

Bioinformatic Analysis

Introduction to the Computational Characterization of Genes and Proteins

Purpose of the Course

This course is an exploration of the emerging field of bioinformatics with a particular emphasis on gaining first-hand experience with how bioinformatics tools are used by investigators in biology. By the end of the course you will have acquired a working knowledge of a diverse range of bioinformatics applications and databases, as well as an understanding of how to interpret the data you find or generate by pursuing the characterization of a hypothetical human gene encoding a protein of unknown function. A major objective is for you to be able to confront the unknown and work through uncertainty, pursue new knowledge using a systematic and thorough strategy, and present what you have learned in a highly effective and professional way. While it's not the primary focus of the class, you should finish the course with a significantly more nuanced appreciation of biological macromolecules, information flow, and cell biology.

Course Description

Bioinformatic analysis is the exploration of molecular sequence, structure and function using online tools and databases. In this class we'll learn to use some of the most powerful tools available for biologists to investigate the nature of genes and proteins. We will each explore a gene and the protein it encodes that no one before us has studied. We will learn to analyze and interpret the diverse forms of bioinformatic data we obtain. And we consider how the data we find allows us to generate and evaluate original hypotheses that can be tested in the laboratory. This is a hands-on course. While the class has no exams, it does require the completion of four problem sets and a summative final project over the course of the semester. It also involves doing some peer review of classmates' work.

This course counts as an elective for the Biology and GCD majors, and fulfills the laboratory course requirement in Biology. This course constitutes a course-based undergraduate research experience (CURE) based on the course project possessing these five qualities: (1) the use of scientific practices, specifically computational biology analyses, (2) discovery of new knowledge and insights, (3) broad relevance of findings with opportunities for impact and action beyond the classroom, (4) collaboration through informal aid, sharing of results, and peer review and, (5) iteration of inquiry, specifically where hypotheses are tested and proposals are developed through the accumulation of evidence over time by repeating studies and by addressing research questions using multiple approaches with diverse methods (see Auchincloss *et al.*, 2014).

Reading

As none of the many currently available books has quite the practical, problems-oriented approach that would be most helpful for a working biologist, there's no assigned text for this class.

I may assign some primary research articles or review articles over the course of the semester and ask you to annotate them with thoughtful questions, answers to others' questions, or comments that illuminate the meaning of the reading for students of bioinformatic analysis. The tool we will use is called Perusall. Like all the analytic tools in this class, it is free for you. It should help facilitate your understanding of the technical material. As this class has no quizzes or exams, I will be looking for two things to come from your reading: (1) significant (≥ 6 substantive questions, answers, or comments on each article assigned and (2) inclusion of concepts from the reading in your solutions to problem set questions. Points will be given for doing each of these.

Problem Solving

There will be four large problem sets over the course of the semester. These problem sets are actually somewhat arbitrary divisions of one single continuous project. Completing these will be essential to the mastery of the subject and the development of confidence in your ability to use bioinformatics in your future work. As much as possible these problems are *authentic* problems, asking you to carry out procedures or analyses which have not been done by others and whose answer is currently unknown (even to your instructor).

Problem sets will be evaluated during individual meetings between each student and me during the week following the Sunday that the problem set is due to the course Canvas site. At these meetings you will talk with me about what you have found by talking with me through a printed copy of your problem set solutions that you will need to bring. Imagine you were making a presentation to your work supervisor or research advisor. *After the meeting I ask that you will submit to Canvas a brief summary of the comments I gave you in the form of a checklist of corrections and improvements you need to make. I use these checklists when I grade the final reports; absence of these post-meeting notes will reduce ones score on that major assignment.*

Laboratory notebook

While we will not be doing wet lab procedures, it is as important as ever for you to keep a record of your work in this course. It will necessarily be a *digital* lab book which you will use to create your problem set solutions. *You should organize list of bioinformatics bookmarks on your browser for each problem set as that is one of the valuable products this class will help you generate.* A good practice is to record in your digital notebook for the class – a word processing file– the *names of files* that contain material worked on a given day.

Workshops

Periodically throughout the semester, in the week preceding a problem set's due date, you will have the chance to discuss your progress on the problem set with small groups of your classmates. It is expected that you bring near-complete solutions to the problem set – as the more you bring the more you will get feedback on – but you need to bring at least partial solutions to receive credit for the workshops. The format for this workshop is that we trade work with the student at your table to the left of you and then with the person at your table to the right of you. You will need to bring two hard copies of your work for them to work from. Feedback will be given in person and also in digital form, uploaded to Canvas. You will be given credit based on the quality of your feedback.

