Special Topics in Data Science:

Algorithms for Computational Biology

Dr. Dan DeBlasio
Department of Computer Science
University of Texas at El Paso

CS 4364/5364 – Spring 2022

Course Description: This course will cover the algorithms that make modern computational biology and bioinformatics possible. The plan is to cover both foundational algorithms such as sequence alignment, as well as their modern applications in solving problems such a genome assembly. The focus of this course is on how computer scientists apply their knowledge to frame a computational problem inspired by a specific real-world problem and to solve such computational problems. In addition to standard algorithm development, the course will cover the influence of convex optimization (mainly integer linear programming) and machine learning on computational biology. The course assumes no previous knowledge in biology or genetics. The course will build on and enhance students' basic understanding of the principle of algorithm design and analysis by applying such principles in the context of bioinformatics.

Course Objectives: This course is designed to study algorithm design and analysis in the context of problems related to computational biology. After the course concludes successful students will not only have a deeper understanding of algorithms, but a taste for the techniques used to convert real-world problems into computational ones; as well as common strategies on solving them.

Prerequisite: CS 2302 or instructor approval.

Knowledge and Abilities Required Before Entering the Course: Students are assumed to be comfortable with basic algorithm design and analysis. One of the major recurring themes will be algorithm running time and memory consumption improvement. Students should also be familiar with common problem solving techniques in particular dynamic programming. A knowledge of basic machine learning concepts (such as training/testing test construction, etc) will be helpful as well, though not required.

Tentative list of topics covered this semester:

- Pairwise Sequence Alignment
- Multiple Sequence Alignment
- Genome Assembly
- Metagenomic & Alignment-free Genomic Analysis
- Phylogenetic Reconstruction
- Integer Linear Programming Applications
- Machine Learning Applications

The remainder of this document is subject to change, notice will be made to students. Changes after the first day of class will be labeled ~~like this~~ as shown here.

Contents

1 Logistics

Synchronous course session times:

- TR 4:30pm-5:50pm

Textbook: We will use a combination of a textbook (details below) and primary literature. The textbook we will use is "Algorithms in Bioinformatics: A Practical Approach" by Wing-Sun Kim. PDFs of additional material will be provided as needed on the course website.

Though not required, other helpful texts are:

- "Algorithms on Strings Trees and Sequences" by Dan Gusfield (on hold at the UTEP library)
- "Algorithms for Next Generation Sequencing" by Wing-Sun Kim
- "Bioinformatics Algorithms: An Active Learning Approach" by Phillip Compeau and Pavel Pevzner
- "Integer Linear Programming for Computational and Systems Biology" by Dan Gusfield
- "Biological Modeling and Simulation" by Russell Schwartz

Communication platforms:

- **Course Website** – `specialtopics.deblasiolab.org/s22/` – Used for course announcements, paper distribution, etc.
- **MS Teams** – `specialtopics.deblasiolab.org/s22/teams` – Used for office hours and intra-class discussions. Several channels will be available in the team for asking and answering questions, the instructional staff will answer questions posted on teams, but other students are encouraged to provide feedback as well.
- **Blackboard** – `specialtopics.deblasiolab.org/s22/blackboard` – Used to disseminate grades and submit assignments.

1.1 Tentative Schedule

Dates	Tuesday Topic	Thursday Topic
January 18 & 20	Introduction & Algorithms Refresher	Algorithms Refresher (cont.) & ILP Intro
January 25 & 27	Molecular Biology Primer (Sung Ch. 1)	Molecular Biology Primer (cont.)
February 1 & 3	Global Pairwise Sequence Alignment (§2.1-2.2)	Global Pairwise (cont.)
February 8 & 10	Local Pairwise Alignment (§2.3)	Gap Penalties (§2.5)
February 15 & 17	Suffix Trees and Applications (§3.1-3.4)	Suffix Trees (cont)
February 22 & 24	Suffix Arrays (§3.5)	LCS & solution by ILP
March 1 & 3	Multiple Sequence Alignment (§6.1-6.2)	MSA Methods (§6.4-6.5)
March 8 & 10	Progressive Alignment (§6.6)	Parameter Configuration
March 15 & 17	**Spring Break**	
March 22 & 24	Genome Alignment (§4.1-4.4)	Database Search (§5.1-5.4)
March 29 & 31	Advanced Database Search (§5.5-5.8)	Advanced Database Search (cont.)
April 5 & 7	Motif Finding (§10.1-10.7)	Motif Finding (cont.)
April 12 & 14	Phylogeny (§7.1-7.3)	Phylogeny (cont.) & ILP for perfect phylogeny
April 19 & 21	Reference-based Genome Assembly	Reference-based Genome Assembly (cont.)
April 26 & 28	*de novo* Genome Assembly	*de novo* Genome Assembly (cont.)
May 3 & 5	Metagenomics	Metagenomics (cont.)

