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 it 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).

Schedule

6/8/21 Tues	1	Introduction to bioinformatic analysis & bioinformation How to find a gene/protein about which little is known and that connects to COVID19
6/10/21 Thurs	2	How to find bioinformation; how to write a Wikipedia page; how to create data figures How to evaluate bioinformation quality; PS grading rubrics; peer review Problem Set 0 due in class Thursday
6/15/21 Tues	3	How to make a global sequence alignment with nucleic acid and protein sequences How to compare sequences using a dot matrix and find internal repeat structure
6/17/21 Thurs	4	How to find similar sequences using BLAST; how to find homologs in the twilight zone How to find similar sequences using BLAT; how to reconstruct homologs from partial sequences Workshop 1 Peer feedback due on PS1 due Saturday noon; PS1 due Sunday at midnight
6/22/21 Tues	5	Symposium 1: Student presentations How to make a multiple sequence alignment; how to choose sequences to compare and the appropriate multiple MSA-making tool Comments received from DJM on PS1 due on Wednesday at midnight
6/24/21 Thurs	6	How to edit and format a multiple sequence alignment for conservation and chemistry How to reconstruct origins; how to calibrate your protein's molecular clock; how to determine the date of gene duplication for paralogs
6/29/21 Tues	7	How to determine a gene's full expression level and pattern in human tissues How to determine a gene's full expression pattern under different physiological conditions; Workshop 2 Peer feedback due on PS2 due Weds noon; PS2 due Weds at midnight
7/1/21 Thurs	8	Symposium 2: Student presentations How to ascertain monoallelic expression and epigenetic regulation Comments received from DJM on PS2 due on Friday at midnight
7/6/21 Tues	9	How to find promoters and other regulatory sequences including a gene's eQTLs How to predict RNA structure and sites of regulation by miRNA and RNABPs
7/8/21 Thurs	10	How to analyze a protein's primary sequence; how to make a protein diagram How to predict post-translational modification; Workshop 3 Peer feedback due on PS3 due Sat. at noon; PS3 due Sunday at midnight
7/13/21 Tues	11	Symposium 3: Student presentations How to predict protein subcellular and tissue-specific localization Comments received from DJM on PS3 due on Wednesday at midnight
7/15/21 Thurs	12	How to predict, explore & annotate 2° structures How to predict, explore & annotate 3° structures using local and global approaches
7/20/21 Tues	13	How to map known mutations to a gene; how to estimate health impact of variants How to identify a protein's functional partners; how to create protein-lifetime network models for a protein's interactions; Workshop 4 Peer feedback due on PS4 due Weds. at noon; PS4 due Weds at midnight
7/22/21 Thurs	14	Symposium 4: Student presentations (All slides due @ midnight) Project and Wikipedia article work time Comments received from DJM on PS4 due on Friday at midnight
7/27/21 Tues	15	Project and Wikipedia article work time
7/29/21	16	Bioinformatics looking forward & course evaluation

Thurs	Feedback on Wikipedia articles due by Friday midnight All Wikipedia articles open to public viewing by Saturday @ 5:00 pm Final reports uploaded to Canvas by Sunday @ 5:00 pm

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.

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 Zoom meetings between each student and me in the days following when the problem set is due to the course Canvas site. At these meetings you will talk with me for about 20 minutes about what you have found by talking with me through your problem set solutions that you submitted. Imagine that you are making a presentation to your work supervisor or research advisor. *After the meeting I ask that you submit to Canvas, within 24 hours, a brief summary of the comments I gave you in the form of a checklist of corrections and improvements you need to make to your work. 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's 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 in your browser for each problem set as that organized set of online resources 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 have near-complete solutions to the

problem set – as the more you bring the more you will get feedback on – but you need to have at least partial solutions completed to receive credit for the workshops. The format for this workshop is that we trade work with other students present that day. Feedback will be given to the person directly and also 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 (Zoom) presentation** of approximately 12 minutes duration during *one* of the four symposia that will be held earlier in the semester.

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 a few components that are not part of the four problem sets completed earlier in the term. The first additional piece is an **executive summary** of your major findings; like an abstract, this will be near the front of your report. The second new component is your discussion of the most interesting / illuminating variants known to exist in your gene. This work will be done in class after the fourth problem set has been turned in so it will only appear in the final report. The third 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 least one paragraph each) that would allow you to test these hypotheses. The final report, which has undergone four rounds of peer-review and four rounds of instructor-review, is typically between 35 and 55 pages long, including data figures.

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 only include information that can be *cited*. Look to list posted at the class website of well-characterized genes for inspiration. See the page for LRRC57 (https://en.wikipedia.org/wiki/LRRC57) 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.

Evaluation

Assignment	Num	Pts each	Points	Due			
Peer Review of Problem Sets	8	3	24	6/19/21	6/30/21	7/10/21	7/21/21
Problem Sets	4	40	160	6/20/21	6/30/21	7/11/21	7/21/21
Comments Received on Problem Sets	4	4	16	6/23/21	7/2/21	7/14/21	7/23/21
Presentation (*one of these/student)*	1	20	20	6/22/21	7/1/21	7/13/21	7/22/21
Peer Review of Presentations	4	2	8	6/22/21	7/1/21	7/13/21	7/22/21
Peer Review of Wiki. Article	3	4	12	7/30/21			
Wikipedia Article	1	30	30	7/31/2021 at 5 pm			
Project Report	1	30	30	8/1/2021 at 4 pm			
Total			**300**				
Extra Credit							
Course Surveys	1	2	2	8/1/2021 at 8 pm			

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 term on which each topic will be explored in class, and the reading that will be relevant to that that topic. Due dates are also listed on the schedule and in the table above. 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.

How to Study for this Course

It is essential that you watch the posted videos *before* the class meeting where we do those analyses in real time, take note of questions you have or specific topics you feel unclear on, and ask questions via email whenever you have them. Schedule sufficient time to watch (and re-watch) the videos and more importantly try out and get good at the analyses you learn along the way. Twenty to thirty hours per week is about average. 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

Instructor Information

David Matthes, working from home this summer but my office (for future visits ☺) is Moos Tower 5-220. It's on the East Bank; in the link between Moos and MCB.

Office hours: You may email me at dmatthes@umn.edu and I will do my best to reply within 24 hours.

Credit: This is a 4-credit course with the expectation of 24 hours of work per week (including in class time) to earn an average grade, which for this course is a B.

Virtual Class Time: Tuesdays & Thursdays, 8:55 am – 12:30 pm