a short description	Jan 23 (or earlier)	
Proposal	Feb 4	5%
Good reference selected	Feb 24 (or earlier)	
Good reference Presentation & Participation on Q&A	March 2	5%
Final presentation & Participation on Q&A	April 8 & 13	5%
Project Paper	April 17	20%
Total of the project		35%

Note: before you start working on the project, please make sure that your choice of project topic is approved by the instructor.

Milestone 1: Project Proposal

Write a document of no more than **three** pages of single spaced **12 point** text, with **1"** margins, including a cover page. Up to 2 additional pages may be added at the end of the text for figures and diagrams.

Your proposal should have the following sections:

1. Cover page: project title, the member's name and student ID.
2. Topic description and motivations (why is this topic important?).
3. Objectives and research methods (research questions to answer; your goals; suggested improvements; how you intend to answer the questions / achieve the goals).
4. Timeline and plan for each objective.
5. References.

Submission

An electronic copy of your proposal must be submitted in PDF format by the due date, as listed on the grading and schedule section. Use moodle for submission. Put your name in the filename (e.g. Your_Name_proposal).

Milestone 1 Marking Criteria

You will be marked on the basis of the following (weights of each item are given in brackets with a total of 100%):

- completion of required items (10%),
- insight into problem definition and statement of goals (20%),
- maturity of the work (i.e., clearly thought through rather than "tossed together" ideas) (30%),
- quality of writing, clarity of exposition, and usefulness of supporting figures (if needed) (30%),
- adhere to the proposal submission criteria (e.g., page length, font size, components, file format, etc.) (10%).
- Late submissions will receive **no** credit.

Milestone 2: Good reference Presentation

Each group will present one or more good reference articles during the scheduled class time. The presentation is 15-20 minutes followed by Q&A. Please practice well before the actual presentations.

Coverage

- Project overview (motivation, problem being addressed and other objectives).
- Introduction to the project topic and the research questions you proposed.
- Demonstrate how the article(s) used data analysis and programming to answer the questions.
- Overview of results.
- What you learn from the article(s). What can be the future work of the project.

Tips: Use the C.R.A.P. principles to help you design an excellent presentation (https://www.instructables.com/id/CRAP-principles/).

Submission

- Presentation slides or notes, if you have one in your showcase video.

Upload your documents on Moodle according to your assigned schedule (will be determined later).

Milestone 2 Marking Criteria

You will be marked both on style and content of the showcase. The marking criteria for the presentation include (each has a weight of 20% to a total of 100%):

- communication skills
- organization and preparation
- presentation layout and visuals
- adherence to the topic in the proposal
- clarity of your work

Milestone 3: Final Project Presentation

Each group will present their completed project during the scheduled class time. The presentation should clearly demonstrate the full research and data analysis process, from problem formulation to final conclusions. The presentation is 20 minutes long, followed by a short Q&A session. Please rehearse carefully to ensure smooth delivery and adherence to the time limit.

Coverage

- Brief recap of the project motivation and research questions.
- Description of the data:
  - data sources,
  - data collection or preprocessing steps,
  - key variables of interest.
- Explanation of the statistical and/or data science methods used, and why they are appropriate for the problem.
- Demonstration of how data analysis and programming were used to address the research questions.
- Presentation of main results, including figures, tables, or visualizations where appropriate.
- Interpretation of results and discussion of limitations.
- Conclusions and potential future directions or improvements.

Tips: Focus on telling a clear and coherent story. Make sure all group members understand the full project and can explain the methodology and results. Use visuals effectively to support your explanations, and avoid overcrowded slides.

Submission

- Final presentation slides (or detailed presentation notes).

Upload your materials to Moodle according to the assigned schedule.

