Bioinformatics Project

Introduction

This document provides the information you need to get started on your group project and is formatted in this order:

1. Learning Objectives.
2. Overview of the project.
3. Problem Statements.
4. Problem Statements Requirements.
5. Presentation Requirements.
6. Written portion.
7. How do I get a high grade?
8. Grade Scheme.
9. Teams.

This project is designed to develop your interpersonal skills (working with team members), test your coding skills, and test your knowledge on technologies you have been exposed to in this course. You will be expected to research and learn new technologies in this project as well. If you have any technical problems, need guidance, or want any feedback on your team's approach to a problem, feel free to contact the TA.

Learning Objectives:

- Learn to collaborate as a team to solve a complex problem and build software for bioinformatics.
- Learn to break down a complex problem into smaller more manageable problems.
- Learn to design and implement a solution to the problem given.
- Learn to research and use existing tools to produce a more robust solution.

Overview of the project

For this project you will have two over aching tasks:

1. Build a software solution to a given problem.
2. Assemble and annotate a genome. Then analyze a subset of proteins from the annotations.

In the next section *Problem Statements* your team will be assigned a genome and a given programming task. The programming task will require a write up about how your team solved the problem and other details which will be discussed in more detail later in this document. For the genome assembly, annotation, and analysis of a subset of proteins your team will need to prepare a powerpoint presentation summarizing your analysis.

Problem Statements

Build a software solution to a given problem.

Your team will be responsible for creating a command line program that will take a fasta file with either 1 or many sequences and process that fasta file based on the problem your team is given below.

Problem	Description	Expected Input	Expected output	Teams
Create a command line program that will find all ORFs in a fasta file with 1 or many sequences.	Create a command line program that will take a fasta file and process the sequences to find all ORFs. You can read a multi-record fasta file with the Biopython SeqIO module.	A fast file with DNA sequence(s).	A fasta file with predicted ORFs.	A
Create a command line program that will align two DNA sequences using a **local** alignment algorithm.	Create a command line program that will take a single fasta file with two DNA sequences. The two DNA sequences will be aligned using a **local** alignment algorithm. Your team can implement the algorithm or use Biopython.	A fasta file with two DNA sequences.	The alignment result printed out to stdout.	T
Create a command line program that will align two DNA sequences using a **global** alignment algorithm.	Create a command line program that will take a single fasta file with two DNA sequences. The two DNA sequences will be aligned using a **global** alignment algorithm. Your team can implement the algorithm or use Biopython.	A fasta file with two DNA sequences.	The alignment result printed out to stdout.	C

Create a command line program that will take a fasta file with a DNA or protein MSA and return the consensus sequence for the given MSA.	You and your team will need to create a command line program that will take a fast file with a protein or DNA MSA and return its consensus sequence. A useful tool for handling MSA formatted fasta files is the Biopython AlignIO module.	A MSA in fast format.	The consensus sequence for the given MSA printed to stdout.	G

Genome assembly, annotation, and analysis of a subset of proteins

Below is a table which assigns a team to a particular organism. Your team's task is to first assemble and annotate the genome. After your team annotates the genome your team will need to focus on a subset of proteins which is indicated in the *Proteins of interest* column. Then after filtering your annotations to a subset of proteins you are tasked with narrowing your search for a specific kind of protein within the subset of proteins, the specific protein is indicated in the *Specific protein to focus on the column*.

SRA Accession	Proteins of interest	Oransim	Specific protein to focus on	TEAM
ERR4319784	Carbohydrate activate enzymes	Bacillus Licheniformis	Xylanase	A
SRR057768	Virulence proteins	mycobacterium tuberculosis	MmaA4	T
SRR1588419	DNA repair proteins	Deinococcus radiodurans	RecA	C
ERR4368600	Ribonucleases	Thermus thermophilus	rnpA	G

Problem Statements Requirements

Below are the requirements for the two tasks discussed above in *Problem Statements*. The requirements given are just the base requirements needed to get a passing grade, but to get a high grade your team will need to be creative and put enough effort that is clearly visible in your team's work. The best presentation will receive extra credit points and to be the best your team will need to go beyond just the requirements given below.

Requirements for *Build a software solution to a given problem*:

1. A name for your team's command line program.
2. Fully working command line program that accomplishes the task given in your team's problem statement.
3. You and your team members must prepare a one page write up of your team's command line program.
4. The write up must discuss the problem that your team was assigned, the solution that your team came up with to solve the problem, the technologies that were used (such as Biopython), any algorithms that the program uses, and how to run your program.
5. Additionally, the write up must have a link to your Teams GitHub repository.

