COURSE ANNOUNCEMENT FOR WINTER 2020

BENG181/CSE 181/BIMM 181
Molecular Sequence Analysis

https://sites.google.com/site/ucsdcse181

Instructor: Pavel Pevzner

- phone: (858) 822-4365
- e.mail: ppevzner@ucsd.edu
- web site: bioalgorithms.ucsd.edu

Teaching Assistants:

- Vikram Sirupurapu (vsirupur@eng.ucsd.edu)
- Adam Jussila (apjussil@eng.ucsd.edu)

Time: 6:30-7:50 Mon/Wed, **Place:** CENTR 109 (seminar Friday 4:00-4:50 PSYNH 106)

Office hours: PP: (Thu 3-5 at EBU3b 4236), TAs (Tue 10-11 AM (CSE B215) and 4-5 PM (CSE B250A) or by appointment)

Prerequisites: The course assumes some prior background in biology, some algorithmic culture (CSE 101 course on algorithms as a pre-requisite), and some programming skills.

Flipped online class. In January 2013, California Governor Jerry Brown announced the initiative to spur online education at UC. Following this initiate, University of California *Innovative Learning Technology Initiative* (ILTI) encourages professors to transform their classes into online offerings available across various UC campuses. Dr. Pevzner is funded by the ILTI and NIH to develop new online approaches to bioinformatics education and to implement flipped classes at UCSD. This class is among the first flipped online classes at UCSD (many more online classes are being produced through the ILTI Program). Lectures in this class are available online rather than presented in the classroom.

The concept of the flipped online class assumes that the students watch lectures at home and do home works before the class starts. This allows the professor and students to interact during the class time when students may ask questions and engage in discussions covering various aspects of the course. We expect that every student will participate in most sessions of the class by either asking questions about the course materials or by answering the questions posed by the instructor or other students.

Automated homework testing. This class provides an automated homework testing environment inspired by the *Rosalind project* (www.rosalind.info) aimed at learning bioinformatics through programming. If you are new to programming, please try the online Coursera Massive Online Open Course (MOOC) *Biology Meets Programming* (https://www.coursera.org/learn/bioinformatics) or EdX MOOC *Introduction to Genomic Data Science* (https://www.edx.org/course/introduction-genomic-data-science-uc-san-diegoxcse181-1x) to prepare yourself for home works that include programming.

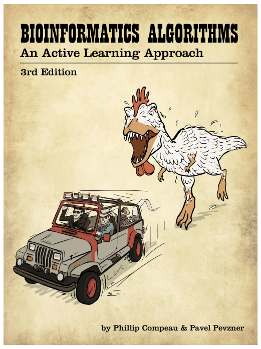
Nearly all HWs in the class will represent programming assignments. Like in real life, there will be no partial credit for programming assignments - you either solve the problem by the deadline (full credit) or not (zero credit). You can use the programming language of your choice to solve the HWs. You will have to submit the code that you developed to a code depository before the deadline.

Quizzes. It is important that you understand the ideas behind each algorithm that you implement in this course. We do not want you to blindly code a “line-by-line" implementation of a pseudocode to pass the automatic grader without learning how the algorithm behind this pseudocode works. That is why we will complement weekly HWs by weekly in-class Quizzes. Similarly to HWs, there will be no partial credit for the Quizzes. You are welcome to discuss your quiz grade (as well as other grades) but make sure that you read the answer key first and submit the request to the TAs within a week of the quiz distribution date.

Avoid anti-correlation between HW and quiz scores! If it turns out that your high HW score anti-correlates with your low quiz score, it likely means that you need to reconsider your approach to completing the HWs. Please note that the Quizzes account for a large fraction of the grade in the class. Moreover, to pass the class, your total quiz score should exceed a minimum threshold, i.e., students with high HW scores but low total quiz score will not be able to pass the class.

Communication skills in bioinformatics (class participation). Communication skills are important in every discipline but they are even more crucial in interdisciplinary fields like bioinformatics. That is why class participation accounts for a large fraction of the total score in this class.

Textbook: Phillip Compeau and Pavel Pevzner. *Bioinformatics Algorithms: An Active Learning Approach*. 3rd edition. Active Learning Publishers 2018.


