## COURSE ANNOUNCEMENT FOR WINTER 2021

# 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:**
- Andrey Bzikadze (abzikadze@eng.ucsd.edu)
- Hsuan-lin (Charlene) Her (hsher@eng.ucsd.edu)

**Time:** 6:30-7:50 Mon/Wed, **Place:** online (seminar Friday 4:00-4:50 online)
Zoom link for the class: https://ucsd.zoom.us/j/99782745100
Zoom link for the seminar: https://ucsd.zoom.us/j/96805484881

**Office hours:** PP: (Th 3-5 online), TAs (online Tue 1-2 PM and 4-5 PM or by appointment online)
PP zoom link: https://ucsd.zoom.us/j/96986851791
Andrey Bzikadze zoom link: https://ucsd.zoom.us/j/94881347266
Hsuan-lin (Charlene) Her zoom link: https://ucsd.zoom.us/j/95134947264

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

**Flipped online class**. Starting in 2014, the *Innovative Learning Technology Initiative* (ILTI) at University of California 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 at UCSD. Since 2014, well before the COVID-19 pandemic, all lectures in this class are available online rather than presented in the classroom.

**Multi-university class.** This class closely follows the textbook *Bioinformatics Algorithms: an Active Learning Approach* that has now been adopted by 140+ instructors from 40+ countries. Since all these instructors cover similar materials, we have decided to launch a multi-university bioinformatics class that will bring various instructors together to enhance the educational experience of students across all participating universities.

In Winter 2021, this online multi-university bioinformatics class will be simultaneously offered by the following instructors:

- Phillip Compeau, Department of Computational Biology, **Carnegie Mellon University**

- Alexey Gurevich, Center for Algorithmic Biotechnology, **Saint Petersburg University**
- Pavel Pevzner, Department of Computer Science and Engineering, **University of California at San Diego**
- Ben Raphael, Computer Science Department, **Princeton University**
- Steven Salzberg and Rachel Sherman, Departments of Biomedical Engineering, Computer Science, and Biostatistics, **John Hopkins University**

This educational experiment will be directed by Professor Niema Moshiri at University of California at San Diego, an expert in online education technologies (see his TEDx talk *The Era of Online Learning*).

**Automated homework testing.** This class provides an automated homework testing environment inspired by the *Rosalind project* aimed at learning bioinformatics through 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 a 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.

This year, we will move from the Rosalind platform to a more advanced *Stepik* platform. You will have to submit all your solutions to the HW programming challenges at Stepik. The code that you used to generate the successful solutions on Stepik should be uploaded to Canvas Assignments section prior to the corresponding deadline.

No hints on how to solve HWs will be provided before the HW deadline.

**Solving HWs.** 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 pseudocode to pass the automatic grader without understanding how the algorithm behind this pseudocode works. That is why the midterm and the final in this class will be designed to test how well you understand your OWN previously submitted HW programs (see below).

**Midterm and Final.** The midterm and final exams will consist of the newly designed programming challenges that represent modifications of problems that have been previously given as HWs in the class. Therefore, to solve a novel problem A* that originated from a HW A, you merely need to slightly modify the original code for A (that you have already submitted). If you got a HW credit for problem A, you **are required** to modify your submitted code for A instead of implementing A* from scratch and mark all new lines where you made changes as compared to the code for A. Thus, you need to have the codes for all your previous HWs available when you start midterm/final. If you haven’t got a HW credit for problem A, you can solve problem A* from scratch.

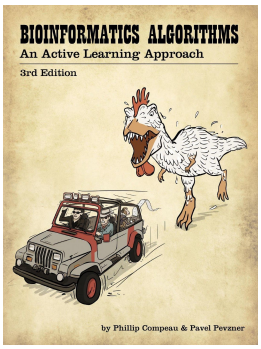
This will be a simple task for students who followed the suggestions in the previous subsection “Solving HWs.” However, it will turn into a difficult task for students who submitted HWs without a deep understanding of the algorithms behind them or (worse!) for students whose HWs do not represent an independent effort. All midterm and final solutions

will be checked for plagiarism.