Final Project

The final project has a three components that are completed at different times of the semester. The first part is the oral presentation. You will present select aspects of your work in an **oral presentation** of approximately 12 minutes during *one* of the four symposia that will be held earlier in the semester. *You should bring a single one-page (two-sided is OK) handout with your name, gene/protein name, and some central findings to distribute to each of your classmates and me.*

The second part is the **final written report** on gene/protein. It is both the sum of and an extension of your four problem sets. It is your chance to submit *corrected, revised, and/or expanded solutions* to the questions that appeared on each of the problem sets.

The final project has two new components that must be added. The first additional piece is an **executive summary** (1 double-sided page) of your major findings; like an abstract, this will be near the front of your report. The second new component is a brief articulation of what you see as **the three most interesting hypotheses** that your work has suggested *and* the **lab experiments** (discussed in at least one paragraph each) that would allow you to test these hypotheses.

The third and most important part of your final project is the creation or expansion of the **Wikipedia article** for your chosen gene. You should include information that can be *cited* and understand, if there is information already there, all the information about your gene appears in those cited sources. You might address issues of classes / paralogs of your gene, their possible functions, the proteins known to bind to them, etc. Look to the sites for well-characterized genes for inspiration. See the page for LRCC57 as an example of a student-generated gene page. You should consider beginning this part of your final project early so that you can get registered with Wikipedia and learn the process of contributing an article to a Wikipedia early in the semester. This part of the final project is your original contribution to the world, a way of publishing what you have learned over the course of the semester. It will be important to follow the guidelines for Wikipedia articles on human genes and their proteins.

Course Prerequisites

The only pre-/co-requisite for the course is an introductory course in genetics and cell biology such as the Foundations in Biology courses.

Materials Needed

- A three-ring notebook with dividers for each section of this project. Identify it on the cover with your name, gene/protein name (the most memorable of the names) and a graphic of your choice. Include your name, phone number and email address inside your notebook following "In case this notebook is found, contact:".
- A jump drive (flash drive, thumb drive) for file transfer.
- *Access to a color printer*: Kinkos or the Coffman print center will have this.
- Website bookmarking capability (part of your browser).
- Abstracting tools (such as command-shift-4 on Macs).
- Citation storing program such as Zotero, Endnote, or Mendeley

Evaluation

Assignment	Num	Pts each	Points	Due			
				1	2	3	4
<i>Project Section</i>				1	2	3	4
<i>Workshop</i>				6/18/20	6/30/20	7/9/20	7/19/20
Peer Review of Problem Sets	8	4	32	6/20/20	7/1/20	7/11/20	7/22/20
Problem Sets	4	40	160	6/21/20	7/1/20	7/12/20	7/22/20
Comments Received on Problem Sets	4	3	12	6/24/20	7/3/20	7/15/20	7/24/20
Presentation (<i>one of these/student</i>)	1	20	20	6/23/20	7/2/20	7/14/20	7/23/20
Peer Review of Presentations	4	3	12	6/23/20	7/2/20	7/14/20	7/23/20
Peer Review of Wiki. Article	3	4	12	7/31/20			
Wikipedia Article	1	60	60	8/1/2020 at 5 pm			
Project Report	1	40	40	8/2/2020 at 4 pm			
Course Surveys (for the pair)	1	2	2	8/2/2020 at 8 pm			
Total			350				

Collaboration with other students in the course is highly encouraged at all times except for during the *writing* of your final solutions to problem sets and the *writing* of the drafts of your project. Setting up resources for sharing, online discussions, face-to-face group study and peer review of writing are *all* welcome forms of cooperation and highly encouraged. Such peer interaction will occur in class throughout the semester and on days set aside for "workshop" your results in progress.