2 Instructional Staff

2.1 Instructor

Dr. Dan DeBlasio
email: dfdeblasio@utep.edu
chat on MS Teams: teamsChat.deblasiolab.org (direct message)
office: CCSB 3.1008 or line via MS Teams
office hours: ~~tbd~~ MW 2-3pm
or appointment calendly.deblasiolab.org

3 Expectations

Communication: Students are expected to consult their emails *daily*, and to answer these as relevant.

Class Participation: Regular attendance is essential and expected. Due to the high emphasis of group discussion and dialogue all students are discouraged from missing classes. Missed course meetings will be noted and chronic absences may impact the students grade if not discussed with the course instructor.

It is the student's responsibility to review the content covered during missed class(es), as well as the assignments given during their absence. Participation points also include completing post-lecture online quizzes (when requested) that are administered as surveys to monitor students' overall progress and potential struggles.

Collaboration The course project and homeworks are meant to expose the student to the topics being discussed. While each student is responsible for their individual projects and homeworks; cooperation and collaboration between students is highly encouraged but plagiarism will not be tolerated.

4 Grading

Grades are communicated to students in a timely manner. It is the students' responsibility to keep track of their grades by compiling the grades they receive. The approximate percentages are as follows:

65%	Homeworks
10%	Midterm Exam
15%	Term Paper/Wikipedia Entry
10%	Participation

The base percentage-score-to-letter-grade conversion for this course is as follows:

90%	or higher is guaranteed an A
80%	or higher is guaranteed a B
70%	or higher is guaranteed a C
60%	or higher is guaranteed a D
	all lower grades are an F

These minimums may be lowered without notice but will not be raised.

4.1 Homework

Grading for homework and exam questions is roughly according to the following scheme:

- correct solution idea and the right technical execution — >90%,
- correct idea but with errors in its execution — >80%.
- wrong idea and errors in its execution, but demonstrating comprehension of the material — >70%.
- wrong idea, errors in execution, and deficiencies in comprehension — ∼60%,
- relevant work that shows no understanding — ∼50%.

Writing an answer that *relates to the question* guarantees at least 50% of the points for the question, no points are awarded for writing nothing (or for anything unrelated to the question asked). *While writing coherent and concise responses to homeworks questions is important, students have the choice (and the instructor reserves the right to request) a discussion of solutions as a supplement to submitted responses.*

On homework, very-high-level ideas can be discussed with friends, but solutions must represent individual work and must be written up separately. Any material from the Internet that is used in a solution must be cited by its URL; to not cite it is plagiarism, which is considered cheating.

Students enrolled in the graduate course will have higher expectations on the homework assignments than those in the undergraduate class. These will be defined in each assignment.

4.2 Exams

Exams will be comprehensive. All material presented in class (including during discussions and project presentations) and those in homework and assigned readings will be included.

4.3 Term Paper/Wikipedia Entry

In place of a final exam students will write a term paper or edit a wikipedia entry.

The term paper should be 7-10 pages plus all relevant citations. Topics can be chosen from the material *related to* the course but cannot be simply a topic we have already covered.

The term paper can also be done as part of the International Society for Computational Biology wikipedia competition. This would need to either be a *significant* improvement in the topic if it already exists or create a new article on a relevant topic.

The level of quality expected for both the term paper and the wikipedia article will be the same.

The possible topics include:

- universal hitting sets
- hashing schemes for comp. bio.
- transcript assembly tools or techniques
- HMMs for alignment
- etc.

The topic can be related to your ongoing research, but should be distinct from any previously developed writing. Your choice will be solicited from the instructor sometime around midterms (when the final project is discussed).

