Statistics 5570 / 6570

Statistical Bioinformatics

Spring 2014

Key Points of Syllabus

- This is a "topics" class and may seem structured quite differently from other STAT classes you've taken.
- Notes posted on class website
- Homework posted on class website, submitted through Canvas
- Classroom behavior will exhibit mutual respect – feel free to ask questions and participate (you may be called on); no off-task wireless usage
- Grades not inflated or curved, and based on homework, 6000-level tutorial project, take-home exam
- Grading only discussed in office hours or via email
- We will use R; help available in office hours and (to some degree) via email
- Students are trusted to read the rest of the syllabus

Class Meetings: LIB 302, MF 11:30-12:45
Course Website: `www.stat.usu.edu/jrstevens/stat5570`

Instructor: John Stevens
Contact: 797-2818, `john.r.stevens@usu.edu`
Office Hours: ANSC 224, MW 3:00-4:30, or by appointment; but on Faculty Senate days (1/6, 2/3, 3/3, & 4/7) office hours are rescheduled for 1:00-1:50

Prerequisites: STAT 5100 or 5200 or equivalent required. Experience with "programming" in some package such as Matlab, Maple, or R will be advantageous, but is not essential.

Reference Text: Because the field of Statistical Bioinformatics changes so quickly, it is difficult to have a single, stable reference. However, a good starting point to see many of the recurring statistical themes in the field is *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (2005), Gentleman et al. eds. We will cover material from almost every chapter, and the text includes a good deal of very helpful R code. However, the text is not required, and you may not be able to return this text to the USU bookstore. We will also make use of published papers and online resources.

Important Dates: Jan. 20 is a holiday (no class). Feb. 18 (Tuesday) Monday schedule; no class on Feb. 17. Mar. 10-14 is spring break (no class).

Course Description: This course will provide an introduction to some of the major **statistical** issues in bioinformatics, a fast-growing, interdisciplinary field of study. At the end of this course students will be able to use the Bioconductor family of packages in R for a variety of bioinformatics-related analyses, with emphasis on gene expression analysis (such as microarray and next-gen sequencing), but also sequence analysis, metabolomics, and mass spectrometry. In addition, students will be able to discuss the limitations and

statistical issues of such analyses.

Student Responsibilities: Students are expected to attend class, participate in class discussions, and complete assignments on time. Students will be responsible for all material presented in class as well as material covered by the assigned homework problems. All students should feel free to ask questions and give answers in a positive, non-threatening classroom environment. Students are also encouraged to take advantage of office hours to receive additional instruction as necessary. Due to the cumulative nature of the course materials, students should take care to not get far behind nor allow concerns to go unaddressed. Students are responsible for monitoring their progress in the course and noting the dates on the syllabus, including the drop deadlines. Students deciding to drop the course are alone responsible for knowing the relevant deadlines and for taking the necessary actions. Students are expected to read the notes and other materials distributed by the instructor.

Office Hours: While questions and discussions are welcome and expected during class periods, all students should feel free to visit office hours for individual assistance with the course material. Questions regarding grades or scores will **only** be answered during office hours. Students unable to attend office hours may easily make an appointment to see the instructor at another time. Students may not access their email on the instructor's office computer.

Grading: Your grade will be based on various assignments. Students registered for 6570 credit have an additional major project to complete. The anticipated grade cut-offs are as follows:

%	Grade	%	Grade
94	A	77	C+
90	A-	73	C
87	B+	70	C-
83	B	65	D+
80	B-	60	D

There is no fixed grade profile for this class. If every student does well, then every student can get an A. Note that scores and grades will not be inflated and are only an indication of what you do, not who you are.

Homework: Some assignments will be worth more than others. Students should submit their completed assignment as a single document (.pdf, .doc, or .docx preferred) through Canvas before 11:59 P.M. on the due date. To access Canvas, go to `http://canvas.usu.edu` and use your BANNER username ("A" number) and password to log on. It is the student's responsibility to know when homework is due. Most of the homework will involve some computer work, mostly in R. Homework assignments should be typed neatly with necessary computer output and graphics placed in order with each corresponding homework exercise. Figures and sub-figures (including fonts) should be clear and readable, and square in shape with no more than four sub-figures within a 3.5 inch height. Each homework should include an appendix with relevant R code. Any unnecessary computer output will result in points deducted. Late homework will receive a 20% deduction for

each calendar day received after the due date. While for most assignments you are encouraged to discuss your work with others, homework handed in must be the student's own work. There will be at least one large assignment that will be more like a take-home exam, with no cooperation allowed amongst students. Another assignment may require an oral presentation to the class.

