# BINF702 BIOLOGICAL DATA ANALYSIS - SPRING 2016

**Instructor** - Jeff Solka, 540-809-9799 (Cell), jlsolka@gmail.com

**Office Hours** - By Appointment

**Schedule** - Mondays 4:30 p.m. - 7:10 p.m. in Prince William Campus, Ocaquan Rm. 204B

**Texts** - Gareth James (Author), Daniela Witten (Author), Trevor Hastie (Author), Robert Tibshirani (Author), An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) Hardcover – August 12, 2013

Peter Dalgaard, Introductory Statistics With R (required)

Wim P. Krijnen, Applied Statistics for Bioinformatics using R, freely under the GNU Free document License

**Grading** - Grades will be based on 4-5 problem sets and an independent final project with an associated 8-10 page paper and 15 minute class presentation. All assessments will be open book and notes. Each of these will contribute to your grade as follows.

Problem Sets (80%), Final Project and Paper (10%) and Final Project and Paper Presentation (10%)

Students will be allowed to work in teams of 2 on their projects.

Weekly homework assignments will be provided but will not be graded.
Solutions to the homework assignments will be provided each week.
Grading will be on the following scale. 97-100 (A+), 93-96 (A), 90-92 (A-), 87-89 (B+), 80-86 (B), less that 80 C, Student averages will be rounded to the closest integer to determine final letter grades.

## Guidence on the Course Project

**Project Proposal**: Student teams must prepare a brief proposal, 2-4 pages, describing the independent project and must submit this proposal no later than March 14, 2016. The proposal should be divided into four sections:

1. Background and objectives: A description of the background of the biological system and the question(s) that you hope to answer. In many cases this might involve reinvestigating a dataset that was already covered in the literature by other authors, i.e. the Golub data.

2. Computational methods: The computational methods that you intend to use to answer the question(s) in your proposal.

3. Discussion: A brief description of how you plan to evaluate the biological significance of the results of your computer analysis. It's very important in science to motivate your audience to care about your work with

its “Impact” or “Significance”.

4. Several references describing the background of your proposed project.

The proposal will not be graded, because its sole purpose is to determine whether the objectives of the project are reasonable and interesting.

Please note that the final project should be designed to test a biological hypothesis. .

**Final Report**: The final report should be in the form of a scientific paper, divided into the following sections: (1) Abstract, (2) Background and objectives, (3) Computational methods, (4) Results and discussion, (5) Conclusions, (6) A brief description of how the conclusions of your analyses could be tested using biochemical or genetic techniques, (7) References.

**References**: Please follow the Cell Journal guidelines for references EXACTLY. I highly recommend that you use a referencing and bibliography software package like EndNote, Zotero, bibtex etc. (It will make your life much easier!) References in the text should include the authors’ names and dates:

- One author: (Pearson, 1996)
- Two authors: (Smith and Waterman, 1981)
- Three or more authors: (Altschul et al., 1990)
- Multiple references: (Pearson, 1996; Smith and Waterman, 1981; Altschul et al., 1990)

The references in the bibliography should also adhere to the Cell Journal format:

- Journal article: Lipman, D.J., Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. Science 227, 1435-1441.
- Book chapter: Schuler G.D. (1998). Sequence alignment and database searching. In: Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, AD Baxevani and BFF Ouellette, eds. Wiley Interscience, New York, NY.

**Organization**: Please try to organize the information and the interpretations as clearly as possible. It is unreasonable to expect the reader to hunt through large numbers of pages to find data supporting a specific conclusion. There are two acceptable ways of organizing the figures. First, the data and text can be integrated into the body of the paper. Second, the data can be compiled into a series of clearly-labeled appendices.

**Figures**: Every figure should have a caption adequately describing the contents of the figure without having to resort to reading the main text. There must be **at least 5 figures** created by the student, and **at least 4 of them must be created in R**.

**Length**: The final report should be 10-12 pages double-spaced, not including computer output or references.

**Presentation**: The last lecture session will be devoted to oral presentations of the projects.

**Course Content**

## Week 1 1/25/16

Biological Data Handout

R Syntax and Development Handout

Golub Handout

Golub Paper

Week 1 HW

Krijnen Chapter 1 R Script

Krijnen Chpt1 Solutions

## Week 2 2/1/16

Data Display and Descriptive Statistics

Microarrays

Descriptive Statistics

Krijnen Chapter 2 R Script

Krijnen Chapter 2 HW Script

## Week 3 2/8/16

Probability

## Week 4 2/15/16

Statistical Distributions

## Week 5 2/22/16

Estimation and Inference

## Week 6 2/29/16

Linear Models

**Spring Break 3/7/16 -- 3/13/16**

## Week 7 3/14/16

Micro Array Analysis

## Week 8 3/21/16

Cluster Analysis

## Week 9 3/28/16

Classification Methods

## Week 10 4/4/16

Resampling Methods and the Bootstrap

## Week 11 4/11/16

Markov Models and KnitR

## Week 12 4/18/16

Experimental Design and Shiny

## Week 13 4/25/16

The Implications of Big Data

## Week 14 5/2/16 Student Presentations