SDS 348—Computational Biology and Bioinformatics

Spring 2015
Unique 57440: TTH 12:30-2pm, W 9am-10am
Unique 57445: TTH 12:30-2pm, W 10am-11am
Unique 57517: TTH 12:30-2pm, W 10am-11am (SDS 385, with graduate course credit)

Instructor: Claus O. Wilke
Email: wilke@austin.utexas.edu
Office: MBB 3.232
Office Hours: Mon. 3pm – 5pm
Twitter: @ClausWilke

Teaching Assistant: Stephanie J. Spielman
Email: stephanie.spielman@utexas.edu
Office: MBB 3.232
Office Hours: Mon. 1pm – 3pm in MBB 3.204

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 comparing gene-expression data among conditions, 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_2015.html
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 handed in in hard-copy during class. 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. Thus, each project contributes 25% to the final grade, and the totality of all homeworks contributes another 25% to the final grade. **There are no traditional exams in this class and there is no final.**

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:

Percentage of total points	Minimum guaranteed grade
90%	A-
80%	B-
70%	C-
50%	D-

Late assignments

Solutions to homeworks will be discussed on the Wednesday after the due date. Homeworks handed in late but before that discussion section will receive a 10% penalty. Once homework solutions have been discussed students who have not yet handed in their homeworks will receive 0 points. For projects, there is a 1 week grace period during which a 10% penalty will be applied. After the grace period, students who haven not 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 2015

Class	Date	Topic		
1	1/20/2015	Introduction		
		Part I: Advanced data analysis and visualization with R		
2	1/22/2015	R review, R markdown		
3	1/27/2015	Data visualization with ggplot2		**HW 1 due**
4	1/29/2015	Data visualization with ggplot2		
5	2/3/2015	Tidy data		**HW 2 due**
6	2/5/2015	Fundamentals of data analysis: one-table verbs		
7	2/10/2015	Fundamentals of data analysis: two-table verbs		**HW 3 due**
8	2/12/2015	Analysis of grouped data		
9	2/17/2015	Logistic regression		**HW 4 due**
10	2/19/2015	Classification, ROC curves		
11	2/24/2015	Differential gene-expression analysis	**Project I due**	
12	2/26/2015	Clustering		
13	3/3/2015	Principal Components Analysis		**HW 5 due**
		Part II: Scripting with Python		
14	3/5/2015	Python: installing and running Python		
15	3/10/2015	variables, assignments, **if**, **for**		**HW 6 due**
16	3/12/2015	functions		
	3/16-3/20	**Spring break**		
17	3/24/2015	Numerical methods: numerical integration		**HW 7 due**
18	3/26/2015	Numerical methods: solving differential equations		
19	3/31/2015	Biopython: working with sequence data	**Project II due**	
20	4/2/2015	Biopython: parsing Genbank data		
21	4/7/2015	Biopython: parsing Genbank data		**HW 8 due**
22	4/9/2015	regular expressions		
23	4/14/2015	regular expressions		**HW 9 due**
		Part III: Misc. topics		
24	4/16/2015	Sequence analysis: finding sequences (BLAST)		
25	4/21/2015	Sequence analysis: alignment and phylogenetic tree building		**HW 10 due**
26	4/23/2015	Working with protein structures		
27	4/28/2015	Protein design		**HW 11 due**
28	4/30/2015	GWAS		
29	5/5/2015	Epidemiology		
30	5/7/2015	Guest lecture	**Project III due**	