**Avoid anti-correlation between HW and midterm/final scores!** If it turns out that your high HW score anti-correlates with your low midterm/final score, it likely means that you have not taken the above subsection seriously or worse, the HWs you submitted do not represent independent work. That is why, to pass the class, your final score should exceed a minimum threshold, i.e., students with high HW scores but low final score will not be able to pass the class. The midterm and final will be designed in such a way that you *will have to* modify the code in one of your previously submitted HWs to solve each of the problems. Therefore, the best way to prepare for the midterm/final and to pass this class is to make sure that you submit your *own HWs* each week and understand how it works.

**Communication skills in bioinformatics.** Communication skills are important in every discipline but they are even more crucial in interdisciplinary fields like bioinformatics. That is why communication skills account 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.




You need the 3rd edition of the book (a chicken-dinosaur on the cover) rather than the outdated 2nd or 1st editions.

**Although it was a required textbook for CSE 181 in the past, this year you DO NOT NEED to buy the hard copy of this textbook.** Since students are now located in various countries where the book may not be available, Active Learning Publishers partnered with Stepik to make the book (along with all additional materials that include videos, programming challenges, etc.) available as an online MOOCBook at the discounted price $69.95.

Please go to BioinformaticsAlgorithms@UCSD at Stepik to enroll in the course.

**Online resources:**

The Stepik platform provides all online resources for this course. You can also find some of these materials at:

- A link to most lessons is available from the Bioinformatics online specialization web

page at coursera.org. Go to the "Interactive Text" tab. This is the main piece of the educational infrastructure for this class.

- Private youtube channel: http://www.youtube.com/user/bioinfalgorithms/
- FAQs: http://bioinformaticsalgorithms.com/faqs.htm
- Rosalind platform

**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, e.g., 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** (55% of the score). Homeworks 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 automatic plagiarism checking. Every HW problem is 1 point.
- **Midterm** exam (15% of the score)
- **Communication skills** in bioinformatics (20% of the score).
- **Quizzes based on invited and research-oriented lectures** (10% of the score).
- **Final** exam (Pass or Fail). IMPORTANT: Failing the Final implies failing the class independently of your other scores.
- **Optional research-oriented class project** (not required). See information about bonus points below.

**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 overall HW score (Communication score) will be computed from your top $n$-1 individual scores, where $n$ is the total number of HWs (Communication sessions) in this class. Thus, you can miss one HW (Communication session) in this class, no questions asked, to account for medical or family-related absences, job, and graduate school interviews, etc. However, you will have to provide an official justification for each missing session if you miss more HWs (Communication sessions) than specified above.

**Learning breakdowns.** An important goal of this class is to teach students how to diagnose their INDIVIDUAL *learning breakdowns* and to resolve them by asking well-formulated questions. A learning breakdown refers to a concept that 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.).

**Communication skills.** Each student who experienced a learning breakdown is required to file a well-formulated question related to their breakdowns by 8 p.m. the day before the Communication Session deadlines specified below. It is important that you invest time in formulating the question so that the instructor can help you to address it based *only on your*

*formulation*, without additional clarification.

There will be nine Communication Session deadlines in this class. Please file your questions at the class website. There will be a different survey link from this page provided in the "Survey" column for each Communication Session 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 communication session because it does not describe your *specific* learning breakdown and does allow an instructor to diagnose what caused it.

The only reason a student should file any breakdown-relevant questions is that this student did not experience any learning breakdowns. These students will be answering questions of the instructor and other students during individual zoom meetings or via email. At a minimum, they should be able to answer all FAQs for the relevant Communication Session session.

**Tandems.** 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 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 OK to work alone.

**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 does this example assume that $K$=5 and not 10?

are not classified as learning breakdowns because they do not affect the 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 Communication Session session**. Each student is required to file a report

by 8 p.m. on the of the “Communication Session 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 “Communication Session deadline:"

- The level of understanding as subjectively evaluated by a student before (variable *U* that varies from 0% to 100%) the class. If you had no learning breakdowns, you have to file *U*=100% (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) and to ask a well-formulated question that will explain to the instructor how to help you to address this specific learning breakdown. Your breakdown report should be self-contained, i.e., 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 textbook 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 textbook) rather than being an open-ended question. For example, we appreciate 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 the follow-up communication session. 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 instructor understands what your question is about.
- If *U* = 100:
    - **A:** if you did not experience any learning breakdowns, your knowledge of the material will be checked either in an individual zoom session or by sending you additional questions via email. You will be given a class participation credit (1 point) if you are able to answer these questions. At minimum, you should be able to answer any question listed in the relevant FAQs (see above for a link to the FAQs).