Starting Homework: For each Homework assignment, think first what needs to be done (conceptually and then statistically), and then construct (maybe using available sample code) the appropriate R code to perform the analysis. If you just blindly modify available sample R code, you will waste a lot of time (and not learn as much). This will require you to first understand what the available sample R code does, so you can find which part(s) of it you will need to modify and use (or not).

6570 Project: Those students who register for this course as STAT 6570 have a major project in addition to the regular assignments required of all students. Students will choose a statistical bioinformatics topic approved by the instructor and not otherwise featured in the course, and make a two-component tutorial presentation of the topic. The first component is a 40-minute presentation to introduce the topic to the class. The second component is a written report submitted to the instructor. Possible topics include certain skipped chapters from the text or other approved research topics. Additional details will be given in class.

Extra Credit: There will be no extra credit.

Handouts and Course Notes: Classroom activities will require students to use handouts and incomplete notes which they will need to print off from the course website and bring to class. The instructor will not bring extra handouts or notes to class. Students are responsible for checking the course website regularly for new notes.

Note: Students with ADA-documented physical, sensory, emotional or medical impairments may be eligible for reasonable accommodations. Veterans may also be eligible for services. All accommodations are coordinated through the Disability Resource Center (DRC) in Room 101 of the University Inn, (435)797-2444. Please contact the DRC as early in the semester as possible. Alternate format materials (Braille, large print, digital, or audio) are available with advance notice.

Computer Work: Because R has emerged as a tremendous platform for bioinformatics-related analyses and because it is freely available, this course will involve substantial use of the R language. Students will be expected to install R in such a way that they can use it on their own – if not on a home computer or laptop, then on a CD or flash drive for use in student computer labs. An introduction to the main syntax of relevant R commands will be given in class, and students should expect to learn more about the features of R through completion of the assigned homework. While the instructor will be willing to provide additional instruction on the use of R, students should recognize that the purpose of the homework is to develop their own proficiency in using the Bioconductor (and related) tools, and not necessarily to arrive at a biological conclusion. Accordingly, students should expect to spend a fair amount of time on the homework.

Nature of Topics Course: This is essentially a "Topics Course" and the pace and content of the course will be somewhat fluid. The following gives a rough outline of the expected order in which major topics will be addressed.

Tentative Order of Units and Course Activities

1. Gene Expression Analysis - Technology and Preprocessing
    - Weeks: 1, 2
    - References: Ch. 1, 2, 4; papers
    - Homework Topics: microarrays, Bioconductor, and preprocessing data
2. Gene Expression Analysis - Visualization
    - Weeks: 3, 4
    - References: Ch. 3, 10, 13; papers
    - Homework Topics: quality checks, clustering, visualization
3. Gene Expression Analysis - Inference (Identifying Candidate Genes)
    - Weeks: 5, 6
    - References: Ch. 14, 15, 23; papers
    - Homework Topics: tests for differential expression, filtering, multiple testing
4. Gene Expression Analysis - Annotation (Characterizing Candidate Genes)
    - Weeks: 7, 8
    - References: Ch. 7-9, 18-22; papers
    - Homework Topics: use of annotation information, pathway visualization
5. Mass Spectrometry Analysis - Technology and Preprocessing
    - Weeks: 9
    - References: Ch. 6; papers
    - Homework Topics: preprocessing and visualizing mass spectrometry data
6. Sequence Analysis and Next-Generation Sequencing Technology
    - Weeks: 10, 11, 12
    - References: papers
    - Homework Topics: pairwise & multiple alignments, hidden Markov models, technology, available tools
7. Discussion Topics
    - Weeks: 13, 14, 15
    - References: papers
    - Homework Topics: ongoing developments in the field