Introduction to Bioinformatics – AS 250.265
Final Project

One of the requirements for completing the course is the completion of a final project. This project will constitute twenty percent of your final course grade, and it replaces a cumulative final exam for the course. You will have three choices for your final project, and these choices are outlined with their requirements below. Whatever option you choose, the project is due Monday, May 15, 2006 at 12:01 am (in the morning). The instructor will begin grading projects on that day, and, because of the strict requirements by the registrar, it is important that you submit your project on time so your grade can be submitted for your transcript or for graduation.

The three projects are described below. It is suggested that you complete a project that both will challenge you and closely match the skills you already have.

Project 1: Identify and Characterize a Novel Gene

To complete this project, you will be expected to use the skills you have learned to identify an as yet novel gene. You will be responsible for ensuring that no proteins have been submitted to GenBank that match your novel gene. Additionally, you are expected to use the tools mentioned in class or described in your book to characterize your novel gene. Some (but not all) of the questions you will want to consider are: What does it do? Can you determine anything about its structure? Where is it localized in the cell? Is it expressed constitutively or only under certain conditions or at certain times? How does it relate to other proteins in other organisms or within the organism itself? Your ability to address these questions will constitute a large percentage of your grade.

You should submit your findings in the style of a research paper. Examples of research papers abound on PubMed, but an example of an acceptable manuscript (stylistically) is provided for you at the course website at the following URL:

`http://roselab.jhu.edu/bioinfo/files/manuscript.pdf`

This is a manuscript your instructor submitted to a journal, and while the content of your work will differ, you are invited format your submission similarly. A part of the project grade will be stylistic, and you should make sure your submitted project does not have spelling or grammatical errors. Because of the problems with printing digital images and incompatible file formats between Macintosh and PC computers, it is recommended that you submit a hard copy of your final paper. Digital submissions are acceptable, by email, but in that case you should make sure that your paper is readable both on Mactintosh and PC computer, particularly any Greek characters you may use. These are real issues that researchers must deal with when submitting papers, and if you submit your document digitally this will prepare you for problems you may face later on.

When attempting to find a new gene, it is recommended that you start simply. For example, rather than searching human chromosomal DNA, which contains a large number of introns, it is advisable to search through expressed sequence tags as your starting point. You may also want to consider other organisms, realizing that humans and mice have been extensively studied. The list of other BLAST servers in your textbook may also bee a good starting point. Your final gene sequence should be within a complete open reading frame (i.e.

between a start codon and a stop codon) on the DNA. In your paper, be sure to submit the final DNA and protein sequence you recommend, as well as the position of any suspected introns. Introns can be identified in the final steps of your search by examining how your novel gene aligns to complete chromosomal DNA stored on the NCBI website.

Project 2: Identify and Characterize an Unknown Gene

This project is very similar to option number one, except that, rather than finding a novel gene, you will be expected to characterize a sequence that the instructor gives you. In this case, the sequence will correspond to a gene or genes of known function, and you will be expected to perform the same analysis as you would if you were characterizing it as a novel gene. You will have to identify the likely class of proteins that are encoded by the DNA sequence, and you will have to identify what it does, what it looks like, etc. Similarly, you will have to write a research paper reporting what you've discovered.

While at first glance this may seem easier than the first project option, keep in mind that your sequence will be derived from some known protein. Therefore, the analysis you perform will have a correct answer, whereas you will be allowed to speculate to a certain degree for the first project. If you make an incorrect statement about the gene you are given, it will count against you for this project.

Project 3: Implement BLAST or Smith-Waterman

The third option for your final project gives you the chance to study the details of a bioinformatics algorithm in depth as opposed to the broad application you will perform in the first two project options. This is also the only project where you will be allowed to work with up to one partner to complete the project, and both of you will receive the same final grade. Your project will basically be to implement one of two algorithms we discussed in class, either the basic local alignment search tool (BLAST) or the Smith-Waterman local alignment program.

For both choices, your final project must be a functional, well-designed web page that takes input sequences from users and displays the resulting alignment(s) upon submission of sequence data. Your implementation will not be graded for efficiency, but because clearly written programs are much easier to maintain, a stylistic element will be part of your final grade. Whether your solutions handle bad input gracefully will also contribute to the grade. It is suggested that you use PHP/HTML for your front end web pages and Python for your actual implementation of the algorithm, but these details can be decided upon as you work with the instructor. Note that you are not limited to the Python programming language, and if you are more comfortable with other languages, you are invited to use them. A (UNIX based) web space will be provided for you if you choose this project.

In addition to writing the web implementation, you and your partner will be expected to write a three to four page (double spaced) document describing how you implemented the algorithm. In this write-up, you should address issues such as the following questions: Did you make any assumptions in your implementation? What options are you not including that may be available at other web sites? What are the speed bottlenecks of your implementation?

If you choose to implement BLAST, you will implement a BLASTP search against a database of human RefSeq sequences. The sequences can be made available to you through a SQL database or as a large text file. You will have to implement at least two scoring matrices, and you will have to evaluate the statistical significance of your results (K and λ will be provided to you based on the matrix you choose). You may limit yourself to a word size of three, and you need not consider gaps. In addition, you may stop your extension during phase three of the algorithm simply at the first point where a residue pairing has a negative score according to the scoring matrix being used (e.g., if F-T alignment has a negative scoring matrix value, you may stop extension in that direction).

If you choose to implement the Smith-Waterman algorithm, you will perform the local alignment between two user-uploaded sequences. Your web page should output the optimal local alignment between two sequences and, using the statistics discussed for BLAST, you should estimate the statistical significance of your result. Your program should also implement at least two scoring matrices, and here you will have to deal with gaps intelligently, providing at least two schemas for scoring gaps (one of which may be no gap penalty).

It is understood that some of you may not have the experience to complete this project, and thus it may be unwise to attempt it. In addition, you may have to do additional reading beyond just your book to understand how the programs are working in detail. However, understanding how the algorithms are implemented on a full scale is just as useful as understanding how to use them to learn what you want to know. Therefore, some guidance from the instructor will be available if you run in to problems.

Making Your Decision

You must let the instructor know what your final project choice will be by March 20, the first day of Spring Break. At this time you should also let him know who your partner will be if you are choosing option three. If you choose option two, you will receive your DNA sequence when you return from break (Monday, March 27). If you choose option three, it is your responsibility to make a meeting with the instructor with your partner to discuss initial strategies and the resources you will need to implement your website. If you choose option one, no further input is needed. Good luck!