Special Topics in Data Science:

Algorithms for Computational Biology

Spring 2021

CS 4364 & 5364

Objective

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.

Base Knowledge

Students should be

- comfortable with basic algorithm design and analysis
  - algorithm running time
  - memory consumption improvement
- 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.

Base Knowledge

Students should be

- comfortable with basic algorithm design and analysis
  - algorithm running time
  - memory consumption improvement
- familiar with common problem solving techniques in particular ***dynamic programming***

If you're rusty on these, thats okay we will be reviewing them this week

A knowledge of basic machine learning concepts (such as training/testing test construction, etc) will be helpful as well, though not required.

Logistics

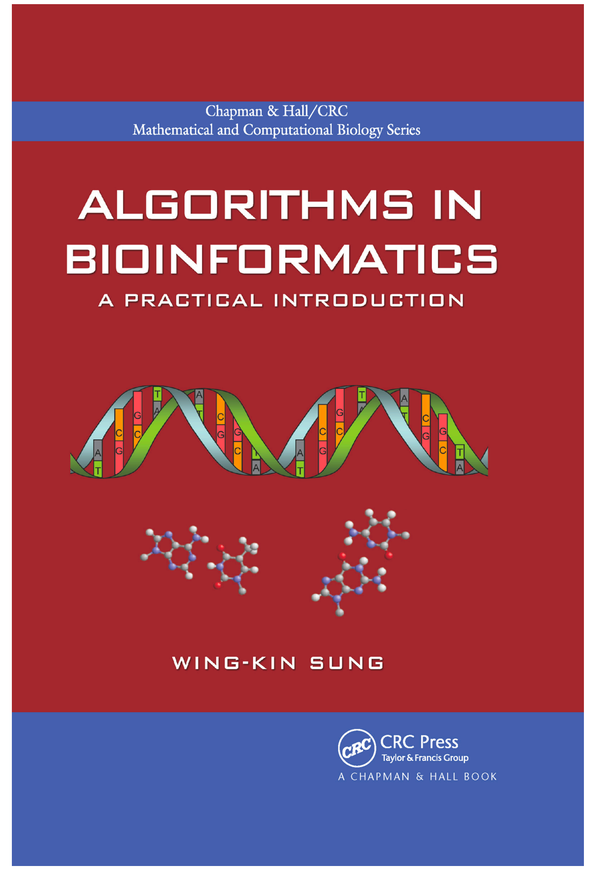
Class: 3pm Tuesday and Thursdays on Zoom
(but you're already here!)

Book: "Algorithms in Bioinformatics: A Practical Approach" by Wing-Sun Kim

- We won't dig into the book until next week at the earliest

Communication

- **email** (yes I'm old)
- specialtopics.deblasiolab.org/s21/
- specialtopics.deblasiolab.org/s21/teams
- specialtopics.deblasiolab.org/s21/youtube
- specialtopics.deblasiolab.org/s21/officehours



Contacting me

Dr. Dan DeBlasio	email:	dfdeblasio@utep.edu
	chat on MS Teams:	teamsChat.deblasiolab.org (direct message)
	office:	~~CCSB 3.1008~~ (not applicable until further notice)
Office Hours	**M 2:00p-3:00p**	specialtopics.deblasiolab.org/s21/officehours
	R 1:00p-2:00p	[or "Office Hours" on the class team]
	or appointment:	calendly.deblasiolab.org

Topics & Tentative Schedule

Dates	Tuesday Topic	Thursday Topic
January 19 & 21	Introduction & Algorithms Refresher	Algorithms Refresher (cont.) & ILP Intro
January 26 & 28	Molecular Biology Primer (Sung Ch. 1)	Molecular Biology Primer (cont.)
February 2 & 4	Global Pairwise Sequence Alignment (§2.1-2.2)	Global Pairwise (cont.)
February 9 & 11	Local Pairwise Alignment (§2.3)	Gap Penalties (§2.5)
February 16 & 18	Suffix Trees and Applications (§3.1-3.4)	Suffix Trees (cont)
February 23 & 25	Suffix Arrays (§3.5)	LCS & solution by ILP
March 2 & 4	Multiple Sequence Alignment (§6.1-6.2)	MSA Methods (§6.4-6.5)
March 9 & 11	Progressive Alignment (§6.6)	Parameter Configuration
March 16 & 18	**Spring Break**	
March 23 & 25	Genome Alignment (§4.1-4.4)	Database Search (§5.1-5.4)
March 30 & April 1	Advanced Database Search (§5.5-5.8)	Advanced Database Search (cont.)
April 6 & 8	Motif Finding (§10.1-10.7)	Motif Finding (cont.)
April 13 & 15	Phylogeny (§7.1-7.3)	Phylogeny (cont.) & ILP for perfect phylogeny
April 20 & 22	Reference-based Genome Assembly	Reference-based Genome Assembly (cont.)
April 27 & 29	*de novo* Genome Assembly	*de novo* Genome Assembly (cont.)
May 4 & 6	Metagenomics	Metagenomics (cont.)

Grading

50% Homework
10% Midterm Exam
30% Term Paper/Wikipedia Article
10% Participation

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).

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.

Term Paper/Wikipedia Entry

Two choices:

- 7-10 page term paper
- create/update a related wikipedia article

Topic should be related to class, but not directly a topic we discuss in detail.

- Should cite primary sources (not textbooks)
- I will ask for topics to be submitted around the midterm

Wikipedia entries can also be entered into the ISCB Wikipedia Competition (cash prize), but this is not required.

- All wiki entries will be drafted on my lab/class wiki first.

Extra Credit

Attending Talks

- Since we're online many Computational Biology seminars are as well
- To earn credit submit: proof and a 1 paragraph description with some details you found interesting
- Possible series:
  - International Society of Computational Biology (ISCBAcademy) – https://www.iscb.org/iscbacademy-webinars
  - #BlackInCompBio Series – https://www.blackwomencompbio.org/events
  - Others with *prior* approval