Milestone 3 Marking Criteria

You will be evaluated on both content and presentation quality. The marking criteria for the final project presentation are as follows (each weighted at 20%, for a total of 100%):

- clarity and correctness of the data analysis
- organization and logical flow of the presentation
- presentation layout and effectiveness of visuals
- interpretation of results and critical thinking
- overall communication skills and professionalism

Milestone 4: Five-page Research Paper

Each group will write a five-page research paper (excluding references) as the final report of the project. The paper must follow the . Springer Lecture Notes in Computer Science template . The templates are available in both Word and LaTex. You are encouraged to use LaTex to write the paper. Here is a tutorial to get started with LaTex using free online editor Overleaf: https://www.overleaf.com/learn/latex/Tutorials

Your paper should have the following sections:

1. Abstract.
2. Introduction: the overall purpose and motivation (why is the topic important?)
3. Methodology: research methods (for example: how you conduct the research to answer the questions, implementation of specific bioinformatics tools).
4. Research results: body of the report – this will correspond to your research presentation.
5. Conclusions: comments on your methods, approach, contribution of this work.
6. References: at least 15 sources from books, research articles, and web.
7. Appendix

Submission

An electronic copy of your paper must be submitted in PDF format by the due date, as listed on the grading and schedule section. Put your name in the filename.

Milestone 4 Marking Criteria

You will be marked on the basis of the following (weights of each item are given in brackets with a total of 100%):f:

- Completion of required items (10%).
- Insight into problem definition and statement of goals (20%).
- Maturity, amount and quality of the work (30%).
- Quality of writing, clarity of exposition, and usefulness of supporting figures (30%).
- Adhere to the ACM SIG Proceedings format (10%). template
- Late submissions will receive no credit.

Options of project topics

Below are some options of topic choice [1]. The purpose of this list is to give you some detailed idea and guidance to choose your topic of interest. You could pick one from this list for your project, but please don't be limited by the list. Feel free to come up with other topics of your choice. Make sure to send your topic to the instructor for approval before Jan 23. Topics that are not approved may result in redoing the project on other topics.

Topic 1: **Bioinformatics history and scope:** By focusing on the generation of sequencing Applied genomics and next-generation sequencing, you will understand how to apply appropriate data types and analyse high-throughput data sets using the latest state-of-the-art approaches. For Example: Summarize

- Human Genome Project development
- Human Disease Genetic Basis Identification
- Health data analytics and Medical Informatics.: Learn how genomics is changing the future of healthcare
- Plant and animal use is
- Learn how genomics is changing the future of healthcare,. . . .
- Current Topics in Human Genetics

template

Topic 2: **Comparative Genomics** Comparative genomics is a field of biological research in which researchers use various tools to compare the complete genome sequences of different species.

For example:

- A review of bioinformatics platforms for comparative genomics
- A review of bioinformatics tools for comparative genomics
- A review of bioinformatics steps for comparative genomics

template

Topic 3: **Substitution models** A number of different Markov models of DNA sequence evolution have been proposed. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. These models are frequently used in molecular phylogenetic analyses. In particular, they are used during the calculation of likelihood of a tree (in Bayesian and maximum likelihood approaches to tree estimation) and they are used to estimate the evolutionary distance between sequences from the observed differences between the sequences. (wiki).

Other good references for starting projects:

- https://evomics.org/resources/substitution-models/nucleotide-substitution-models/
- https://beast.community/custom_substitution_models
- http://www.iqtree.org/doc/Substitution-Models

[1] Some ideas from the textbook: Concepts in Bioinformatics and Genomics, by Jamil Momand and Alison McCurdy, published by Oxford University Press.

Topic 4: Develop a bioinformatics tool. There could be so many project ideas under this topic. For example, you can develop a sequence alignment program to generate the scoring matrix (M), the traceback matrix (TB), the alignment score, and the pairwise alignment of a given alignment problem, such as

- a global sequence alignment using the PAM150 substitution matrix and a linear gap penalty with $w = -4$,
- a local alignment using the BLOSUM62 substitution matrix and a linear gap penalty with $w = -4$.

Note that the first step in program development is to understand the problem, its inputs, and its outputs. A good way to do that is to generate a solution by hand. In addition to helping you understand how to solve the problem, the solution can also be used to test your program.

Compare to other publicly available tools or algorithms already in existence. Determine the advantages and disadvantages of various tools or algorithms. Which is computationally efficient, fast, and accurate?

Topic 5: Phylogeny is the study about the history of the evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms. Study about the various approaches to constructing Phylogenetic trees and when (and with what kinds of datasets) each one is most effective.

