

SDS 348/385—Computational Biology and Bioinformatics

Spring 2018

Unique 56800: TTH 9:30–11am UTC 3.112, W 9am–10am FAC 101B

Unique 56805: TTH 9:30–11am UTC 3.112, W 10am–11am FAC 101B

Unique 56865: 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: Tues. 11:30am–1:30pm, or by appointment

Twitter: @ClausWilke

Teaching Assistant: Dariya Sydykova

Email: dariya.k.sydykova@gmail.com

Office: MBB 3.232

Office Hours: Mon. 2pm–4pm, or by appointment

Purpose and contents of the class

Over the last decade, advances in high-throughput measurement techniques have transformed biology into a data-driven science. It is now routine to measure the abundances of thousands of RNAs or proteins in a cell, carry out many hundreds of experiments using robotic liquid-handling tools, or sequence multiple genomes in just a single experiment. All these high-throughput techniques produce massive amounts of data, and the biologist of the 21st century frequently spends substantially more time and effort analyzing these data than generating them in the first place.

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 using their own computer.

Course site

All materials and assignments will be posted on the course webpage at:

http://wilkelab.org/classes/SDS348_spring_2018.html

Assignments will be submitted and grades will be posted on Canvas at:

<https://utexas.instructure.com>

Assignments and grading

This class will have 11 graded homeworks and 3 graded projects. Every Tuesday 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 immediately after the due date. **Late homework submissions will not be graded and will receive 0 points.**

Project submissions have a 2-day grace period. Projects submitted during the grace period will have 20 points deducted from the obtained grade. After the grace period, students who haven't submitted their project will receive 0 points.

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.

*For project III, all students except those taking the class for graduate credit are allowed to work in pairs.

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 10am and 11:30am or Fri. before 11am or after 5pm.

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 2018

Class	Date	Topic	
1	1/16/2017	Introduction	
Part I: Advanced data analysis and visualization with R			
2	1/18/2017	R review, R markdown	
3	1/23/2017	Data visualization with ggplot2	HW 1 due
4	1/25/2017	Data visualization with ggplot2	
5	1/30/2017	Working with tidy data	HW 2 due
6	2/1/2017	Working with tidy data	
7	2/6/2017	Working with tidy data	HW 3 due
8	2/8/2017	Rearranging data tables with tidyr	
9	2/13/2017	Principal Components Analysis (PCA)	HW 4 due
10	2/15/2017	k-means clustering	
11	2/20/2017	Binary prediction/logistic regression	Project I due
12	2/22/2017	Sensitivity/Specificity, ROC curves	
13	2/27/2017	Training and test data sets, cross-validation	HW 5 due
Part II: Scripting with Python			
14	3/1/2017	Installing and running python, basic data structures	
15	3/6/2017	Control flow (if/for) in python	HW 6 due
16	3/8/2017	Functions in python	
3/12-3/17 Spring break			
17	3/20/2017	More on python data structures, classes	HW 7 due
18	3/22/2017	Working with files	
19	3/27/2017	Introduction to Biopython	Project II due
20	3/29/2017	Working with gene features and genomes	
21	4/3/2017	Running queries on Entrez	HW 8 due
22	4/5/2017	Regular expressions	
23	4/10/2017	Using regular expressions to analyze data	HW 9 due
24	4/12/2017	Using regular expressions to analyze data	
Part III: Misc. topics			
25	4/17/2017	Aligning sequences	HW 10 due
26	4/19/2017	Global and local alignments, BLAST	
27	4/24/2017	Multiple sequence alignments and phylogenetic trees	HW 11 due
28	4/26/2017	Working with protein structures	
29	5/1/2017	Protein structure and PyMOL scripting	
30	5/3/2017	Molecular evolution of proteins	Project III due