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

Instructor: Claus O. Wilke
Email: wilke@austin.utexas.edu
Office: MBB 3.232
Twitter: @ClausWilke

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.

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