SDS 348/385—Computational Biology and Bioinformatics

Spring 2020
Unique 55615: TTH 9:30–11am UTC 3.112, W 9am–10am FAC 101B
Unique 55620: TTH 9:30–11am UTC 3.112, W 10am–11am FAC 101B
Unique 55700: TTH 9:30–11am UTC 3.112, W 10am–11am FAC 101B (SDS 385, with graduate course credit)

Instructor: Claus O. Wilke
Email: wilke@austin.utexas.edu
Office: MBB 3.232
Office Hours: Thurs. 11:30am–1:30pm, or by appointment
Twitter: @clauswilke

Teaching Assistant: Rachael Cox
Email: rachaelcox@utexas.edu
Office: MBB 3.304
Office Hours: Fri. 9am–11am, or by appointment
Twitter: @rachaelmcox

Purpose and contents of the class

In this class, students will learn the basic skills required to handle the kind of data sets current-day working biologists will encounter. Because any kind of large-scale, automated data analysis requires programming skills, a substantial component of this class will be dedicated to learning how to program in the two languages most commonly used by computational biologists, R and Python. The class will also put substantial emphasis on good data management practices, on data visualization, and on interpreting the patterns that are seen in the data. Finally, several commonly encountered data-analysis problems in computational biology will be discussed, such as clustering data into groups, searching for gene sequences in related organisms, or building phylogenetic trees.

Prerequisites

The class requires no prior knowledge of programming. However, students are expected to have successfully completed SDS 328M Biostatistics before taking this class, and materials from SDS 328M will be considered known. In particular, students are expected to have some basic familiarity with the statistical language R.

Textbook

There is no textbook for this class. All reading assignments will be documents that are freely available online. Students will also be expected to find relevant materials using Google as well as online help forums such as stackoverflow.com.

Computing requirements

Computational biology needs to be learned by doing, and much of the classroom time will be dedicated to working through simple problems. Therefore, students will be strongly encouraged to bring their own laptops into the classroom and to follow along as the material is presented. While no graded assignments in this class will require having a laptop, the overall learning experience will be much less rewarding for students who cannot participate in in-class activities. Both R and python will be available through a web-based system, so the only system requirement for student laptops is a modern web browser.

Course site

All materials and assignments will be posted on the course webpage at:
http://wilkelab.org/classes/SDS348_spring_2020.html

Assignments will be submitted and grades will be posted on Canvas at:
https://utexas.instructure.com

R and python compute sessions are available at:
https://educcomp01.ccbb.utexas.edu/
https://educcomp02.ccbb.utexas.edu/
https://educcomp03.ccbb.utexas.edu/
https://educcomp04.ccbb.utexas.edu/
(Please choose one of them arbitrarily. E.g., roll a die and pick the compute server corresponding to the number you roll. If you roll 5 or 6, roll again.)

Assignments and grading

This class will have 11 graded homeworks and 3 graded projects. Every Monday of each week, either a homework or a project will be due. Both homeworks and projects need to be submitted as pdf files on Canvas. Homeworks are worth 10 points and projects are worth 100 points. The lowest-scoring homework will be dropped, so that a maximum of 100 points can be obtained from the homeworks. Finally, you earn 4 points for each attended Wednesday lab section, up to 52 points total. (There are 15 lab sections, so you can miss 2 and still receive full points.)

Thus, in summary, each project contributes 22% to the final grade, the totality of all homeworks contributes another 22% to the final grade, and lab attendance contributes 12%. **There are no traditional exams in this class and there is no final.**

Assignment type	Number	Points per assignment	Total points
Homework	10 (+1)	10	100
Project	3	100	300
Lab attendance	13 (+2)	4	52
		Total	452

The class will use +/- grading, and the exact grade boundaries will be determined at the end of the semester. However, the following minimum grades will be guaranteed:

Total points achieved	Minimum guaranteed grade
407 (90%)	A-
362 (80%)	B-
316 (70%)	C-
226 (50%)	D-

Late assignments

Solutions to homeworks will be discussed on the Wednesday of the week in which they are due. Late homework submissions will receive a mandatory penalty of 3 points. Homeworks that are received after 9am on Wednesday will not be graded and will receive 0 points.

Project submissions similarly have a 2-day grace period. Projects submitted during the grace period will have 25 points deducted from the obtained grade. After the grace period, students who haven not submitted their project will receive 0 points.

Extra assignment for graduate course credit

Graduate students who are taking this class for graduate course credit (i.e., are enrolled in SDS385) will have to complete one additional assignment. The assignment will be to write a brief report (4-5 pages, no more than 3 figures) applying concepts from this class to a dataset of the student's choice. This assignment will be graded pass/fail, and a failing grade on this assignment will result in a 45 point penalty on the total points obtained in the class. **This assignment will be due on April 14, 2020.** Students who receive a failing grade can submit a revised assignment for regrading. The last day by which a revised assignment can be submitted is the last day of class in the semester.