For example: One can construct a phylogenetic tree using the mitochondrial gene that codes for 16S ribosomal RNA from different organisms, such as modern *H. sapiens*, chimpanzee, pygmy chimpanzee, Neanderthal, Denisovan, gorilla, orangutan, mitred leaf monkey, and hanuman langur.

Topic 6: Check whether the alignment results (more than two) match your expectations? If not, investigate the reason(s).

For example: Peyton Rous (1879–1970) was a relatively young man when, in 1911, he discovered a virus that causes sarcomas in chickens. The virus was named *Rous sarcoma virus* (RSV). Later, it was found that RSV is a retrovirus that contains an oncogene, $v - src$, in its RNA genome.

One can perform pairwise sequence alignments between $v - src$ protein and chicken $c - src$, human $c - src$, and mouse $c - src$ proteins. Note that the “c-” prefix is short for cellular. Sometimes protooncogenes are distinguished from viral oncogenes with the prefix “c-” and “v-” respectively. By reporting the identities for each pairwise alignment, you can discuss about the origin of $v - src$. Does the alignment result match your expectations?

Chicken: 94Human: 88Mouse: 87

The high identity shared between v-src and chicken c-src (94The resulting virus, called RSV, was able to out-compete the parental retrovirus for survival. RSV confers a growth advantage onto cells it infects because the c-src promotes cell growth. As the virus continues to regenerate itself, it creates a few mutations that make v-src a more powerful cell growth inducer than its c-src counterpart.

Topic 7: One important exercise a bioinformatician performs is to compare amino acid sequences. One reason to make comparisons is to determine the parts of the proteins that are critical for function. These regions are generally conserved within proteins that perform the same duties. Conserved regions are those that have nearly the same amino acid sequences. Proteins that perform the same duties are called homologs and can be found in different species. For example, $p53$ from humans and $p53$ from frogs perform the same functions. There are some regions within these proteins that will be similar in both humans and frogs. We call these regions conserved sequences. A **multiple sequence alignment** allows the bioinformatician to readily line up amino acid sequences of related proteins. The conserved regions are identified in the alignment.

For example, you can perform a multiple sequence alignment of three homologs of cytochrome C from human, yeast, and dog (or other organisms). Based on your multiple sequence alignment result, you could explain how the alignment gives you a clue as to which parts of the cytochrome C protein you would hypothesize are most important to its function. (The function is the same in all three organisms.) Then, find some evidences to support your results.

Topic 8: Genes in eukaryotes are often organized into **exons** and **introns**, which require splicing to produce an mRNA that can be translated. The gene organization is the order of the DNA segments that comprise the gene starting with the promoter, the first exon, the first intron, the second exon, and so on. The interspersed introns can make gene identification difficult in eukaryotes — particularly in higher eukaryotes with many introns and **alternative spliced mRNAs**. Prediction of many genes and their organization has been based on similarity searches between genomic sequence and known protein amino acid sequences and genomic sequence and the corresponding full-length cDNAs. cDNAs are reverse-transcribed mRNAs and therefore generally do not contain intron sequences. cDNAs (i.e., copied DNA) can be considered mRNAs. A comparison of a genomic sequence (with introns) to its corresponding cDNAs will reveal where introns begin and end. GenBank will contain the genomic sequence and the cDNA sequence. To find out the structure of the gene (i.e., the arrangement of the exons and introns) we simply need to perform a sequence comparison between the genomic sequence and the cDNA sequence.

The Basic Local Alignment Sequence Tool (BLAST) can be used to elucidate part of the **gene organization** (arrangement of exons and introns) of a genomic sequence. BLAST can be used to compare genomic DNA sequence with all RNA sequences (i.e., cDNA sequences) in GenBank. The top hit of the output will be a sequence comparison between your sequence (the query sequence) and the most similar sequence in the database (subject sequence). Subsequent hits will display sequence comparisons between the query sequence and subject sequences that are increasingly less similar. If all hits have 100% identity, then use the hit with the most extensive percentage coverage.

You can use the nucleotide BLAST tool and appropriate databases to construct a schematic diagram that shows the arrangement of introns and exons in the genomic sequence. For example, here is an example genomic sequence from the species *Caenorhabditis elegans* that you can start looking at.