The instructor may give +1 extra credit for best questions and best answers and subtract -1 if a student filed 100% but was not able to answer a basic question about the material.

**Filing a breakdown report.** Please address the following points when you file the breakdown report:

- Specify the exact paragraph in the book where you experienced a learning breakdown. You can copy the first sentence of this paragraph from the online book or (if you have the hard copy) provide the page and paragraph numbers.

- Confirm that you checked all FAQs to see whether one of them addresses your individual learning breakdown.
- Provide a concise and polished description of your learning breakdown. If the instructor does not understand your description, you failed to communicate what went wrong. In this case, the instructor cannot help you to resolve the breakdown and you get no points for the communication session.
- Specify which specific follow-up concept you failed to understand because of your learning breakdown.

**Reviewing online materials.** Students can review the online materials at their convenience (all materials are available on the first day of classes) but should be prepared to answer questions about the materials by the Communication Session deadlines specified below. The Study periods below are merely the 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.

**Academic Integrity.** To detect instances of academic integrity violations in programming assignments we will use a third-party plagiarism detection software. You may find the tutorial "Plagiarizing the source code" 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 the 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 homework. Please note that if you solved a HW before the start of the class (e.g., in Fall 2020) 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.

**Class Project (optional).** This class offers an optional project aimed at genome assembly using the recently emerged HiFi technology based on long and accurate reads. See the multi-university class web page for more information on the class project. If you plan to form a project team, we recommend that you cover Chapter 3 (Genome Assembly) in the very beginning of the quarter. If you decide to participate in the class project, please communicate the list of students and the name of your team to TAs by February 10.

The team that developed the best assembler will get 40% bonus points, and the remaining teams will get up to 20% bonus points depending on the results.

**Research-oriented lectures.** In addition to lectures covering the textbook material, we will have research-oriented lectures by professors in the multi-university class:

- Phillip Compeau (CMU): ***How do we measure gene expression: transcript assembly and quantification.***
    - This lecture is related to the topics of sequence comparison (chapter 5), clustering (chapter 8), and read mapping (chapter 9).
    - Time: March 1
- Alexey Gurevich (SPBU): ***How do we compare LONG genomic sequences?***
    - This lecture is related to the topics of genome assembly (chapter 3) and sequence comparison (chapter 5).
    - Time: February 1
- Pavel Pevzner (UCSD): ***The long-read revolution in genome sequencing.***
    - This lecture is related to the topic of genome assembly (chapter 3).
    - Time: January 20
- Ben Raphael (Princeton): ***Cancer evolution.***
    - This lecture is related to the topics of genome assembly (chapter 3), genome rearrangements (chapter 6), and evolutionary tree construction (chapter 7)
    - Time: February 17
- Rachel Sherman (JHU): ***What's in a mutt: an intro to dog DNA analysis.***
    - This lecture is related to the topic of the Hidden Markov Models (chapter 10).
    - Time: March 8

In addition, we will have two COVID-19-related lectures that discuss how bioinformatics contributes to emerging problems in personalized immunogenomics and antibody discovery:

- Pavel Pevzner (UCSD). ***Personalized immunogenomics***
    - Time: January 20
- Stefano Bonissone (Digital Proteomics). **Immunoinformatics: how informatics can be central to antibody discovery and characterization.**
    - Time: February 8

Last but not least, we will have a lecture that describes an open-ended biological problem and raises an important question on how to transform it into a well-defined computational problem:

- Arcady Mushegian (National Science Foundation). ***The Minimal Genome Size problem***
    - Time: January 13

**Course schedule** (subject to change). Below is the class schedule (information about the schedule of the research-oriented lectures will be added later).

Homework deadlines are at 11:59 pm on the specified dates. Please note that HWs do not necessarily include ALL PROBLEMs from a given chapter. Read information below for the list of excluded problems for each chapter.