BIOINFORMATICS ALGORITHMS
An Active Learning Approach
3rd Edition
by Phillip Compeau & Pavel Pevzner

*You need the 3rd edition of the book (*a chicken-dinosaur on the cover) *rather than the outdated 2nd edition.*

The book is available at bioinformaticsalgorithms.org, UCSD bookstore or (a more expensive option) amazon.com. The flipped class assumes that you have instantaneous access to a specific page of the book that is being discussed at a given moment since the questions in the class will be linked to specific pages of the book.

Pens/pencils/laptop. You will need pens/pencils of three different colors and a fully charged laptop with an internet connection for each session of the class. If your laptop has a poor battery, please bring the extension cord and sit near one of the power outlets.

Online resources:

- A link to most lessons is available from the Bioinformatics online specialization web page at coursera.org. Go to "Interactive Text" tab. This is the main piece of the educational infrastructure for this class.
- Our private youtube channel: http://www.youtube.com/user/bioinfalgorithms/
- FAQs: http://bioinformaticsalgorithms.com/faqs.htm
- The link for enrolling in the class on the Rosalind platform is available at the class website:

Use of external packages to solve HWs. You are not allowed to use any external packages (e.g. Numpy, JGraphT, etc.) to solve the HW programming challenges. Using the native implementations of basic data structures (e.g. hashmap/dictionary, array/list, queue, stack, heap, etc.) is fine, but using things like full-fledged graph libraries is not allowed. That being said, you are free to implement your OWN data structures. If you are comfortable approaching problems using an object-oriented coding approach, you can implement your own Node/Edge/Graph classes as you see fit. Make sure that, outside of things you implement yourself, you only use native basic data structures to solve the problems.

Course Website: https://sites.google.com/site/ucsdcse181

Grading: The total score will be composed of the following components:

- HWs (45% of the score). Home works are assumed to be the result of individual work. You are NOT allowed to search for solutions of home works on any online resources. HW submissions will be subjected to the automated plagiarism checking system. Every HW problem is 1 point.
- Quizzes (25% of the score). To pass the class, your total quiz score should exceed the minimum threshold set by the instructor.
- Midterm exam (15% of the score)
- Communication skills in bioinformatics (15% of the score).
- Final exam (Pass or Fail). IMPORTANT: Failing the Final implies failing the class independently of your other scores. The Final will be waived for students with active class participation unless their HW scores anti-correlate with their Quiz/Midterm scores.

Missing classes. We understand that you may miss some sessions of the class due to unforeseen circumstances, illness, or graduate school interviews. To help you deal with these circumstances, your HWs will be computed from your top *n*-1 individual scores where *n* is the total number of HWs in this class. Thus, you can miss one HW in this class, no questions asked, to account for medical or family-related absences, job, and graduate school interviews, etc. Your Quiz and Communication scores will be computed from your top *m* - 2 individual scores where *m* is the total numbers of Quizz and Communication sessions for this class.

Thus, you can miss up to two Quizzes and communication sessions, no questions asked, to account for medical or family-related absences, job, and graduate school interviews, etc. However, you will have to provide official justification for each missing session if you miss more sessions than specified above.

Communication skills (class participation). In a flipped class, the class participation is measured not only by some spontaneous questions that students ask in the class but rather by the questions that students prepare BEFORE the class. The students are expected to learn materials before the class and to prepare a question every time they experience a *learning breakdown.*

An important goal of this class is to teach students how to diagnose their own learning breakdowns and to resolve them by asking well-formulated questions. Learning breakdowns refer to concepts students have struggled with even after spending significant time trying to address this breakdown (e.g., thinking deeply about this concept, checking FAQs and other learning materials, etc.).

Each student who experienced a learning breakdown is required to file a well-formulated question related to his/her breakdowns by 8 p.m. the day before the Discussion deadlines specified below (you will be presenting your questions during the class). It is important that you invest time in formulating the question so that it is clear to all students in the class. During the class, make sure that you present your question in a polished and concise way and speak loud so that all students can hear your question.

When you try to address your learning breakdowns you can form a *tandem* – a group of two students who discuss their learning breakdowns together and try to address them. When you file a question, this activity is assumed to be the result of individual work or a tandem work - please do not share your questions with your classmates outside of your tandem. If you filed a question as a tandem (and in this case, you have to file a single question and include both names), the same score will be assigned to both students in the tandem. We encourage all students to form tandems (so that you can help each other to resolve your learning breakdowns) but it is also perfectly fine to work alone.