Academic dishonesty

This course is built upon the idea that student interaction is important and a powerful way to learn. We encourage you to study together often. However, there are times when you need to demonstrate your own ability to work and solve problems. In particular, your homeworks and projects are independent work, unless explicitly stated otherwise. You are allowed to confer with fellow students about general approaches to solve the problems in the assignments, but you have to do the assignments on your own and describe your work in your own words. Students who violate these expectations can expect to receive a failing grade on the assignment and will be reported to Student Judicial Services. These types of violations are reported to professional schools, should you ever decide to apply one day. Don't do it—it's not worth the consequences.

Special accommodations

Students with disabilities. Students with disabilities may request appropriate accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 512-471-6259, http://www.utexas.edu/diversity/ddce/ssd/

Religious holy days. Students who must miss a class, a homework assignment, or a project in order to observe a religious holy day will be given an opportunity to complete the missed work within a reasonable time after the absence. According to UT Austin policy, such students must notify me of the pending absence at least fourteen days prior to the date of observance of a religious holy day.

Office hours

Both the graduate TA and myself are available during posted office hours or at other times by appointment. Do not hesitate to request an appointment if you cannot make it to the posted office hours. The most effective way to request an appointment for office hours is to suggest several times that work for you. I would suggest to write an email such as the following:

> Dear Dr. Wilke,
>
> I would like to request a meeting with you outside of regular office hours this week. I am available Thurs. between 2pm and 3:30pm or Fri. before 11am or after 4pm.
>
> Thanks a lot,
> John Doe

Note that I will usually not make appointments before 10am or after 6pm.

Email policy

When emailing about this course, please put "SDS348" into the subject line. Emails to me or the TA should be restricted to organizational issues, such as requests for appointments, questions about course organization, etc. For all other issues, please see us in person.

Specifically, we will not discuss technical issues related to assignments over email. Technical issues are questions concerning how to approach a particular problem, whether a particular solution is correct, or how to use the statistical software R. It is acceptable to inquire per email if you suspect that a problem set has a typo or if you find the wording of a problem set ambiguous.

We will also not discuss grades or grading issues per email. According to state law and UT regulations, all grading information must be kept confidential, and email is not a confidential communication medium. If you have concerns about your grade, talk to the TA or me in the office hours. It is OK to send an email inquiring about grading issues that affect all students. For example, the question "Do I understand correctly that question 2 is worth 3 points" would be fine. However, please do not send an email that states your grade, and please do not expect us to send you an email that states your grade either.

Schedule, SDS 348, Spring 2020

Class	Date	Topic	
1	1/21/2020	Introduction	
		Part I: Advanced data analysis and visualization with R	
2	1/23/2020	R review, R markdown	
3	1/28/2020	Data visualization with ggplot2	**HW 1 due 1/27**
4	1/30/2020	Data visualization with ggplot2	
5	2/4/2020	Working with tidy data	**HW 2 due 2/3**
6	2/6/2020	Working with tidy data	
7	2/11/2020	Working with tidy data	**HW 3 due 2/10**
8	2/13/2020	Rearranging data tables with tidyr	
9	2/18/2020	Principal Components Analysis (PCA)	**HW 4 due 2/17**
10	2/20/2020	k-means clustering	
11	2/25/2020	Binary prediction/logistic regression	**Project I due 2/24**
12	2/27/2020	Sensitivity/Specificity, ROC curves	
13	3/3/2020	Training and test data sets, cross-validation	**HW 5 due 3/2**
		Part II: Scripting with Python	
14	3/5/2020	Installing and running python, basic data structures	
15	3/10/2020	Control flow (if/for) in python	**HW 6 due 3/9**
16	3/12/2020	Functions in python	
	3/16–3/20	**Spring break**	
17	3/24/2020	More on python data structures, classes	**HW 7 due 3/23**
18	3/26/2020	Working with files	
19	3/31/2020	Introduction to Biopython	**Project II due 3/30**
20	4/2/2020	Working with gene features and genomes	
21	4/7/2020	Running queries on Entrez	**HW 8 due 4/6**
22	4/9/2020	Regular expressions	
23	4/14/2020	Using regular expressions to analyze data	**SDS385 report due, HW 9 due 4/13**
24	4/16/2020	Using regular expressions to analyze data	
		Part III: Misc. topics	
25	4/21/2020	Aligning sequences	**HW 10 due 4/20**
26	4/23/2020	Global and local alignments, BLAST	
27	4/28/2020	Multiple sequence alignments and phylogenetic trees	**HW 11 due 4/27**
28	4/30/2020	Working with protein structures	
29	5/5/2020	Geospatial data (maps)	
30	5/7/2020	Animations	**Project III due 5/7**