Requirements for *Genome assembly, annotation, and analysis of a subset of proteins*:

As your team is preparing the presentation try to relate the requirements given below to the organism you are given, the proteins of interest, and the specific proteins. For example, if your team is focusing on virulence proteins then try to research about such proteins and figure out what may be important to know about them:

6. Your team's powerpoint should give a brief introduction to the assigned

organism, the proteins of interest, and the specific proteins the analysis will consist of.

7. Include a summary of how your team assembled and annotated the genome. In other words, what steps did your team perform to assemble and annotate the genome?
8. Ensure that your team's presentation includes a summary of genome assembly and annotation results; size of the genome, number of CDS, GC content, number of RNAs, number of hypothetical proteins, a summary of CDS pathway analysis (i.e what percent of CDS were predicted to be virulence proteins, carbohydrate degrading, DNA repair, etc.), etc.
9. Discuss how your team filtered for the proteins of interest and how your team narrowed down to the specific proteins.
10. With the specific proteins create a phylogenetic tree to observe the phylogenetic relationship with other proteins.
11. Use a multiple sequence alignment to supplement your phylogenetic tree.
12. What is the 3D structure of the specific proteins? How did your team obtain the protein structure? Was the structure already in a database? Did your team have to predict the structure with bioinformatic software? If so, what technique did your team employ?
13. Are there any notable domains, motifs, or physicochemical characteristics of the specific protein?
14. There is a lot you can do with a protein structure. You can superimpose the structure with other similar protein structures, highlight the domains, motifs, and physicochemical characteristics (Such as coloring all hydrophobic amino acids).

Presentation Requirements

- Your team will be given 20 minutes to present which is about 5 minutes per person. Dr.Balan will time each person and expect that your team's allotted time be used fully. Your team should not finish before the 20 minutes are up or go over the 20 minutes.
- Practice going over your slides and what you will discuss as a team.
- Each member in your team must present a few slides.
- All slides must be numbered and have the name of the person who will present them.
- Do not put only text on a slide. Be more creative than that. Use pictures and some accompanying bullet points.
- Do not read off your slides.
- White background.
- Uniform format.
- The slides should guide you on what you will talk about.

How do I get a high grade?

1. “Be able to explain everything about your slides and how your team was able to solve a particular problem. Be prepared to discuss the problems you come across and the bioinformatic tools you used?” - WORDS OF BALAN
2. Go beyond what is expected of you. Look back at the previous tutorials and lectures to utilize all the skills and technologies you have covered throughout this course.

Grade Scheme

Presentation

Category	**2**	**1**	**0**
Preparedness	Student is completely prepared and has obvious rehearsed	Student seems pretty prepared but might have needed a couple more rehearsals	The student is somewhat prepared, but it is clear that rehearsal was lacking.
Content	Shows a full understanding of the topic.	Shows a good understanding of the topic.	Shows a good understanding of parts of the topic.

Total Score = 15 points
Preparedness = 7.5
Content = 7.5

Software solution write up

Category (3.3pts each)	2	1	0
Discusses the given problem assigned to the team.	Understands the problem given and has a clear solution to the problem.	Does not fully understand the problem and the given solution is not well thought out and ambiguous.	Does not understand the problem given and unable to come up with a solution.
Gives proper credit to utilized software and algorithms.	Explains the different technologies used and the algorithms implemented to solve the problem with clear explanation about the algorithms used.	Explains only some of the technologies and algorithms used with unclear explanation about the algorithms used.	Evident that the code written was copy and pasted without understanding the software and algorithms being used. No explanation about the software used and no explanations about the algorithms used.
Link to Teams GitHub repository.	Provides a link to the teams command line program on GitHub. Team's GitHub repository contains a README and explains how to install and use the software for anyone to use.	Provides a link to the teams command line program on GitHub.	No link to GitHub repository.

Teams

Group Name	Teams (Team leaders are highlighted)
A	• Chantera Lazard • Hasfa Amir • Mohammed Ahmed • Anahi Arellano Esparaza
T	• Dhayni Rana • Jesse Hardin • Duy Nguyen • Mahat Shah
C	• Lauren Harris • Gabriela Villegas • Raymond Sandoval • Tuyet Nguyen
G	• Kenny Dao • Chinaza Nwosu • Atish Sanghavi • Claryssa Casarez