*Replication Origin* (Chapter 1)

- Communication Session survey deadline: Friday Jan 8, 8pm

- HW deadline: Tuesday, Jan 12
- Study: Mon Jan 4 - Tuesday, Jan 12
- Excluded HW problems: 1L (PatternToNumber) and 1M (NumberToPattern)
- Wed, January 6. Pavel Pevzner (UCSD) ***Personalized immunogenomics***

*Regulatory Motifs* (Chapter 2)
- Communication Session survey deadline: Friday, Jan 15, 8pm
- HW deadline: Tuesday, Jan 19
- Study: Tuesday, Jan 12 - Tue Jan 19
- Wed. January 13. Arcady Mushegian (Program Director, National Science Foundation). ***The Minimal Genome Size problem***

*Martin Luther King Day*. Monday, January 18

*Assembly* (Chapter 3)
- Communication Session survey deadline: Friday, Jan 22, 8pm
- HW deadline: Tuesday, Jan 26
- Study: Tuesday, Jan 19 - Tue Jan 26
- Excluded HW problems: 3K (Generate contigs) and 3L (a string spelled by a gapped genome path)
- Wed, January 20. Pavel Pevzner (UCSD) ***Personalized immunogenomics.***
- Mon, January 25. Pavel Pevzner (UCSD) ***The long-read revolution in genome sequencing.***

*Alignment, Part 1* (Chapter 5, ending before (not including) “Penalizing Insertions and Deletions")
- Communication Session survey deadline: Friday, Jan 29, 8pm
- HW deadline: Tuesday, Feb 2
- Study: Tuesday, Jan 26 - Tue Feb 2
- Mon, February 1. Invited talk: Alexey Gurevich (Saint Petersburg University) ***How do we compare LONG genomic sequences?***

*Alignment, Part 2* (Chapter 5, starting from (including) “Penalizing Insertions and Deletions")
- Communication Session survey deadline: Friday, Feb. 5, 8pm
- HW deadline: Tuesday, Feb 9
- Study: Tuesday, Feb 2 - Tue Feb 9
- Mon. Feb 8. Stefano Bonissone (Digital Proteomics). ***Immunoinformatics: how informatics can be central to antibody discovery and characterization.***

***Midterm* on Wed Feb 10**

**The last day to register your team for the optional HiFiAssembler project**

*President’s Day*. Monday, February 15

*Rearrangements* (Chapter 6)
- Communication Session survey deadline: Friday, Feb 12, 8pm

- HW deadline: Tuesday, Feb 16
- Study: Tuesday, Feb 9 - Tue Feb 16
- Excluded HW problems: 6J (2-BreakOnGenomeGraph) and 6K (2-BreakOnGenome)
- Wed, February 17. Invited talk: Ben Raphael (Princeton) ***Cancer Evolution***

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

- Communication Session survey deadline: Friday, Feb 19, 8pm
- HW deadline: Tuesday, Feb 23
- Study: Tuesday, Feb 16 - Tue Feb 23

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

- Communication Session survey deadline: Friday, February 26, 8pm
- HW deadline: Tuesday, March 2
- Study: Tuesday, Feb 23 - Tue March 2
- Excluded HW problems: 9P (TreeColoring), 9Q (Partial Suffix Array of a String), and 9R (Suffix Tree from a Suffix Array)
- Mon, March 1. Invited talk: Phillip Compeau (Carnegie Mellon University) ***How do we measure gene expression: transcript assembly and quantification.***

*Clustering* (entire Chapter 8) and *Hidden Markov Models* (Chapter 10) before (not including) "Classifying proteins with profile HMMs."

- Communication Session survey deadline: Friday, March 5, 8pm
- HW deadline: Tuesday, Mar 9
- Study: Tuesday, March 2 - Tue Mar 9
- Excluded HW Problems. Implement all problems from Chapter 8 except for 8E (Hierarchical Clustering). **Do not** implement any problems from Chapter 10 except for the only problem 10C (Viterbi algorithm).
- Mon, March 8. Invited talk: Rachel Sherman (Johns Hopkins University) ***What's in a mutt: an intro to dog DNA analysis.***

**Final:** Wed, March 17, 7 pm -10 pm