

CSC / BIO 310
Bioinformatics
Instructor: Dr. Laurie J. Heyer
Project #1
Due Thursday, Feb 28

This bioinformatics research project will require interdisciplinary teams to pool your knowledge, skills and creativity to create a tool for researchers in synthetic biology. The tool you will build is needed and will be used by the Davidson College / Missouri Western State University research team, so it is very cutting edge and real, not just a made-up exercise.

The project will consist of a set of web pages, at least one of which must call a Perl program to analyze an input set of sequences and return the results on a new web page. I will show you how to set up this type of system in class, and link to the Perl script that runs behind the scenes of the GC-skew program you looked at earlier in the course. This version is on the gcat server, where you will be working:

http://gcat.davidson.edu/DGPB/gc_skew/gc_skew.html

The Davidson / Western team competed in the 2007 international Genetically Engineered Machines competition last November, with a project that is described here:

http://parts.mit.edu/igem07/index.php/Davidson_Missouri_W

Your team should get to work right away on understanding their project. The key aspects to your work will be discussed briefly in class. I will help teams with little previous exposure to synthetic biology to understand the project, but **only up until class time on Tuesday, Feb. 19.**

Now we have sequencing data for a large number of potential solutions to the Hamiltonian Path Problem. They are contained in two archived sets posted online. The sequences come from three different primers. Your task is to analyze the sequences and determine their composition relative to the expected solution. For example, if the solution can be AB, ABC or ABC', your results web page should clearly show which of the three possibilities is supported by the sequence data.

You may want to do some sequence alignments with the Smith Waterman algorithm, which will be provided to you. Every project must also include appropriate dot plot visualizations of the alignments. Basic code for dot plots is available at <http://www.oup.com/uk/orc/bin/9780199277872/01student/programs/ch04/>, for example, but you will likely want to improve this display by using color, fonts, or different printed characters to indicate the locations of different sequence elements. You might also want to modify the algorithm for the dot plot to allow you to show similarity between one sequence and the reverse complement of the other. Keep in mind that we only have one strand of sequence data for each primer, and we know the orientation for two of the

primers, but we do not know in advance the orientation of the sequence amplified by the third primer.

You may use any inanimate resources to complete this project, provided that they are properly cited. This includes reading and using openly available code online. You must cite the source of any code you use, both inside the code and on your project web page. It is **plagiarism** to use even a few lines of code from someone else without citing the source. It is **stealing** to take code from someone else who doesn't make it explicit that you may do so.

Your web pages must be self-explanatory and provide links and references that would help the user learn more or dig deeper. The user should not have to be familiar with the Hin-Hix system, so you need to provide background on this system and its application in this project. You should allow the user to make choices and provide inputs to the system that will make it extensible to new situations. This is a tool we hope to be able to share with the international synthetic biology community, so style, clarity, and accuracy are essential. Other examples of tools we have created for the community are the following:
<http://gcat.davidson.edu/IGEM06/oligo.html>
<http://gcat.davidson.edu/iGEM07/genesplitter.html>

You will be evaluated on the content on your web page, the functionality and efficiency of your code, and a 10-minute team presentation to the class that describes the unique aspects of your project. The final exam will include questions about both team projects that you will do this semester, so all members of the team should ensure they understand all aspects of the project as it unfolds.

Each of you should plan to spend as many as 20 hours on this project over the next two weeks. You will not have any other work due for this class during that time, and some class time will be devoted to working on the project. DO NOT try to do the whole project in two or three days; your project will be a **disaster**. You can, and probably should, divide the work among you, but you will need to meet and communicate regularly to be sure you are all on track.