Grading

Percentage of total course points required for each grade is as follows:

92-100% = A 90-91.9% = A- 88-89.9% = B+ 82-87.9% = B 80-81.9% = B-
78-79.9% = C+ 72-77.9% = C 70-71.9% = C- 60-69.9% = D <60% = NC

How to Use the Schedule

The syllabus identifies the day of the semester on which each topic will be explored, and the reading that will be relevant to that that topic. The Sunday night before the dates marked in red are when the final version of problem sets are due. It's a good idea to transfer these dates to your personal planner / calendar and also to schedule in extra time to do any necessary catch up. You will have one 20' meeting time to meet with me to discuss the problem set solutions you turned in the preceding Sunday

How to Study for this Course

It is recommended that you read the reading *before* the appropriate class meeting, take note of questions you have or specific topics you feel unclear on, ask questions in class when the topic comes up or the questions appear relevant, come to office hours as often as you can, particularly if you have any questions or any topics that you are confused about, and solve problems related to the week's topic as soon after as possible. Schedule sufficient time to read and more importantly try out and get good at the analyses you learn along the way. Ten hours per week outside of class is about average. Some spend as few as five or six. Others give the class 12 to 15 hours. In any case, it is very likely that you'll enjoy the analyses you do so the time will pass quickly.

Learning Tools

The National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>) provides press access to *PubMed* (abstracts of most professional articles in genetics & molecular biology, and full text versions of many articles).

The NCBI "bookshelf" has electronic (and searchable) versions of several useful molecular biology and bioinformatics texts:

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books>

Special Requests

Cell phones, e-mail, messaging, etc. Please turn off your cell phones before you come to class. We will frequently be online in class. Also, please do not send or read e-mail or instant messages unrelated to this course while you're in class. It's important for all of us to stay focused, especially while learning such a challenging, non-linear skill as bioinformatics analysis.

Instructor Information

David Matthes, Moos Tower 5-220 (Mpls East Bank; in the link between Moos and MCB). Regular office hour: every Tuesday 12:30-1:30 pm. You may email me at dmatthes@umn.edu to arrange to meet with me at other times.

Class Time: Tuesdays & Thursdays, 9:45 am – 12:30 pm

Bioinformatic Analysis

	Day	Topic
6/9/2020 Tues	1	Introduction to bioinformatic analysis & bioinformation How to find a gene/protein about which little is known
6/11/2020 Thurs	2	How to find bioinformation; how to write a Wikipedia page How to evaluate bioinformation quality; peer review <i>Problem Set 0 due in class Thursday</i>
6/16/2020 Tues	3	How to make a global sequence alignment How to compare sequences using a dot matrix
6/18/2020 Thurs	4	How to find similar sequences using BLAST How to find similar sequences using BLAT; <i>Workshop 1</i> <i>Peer feedback due on PS1 due Saturday noon; PS1 due Sunday at midnight</i>
6/23/2020 Tues	5	Symposium 1: Student presentations How to make a multiple sequence alignment <i>Comments received from DJM on PS1 due on Wednesday at midnight</i>
6/25/2020 Thurs	6	How to edit and format a multiple sequence alignment How to reconstruct origins; molecular clocks
6/30/2020 Tues	7	How to determine a gene's full expression pattern How to determine a gene's full expression pattern; <i>Workshop 2</i> <i>Peer feedback due on PS2 due Weds noon; PS2 due Weds at midnight</i>
7/2/2020 Thurs	8	Symposium 2: Student presentations How to find transcribed, spliced & translated sequences <i>Comments received from DJM on PS2 due on Friday at midnight</i>
7/7/2020 Tues	9	How to find promoters and other regulatory sequences How to predict RNA structure and sites of regulation
7/9/2020 Thurs	10	How to analyze a protein's primary sequence How to predict post-translational modification; <i>Workshop 3</i> <i>Peer feedback due on PS3 due Sat. at noon; PS3 due Sunday at midnight</i>
7/14/2020 Tues	11	Symposium 3: Student presentations How to predict protein localization <i>Comments received from DJM on PS3 due on Wednesday at midnight</i>
7/16/2020 Thurs	12	How to predict, explore & annotate 2° structures How to predict, explore & annotate 3° structures
7/21/2020 Tues	13	How to map known mutations to a gene How to identify a protein's functional partners; <i>Workshop 4</i> <i>Peer feedback due on PS4 due Weds. at noon; PS4 due Weds at midnight</i>
7/23/2020 Thurs	14	Symposium 4: Student presentations (<i>All slides due @ midnight</i>) Project and Wikipedia article work time <i>Comments received from DJM on PS4 due on Friday at midnight</i>
7/28/2020 Tues	15	Project and Wikipedia article work time
7/30/2020 Thurs	16	Bioinformatics looking forward & course evaluation <i>Feedback on Wikipedia articles due by Friday midnight</i> <i>All Wikipedia articles open to public viewing by Saturday @ 5:00 pm</i> <i>Final reports uploaded to Canvas by Sunday @ 5:00 pm</i>