Term Project

General Rules for the Project

A) Teams:

You are to work in teams of two students. Both members of the team are expected to participate equally in the project and will receive the same grade on the project.

B) Project Requirements:

Each group is required to choose a chapter from the textbook [ZB08] from Chapters 13, 14, 15, 16, and 17, read the chapter, and:

a) Perform some of the bioinformatics analyses described in the chapter you choose.
b) Prepare a hands-on exercise on the material you covered in part a).
c) Prepare a PowerPoint presentation (of at least 20 slides) on part a).

Your final project grade will be computed in the following way:

a) Team Formation: 10% b) Progress Report: 20% c) Project report: 70%

C) Important Dates:

a) Team Formation

Tuesday, March 5: Each team submits a paper containing the names of the team members and the chosen chapter. If the topic is not the one described under "Project Topics" below, then the team includes a detailed description of their proposed project, even if it is chosen from "Alternative Projects" on pages 4 to 6.

b) Progress Report

Tuesday, April 9: Two-page progress report by each team is due at the beginning of the lecture. Include the title of the project, description of the problem in your own words, tools that you have used so far in the analysis, the programming language used, the description of the data sets, and references. Also include the questions you have answered so far.

c) Project Report

Thursday, May 2: Printed copies of the final project are due at the beginning of the lecture. Do not forget to include all pertinent documents, such as screen dumps, references, and copy of articles you reference. Also, submit a CD or USB key containing all the files described in Section D below.

I) Programming Project (for Computer Science Majors)

Submission Requirements:

Please include references for any material you have used, including source code.

A) Submit a hard copy of the project containing all of the following:

1) Title page.
2) Table of contents (with page numbers).
3) An essay of not more than 6, 1 and 1/2 spaced pages (font size: 11 or 12 points) describing the problem, the overall organization, design of your program. The essay should include a detailed analysis of your results and comparison with the findings presented in the chapter (or original article). Use MS Excel or a similar package to generate tables and graphs. The essay should give the user an overall roadmap of your code, and would be read by a maintenance programmer before he/she began reading your code. Please do not include definitions, explanations of topics we have covered in the course and do not simply copy the original article. Include a flowchart, or UML or a structure chart to show the design of your program. Do not forget to number pages!
4) A one-page description of the test data including accession numbers. You must test your program with the data sets described in the articles or downloaded from databanks such as GenBank at NCBI. You may use additional data sets if you wish.
5) Include one sample output of your program. Make sure that your output is readable and well formatted.
6) Instructions for running your program, in other words, explain how to compile, and execute your program.
7) A list of references.

B) Submit a CD or USB key, labeled with your names, course number and semester, and containing:

a) The source code (fully documented)
b) The input files
c) The document specified in part A

Please make sure that all the files on the CD or USB key are readable.

Project Topics:

Each group is required to choose a chapter from the textbook [ZB08] from Chapters 13, 14, 15, 16, and 17, read the chapter, and:

d) Perform some of the bioinformatics analyses described in the chapter you choose.
e) Choose an algorithm, implement, run, and test it. If possible, compare its performance to a publically available package.

[ZB08] "Understanding Bioinformatics", by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008.

II) Non-Programming Project (for non CS Majors)

Submission Requirements:

Please include references for any material you have used.

A) Submit a hard copy of the project containing all of the following:

1) The cover page with the appropriate fields filled in by you.
2) Table of contents (with page numbers).
3) A 10-page summary of the topics you chose from the chapter, including the results of the bioinformatics analyses. Use single space and 12pt font. Please do not include definitions, explanations of topics we have covered in the course and do not simply copy the original articles. Use MS Excel or a similar package to generate tables and graphs. Include screen dumps were appropriate. Do not forget to number pages!
4) A copy of your PowerPoint presentation, 6 slides per page. The PowerPoint presentation should be on the topics you have decided to concentrate on.
5) A copy of the Hands-On exercise you have prepared. The Hands-On exercise should be on the topics you have decided to concentrate on.
6) A list of references.

B) Submit a CD or USB key, labeled with your names, course number and semester, and containing:

a) An MS Word document containing the summary mentioned under 3) above.
b) The PowerPoint presentation you have prepared.
c) The Hands-On exercise you have prepared.
d) Additional articles (in pdf) used in your project.

Please make sure that all the files on the CD or USB key are readable.

Project Topics:

Each group is required to choose a chapter from the textbook [ZB08] from Chapters 13, 14, 15, 16, and 17, read the chapter, and:

a) Perform some of the bioinformatics analyses described in the chapter you choose.
b) Prepare a hands-on exercise on the material you covered in part a).
c) Prepare a PowerPoint presentation (of at least 20 slides) on part a).

[ZB08] "Understanding Bioinformatics", by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008.

III) Alternative Topics (for all majors)

A) Genome-Wide Association Studies

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses. [http://www.genome.gov/20019523]

1) Please go to http://www.genome.gov/20019523 and read and understand the following short sections that were last updated on August 17, 2010:

- What is a genome-wide association study?
- Why are such studies possible now?
- How will genome-wide association studies benefit human health?
- What have genome-wide association studies found?
- How are genome-wide association studies conducted?
- How can researchers access data from genome-wide association studies?
- What is NIH doing to support genome-wide association studies?

2) Please go to http://www.genome.gov/26525384 and read and understand the first few paragraphs which describe how the articles that are part of "A Catalog of Published Genome-Wide Association Studies" were chosen. You can also view a figure of the genes and traits (and their locations on the chromosomes) of "Published Genome-Wide Association" through September 2009 at http://www.genome.gov/images/illustrations/GWAS2009-9.pdf.

The project consists in picking one (or two) of the genes or traits mentioned at http://www.genome.gov/26525384 and writing an essay on it (them).

- Prepare a hands-on exercise on the gene/trait you chose.
- Prepare a PowerPoint presentation (of at least 20 slides).
- Optional: Perform some of the bioinformatics analyses described in the articles.

Programming Component: Choose a GWAS algorithm to implement, run, and test it with a suite of input data. If possible, compare your program's performance with a publically available package. Or alternatively, use the same data set mentioned in the article and compare your results to the ones reported in the article.

B) Next-Generation Sequencing

"Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventional methods." [MM2010]
The article presents "a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments." It also outlines "the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest."

The project consists in writing an essay on the NGS. Describe the methods and applications explained in the article, but also go beyond the article. Use the reference section to further explore these methods and their applications.

- Prepare a tutorial (hands-on exercise) that explains how some of these methods work.
- Prepare a PowerPoint presentation (of at least 20 slides).
- Optional: Perform some of the bioinformatics analyses described in the articles.

[MM2010] "Sequencing technologies - the next generation" by Michael Metzker, Nature Reviews Genetics, January 2010.

Programming Component: Choose a NGS algorithm to implement, run, and test it with a suite of input data. If possible, compare your program's performance with a publically available package. Or alternatively, use the same data set mentioned in the article and compare your results to the ones reported in the article.

C) SNPs, Haplotypes, and Copy Number Variations (CNVs)

D) Microarray Analysis

E) Personalized Medicine

F) Machine Learning Techniques

Choose a supervised or unsupervised method and a bioinformatics problem, implement the method and test it on different data sets. Compare your results by running the data sets with publically available software packages.

G) Hidden Markov Models