There will be nine Discussion deadlines in this class. Please file your questions at the class web site. There will be a different survey link from this page provided in the "Survey" column for each Discussion Deadline. A "well-defined question" means that your peers (and the instructor!) are able to understand the specific difficulty you are having and to help you to overcome the learning breakdown. For example, "I don’t understand how this algorithm works, can you please explain it again?" is not a well-formulated question and will not be given credit for the class participation because it does not describe your *specific* learning breakdown and does allow an instructor to diagnose what caused it.

The only reason a student did not file any breakdown-relevant questions before the deadline is because this student did not experience any learning breakdowns. These students will be answering questions of the instructor and other students during the class. They should be able to answer all FAQs for the relevant Discussion session.

Learning breakdowns versus curiosity questions. Learning breakdowns reflect challenges that make it difficult for a student to understand the follow-up materials. “Curiosity questions" like:

- Are there any alternative ways of estimating the location of the replication origin?
- How do we select the size of the window in the Clump Finding Problem?
- How do we select the constant K for the partial suffix array?
- Why this example assumes that K=5 and not 10?

are not classified as learning breakdowns because they do not affect understanding of the follow-up materials. If you only face curiosity questions while going through the chapter it means you have not really had a breakdown. Also, you cannot file questions that simply repeat “Exercise Breaks," “STOP and Think" boxes, FAQs, or indirectly ask to provide hints for homework problems.

All books have small errors that should not be filed as learning breakdowns unless this error prevents understanding of follow-up materials. If you want to file an error in the book as a learning breakdown, please specify how this error prevented you from understanding the follow-up materials.

Preparing for a discussion session. Each student is required to file a report by 8 p.m. the day before the “Discussion deadlines" specified below. This report describes the level of understanding a student has for each chapter and specifies a learning breakdown that a student is struggling with. The following information is required for each “Discussion deadline:"

- The level of understanding as subjectively evaluated by a student before (parameter *U*) and after (parameter U^*) the class (these parameters vary from 0 to 100 percent). Please update the parameter U^* after the class is over to let us know if the class helped you to better understand the material. If $U^*< 100$, please specify a learning breakdown that remains unresolved. If you had no learning breakdowns, you have to file *U*=100 percent (even if you had “curiosity" breakdowns").
- If $U < 100$:
  - **A:** You are REQUIRED to file a detailed description of the learning breakdown (pointing to a specific page/paragraph in the textbook) and to ask a well-formulated question that will explain to other students how to help you to address your specific learning breakdown. Your breakdown report should be self-contained, i.e., your classmates and the instructor should be able to understand the cause of the breakdown without additional verbal clarifications from you. If you file a breakdown/question that has been already addressed in the available resources, there will be no credit for the class participation.
  - **B:** It is a student’s responsibility to check the FAQs and Charging Stations (not to mention the text of the entire chapter) to ensure that this question has not been addressed yet (otherwise, there will be no credit for the class participation).
  - **C:** Your questions should refer to the book rather than the videos (or power points) since videos represent incomplete and error-prone versions of the learning materials. It is important that the question relates to the specific learning breakdown (and a specific page/paragraph in the book) rather than being an open-ended question. For example, we appreciate the questions like “What is the future of this sequencing technology?” or “Can I apply Hidden

Markov Models to gene prediction? and they will be answered in class. However, no credit for the class participation will be given for such general open-ended questions.

  - **D:** You will be given a class participation credit (**1 point**) for good questions that comply with the above description provided the students and the instructor understand what your question is about.

- If *U* = 100:
  - **A:** If you experienced a learning breakdown that you were able to resolve on your own during preparation to the class, you may file the breakdown report. If you feel that this learning breakdown was caused by a deficiency of the text, file a detailed description of the learning breakdown (breakdown report). Specify the page number /paragraph where the breakdown happened and suggest how the text should be improved to remedy this learning breakdown.
  - **B:** if a student did not experience any learning breakdowns, she/he may file an open-ended question (e.g., What is the future of this sequencing technology?). You will be given a class participation credit (1 point) if you answered questions during the class or provided a well-formulated breakdown report that revealed a deficiency in the textbook.

The instructor may give +1 point extra credit for best questions and best answers presented in the class and subtract -1 if a student filed 100% but did not answer a basic question about the material.

Reviewing online materials. Students can review the online materials at their convenience but should be prepared to answer in-class questions about the materials by the following Discussion deadlines (to be ready for the Q&A session in class). The Study periods below are merely our suggestions to help you get organized for this class - you can work on whatever schedule you find convenient, for example, you can solve all HWs in the first week of classes.