Students enrolled in the graduate course will have higher expectations on this item than those in the undergraduate class.

In leu of the paper, an alternate hands-on project can be developed *at the discretion of the instructor*.

4.4 Extra Credit

Related Talks — Attending seminars related to the course can count as extra points toward a student's participation grade. Up to 10% of credit can be applied to homework grade, up to a maximum of 70% of the student's total grade. That is the homework grade will be equal to max {(average of homework assignments + extra credit) , 70%}.

Sources of talk opportunities can be found via:

- International Society of Computational Biology (ISCBAcademy) – https://www.iscb.org/iscbacademy-webinars
- #BlackInCompBio Series – https://www.blackwomencompbio.org/events

To receive credit a student must give proof of attendance (screenshot of the talk, etc.) and submit a 1 paragraph summary of the major points presented. Talks from other sources can be considered with prior approval of relevance.

5 Standing in the course

Students will have access to their grades for all assignments so that they can self-monitor their standing and progress. However, it is also completely fine for any student to come and talk to their instructor about their standing and work together to make sure the student is as successful as can be.

Dropping the Course: Every semester, some students drop the course. We, instructors, completely understand and respect that. We only hereby ask students to inform us, ideally before, but in the worst-case right after, of their intention to drop the course. This is

really important for us as it possibly informs us of ways in which to better serve our students.

6 Special notices for COVID-19

The following are a summary of the universities policies regarding COVID-19.

You must STAY OFF CAMPUS and REPORT if you: (1) have been diagnosed with COVID- 19, (2) are experiencing COVID-19 symptoms, or (3) have had recent contact with a person who has received a positive coronavirus test. Reports should be made at screening.utep.edu. If you know anyone who should report any of these three criteria, encourage them to report. If the individual cannot report, you can report on their behalf by sending an email to COVIDaction@utep.edu.

For each day that you attend campus—for any reason—you must complete the questions on the UTEP screening website (screening.utep.edu) prior to arriving on campus. The website will verify if you are permitted to come to campus. Under no circumstances should anyone come to class when feeling ill or exhibiting any of the known COVID-19 symptoms. If you are feeling unwell, please let me know as soon as possible, and alternative instruction will be provided. Students are advised to minimize the number of encounters with others to avoid infection.

Wear face coverings when in common areas of campus or when others are present. You must wear a face covering over your nose and mouth at all times in this class. If you choose not to wear a face covering, you may not enter the classroom. If you remove your face covering, you will be asked to put it on or leave the classroom. Students who refuse to wear a face covering and follow preventive COVID-19 guidelines will be dismissed from the class and will be subject to disciplinary action according to Section 1.2.3 Health and Safety and Section 1.2.2.5 Disruptions in the UTEP Handbook of Operating Procedures.

7 Resources

Special Accommodations: If you have a disability and need classroom accommodations, please contact the Center for Accommodations and Support Services (CASS) at 747-5148 or by email to cass@utep.edu, or visit their office located in UTEP Union East, Room 106. For additional information, please visit the CASS website at www.sa.utep.edu/cass. CASS' staff are the only individuals who can validate and if need be, authorize accommodations for students with disabilities.

Scholastic Dishonesty: Any student who commits an act of scholastic dishonesty is subject to discipline. Scholastic dishonesty includes, but not limited to cheating, plagiarism, collusion, and submission for credit of any work or materials that are attributable

to another person.

Cheating is:

- Copying from the test paper of another student
- Communicating with another student during a test to be taken individually
- Giving or seeking aid from another student during a test to be taken individually
- Possession and/or use of unauthorized materials during tests (i.e. crib notes, class notes, books, etc.)
- Substituting for another person to take a test
- Falsifying research data, reports, academic work offered for credit

Plagiarism is:

- Using someone's work in your assignments without the proper citations
- Submitting the same paper or assignment from a different course, without direct permission of instructors

  To avoid plagiarism, see:
  https://www.utep.edu/student-affairs/osccr/_Files/docs/Avoiding-Plagiarism.pdf

Collusion is:

- Unauthorized collaboration with another person in preparing academic assignments

Important!

When in doubt on any of the above, please contact your instructor to check if you are following authorized procedure. Also, please check the UTEP's Handbook of Operating Procedures at: hoop.utep.edu.