Instructor: Kerby Shedden, 461 W. Hall, kshedden@umich.edu

Course webpage: www.stat.lsa.umich.edu/~kshedden/Courses/Stat545

Overview: This is a course in statistical analysis of experimental data from biological experiments, with an emphasis on high throughput assays and multivariate analysis techniques relevant in modern molecular biology. During this course we will look at a number of popular analysis strategies, consider them in the context of their scientific aims, and critically analyze their performance from both statistical and scientific viewpoints.

Format: Formal lectures will generally occupy around 50-75% of each class session, with the remainder devoted to group discussions and data analysis exercises.

Topics: See the course webpage for a tentative schedule of topics to be covered.

Prerequisites: Students should have a working mathematical background equivalent to single-variable calculus, and a statistics background including at minimum a familiarity with basic concepts in probability, hypothesis testing, and point estimation. Students with little or no background in biology should be prepared to do additional reading early in the course.

Course work: Students are expected to attend class regularly and participate in class discussions. In particular, all students should attend the group presentations and discussions on December 1, 6, 8, and 13. The coursework will have four components: two problem sets due approximately in early October and early November; a two-page written critical paper review due December 1; and a group research presentation to be given on either December 8th or 13th. More details about these components follow.

- The problem sets will consist of computer data analysis exercises based on the lecture material.
- The critical paper review will be a two-page critical summary of a published research article, focusing on the appropriateness and completeness of the data analysis strategy. Each student will be assigned one of around four papers to evaluate and will be responsible for submitting an individual evaluation of the paper. Then we will have a round table discussion of the papers in class on December 1st and 6th.
- Students will be randomly assigned to project groups of around three students. Each project group will formulate a research aim relevant to the course topics, prepare a brief strategy for meeting the aim (as in a grant proposal), and will carry out preliminary studies using publicly-available data. The group as a whole will submit a 2-3 page summary and give a 10-15 minute presentation on its work. The presentations will take place on December 8th and 13th.
**Grading:** Course grades will be determined by an equally weighted average of four individual grades for the two problem sets, the critical paper review, and the group analysis project.

**Computing:** Three of the four graded components of the course will require students to program a computer. In the two problem sets, each student must demonstrate his or her ability to do this individually (although conceptual discussion with other students is encouraged). Programming in the relevant languages will be taught from first principles, but at a rapid pace. Students with little or no relevant programming background should be prepared to invest significant additional time in this area.

Students are encouraged to use R, Matlab, or Python as appropriate for course programming work. One important aspect of this course will be the development and implementation of data analysis procedures. Therefore we will avoid using pre-built packages of analysis routines such as the Bioconductor package. Code examples given in class will be written in either R or Matlab. The course will also cover the programming language Python which is useful for data file cleaning, text processing, and elementary chemical structure calculations.