By default, each student should either ask an in-class question or to answer a question posed by the instructor or other students for each chapter. All questions except for the questions on how to solve the upcoming HWs will be answered in class. No hints on how to solve HWs will be provided in the class or during office hours before the HW deadline.

Academic Integrity. To detect instances of academic integrity violations in programming assignments we will use a third-party software. You may find the plagiarism tutorial at the link: https://libraries.ucsd.edu/assets/elearning/cse/cseplagiarismexternal/story.html

All the work in the course should be your own. Since plagiarism was detected in previous sessions of this class (with serious long-term consequences for the students involved), we invest significant effort in checking your code and comparing it with a database of existing solutions. Using various web resources (that provide solutions to coding challenges) for solving HWs is considered a violation of the academic integrity policy. Please do not post your solutions on Internet and do not share your solutions with classmates since it may trigger a violation of the academic integrity policy, for example in the case when your schoolmate uses your solution in a home work. Please note that if you solved a HW before the start of the class (e.g., in the Fall 2019) and used web resources for solving it, it may also trigger a violation of the academic integrity policy. If it is the case, you have to redo the program from scratch since otherwise it may be marked as a violation by our plagiarism checking tool.

Course schedule (subject to change).

Homework deadlines are at 11:59 pm on the specified dates.

Discussion deadlines are at 8 p.m. the day before this communication session starts in the class.

Replication Origin (Chapter 1)

- Discussion deadline: Wed Jan 8 (note a short time to prepare for this session caused by the logistic of the flipped class)
- HW deadline: Tuesday Jan 14
- Quiz: Mon Jan 13
- Study: Mon Jan 6 - Sun Jan 12
- Excluded HW problems: 1L (PatternToNumber) and 1M (NumberToPattern)

Regulatory Motifs (Chapter 2)

- Discussion deadline: Mon Jan 13
- HW deadline: Mon Jan 20
- Quiz: Wed Jan 15
- Study: Thursday Jan 9 - Tue Jan 20

Martin Luther King Day. Mon January 20

Guest lecture *"Computational Analysis of Human Microbiome*" given by Professor Rob Knight at 6:30 pm, Wed January 22 (CENTR 109)

Assembly (Chapter 3)

- Discussion deadline: Wed Jan 22
- HW deadline: Sun Jan 26
- Quiz: Mon Jan 27
- Study: Th Jan 16 - Tue Jan 21
- Excluded HW problems: 3K (Generate contigs) and 3L (a string spelled by a gapped genome path)

Alignment, Part 1 (Chapter 5 before (not including) "Penalizing Insertions and Deletions")

- Discussion deadline: Wed Jan 29
- HW deadline: Sun Feb 2
- Quiz: Mon Feb 3
- Study: Wed Jan 22 - Tue Jan 28

Midterm on Mon Feb 10

Alignment, Part 2 (Chapter 5 from (including) "Penalizing Insertions and Deletions")

- Discussion deadline : Wed Feb. 5

- HW deadline: the standard Sunday deadline for HWs will be postponed till Tuesday Feb 11 to not interfere with the midterm
- Quiz: Wed Feb 12
- Study: Wed Jan 29 - Tue Feb 4

President's Day. February 17

Rearrangements (Chapter 6)

- Discussion deadline: Wed Feb 19
- HW deadline: Sun Feb 23
- Quiz: Mon Feb 24
- Study: Wed Feb 12 - Tue Feb 18
- Excluded HW problems: 6J (2-BreakOnGenomeGraph) and 6K (2-BreakOnGenome)

Detecting Mutations, Part 1 (Chapter 9 before (not including) "Inverting Burrows-Wheeler Transform")

- Discussion: Wed Feb 26
- HW deadline: Sun March 1
- Quiz: Mon March 2
- Study: Wed Feb 19 - Tue Feb 25

Detecting Mutations, Part 2 (Chapter 9 from (including) "Inverting Burrows-Wheeler Transform")

- Discussion deadline: Wed March 4
- HW deadline: Sun March 8
- Quiz: Mon March 19
- Study: Wed Feb 27 - Mon March 2.
- Excluded HW problems: 9P (TreeColoring), 9Q (Partial Suffix Array of a String), and 9R (Suffix Tree from a Suffix Array)

Clustering (Chapter 8)

- Discussion deadline and quiz: Wed March 11
- HW deadline: Friday Mar 13
- Study: Wed March 4 - Tue Mar 10