BIOS 60576 (01) Topics in Bioinformatics: Introduction to R

Steven Buechler, Department of Mathematics

MW 3:00 – 4:15, August 28 to October 17, 2007
(1.5 credits)

Recommended text: **Statistics: An Introduction using R.** Michael Crawley, J. Wiley (2005). (about $33 through Amazon).

Required manual: http://cran.r-project.org/doc/manuals/R-intro.pdf

This seven-week course module is an introduction to the data analysis and statistical modeling language R. A list of topics to be covered is given below. R is an open-source computer application designed for organizing and visualizing experimental data and performing statistical analysis of this data. Numerous add-on packages exist for applications ranging from microarray data analysis to ecological modeling. Many of these packages are part of the Bioconductor system. See http://www.r-project.org/ and http://www.bioconductor.org/.

While I will not assume any familiarity with R, I will assume a basic working knowledge of statistics, such as would be learned in an undergraduate biostatistics course. Homework assignments using R will be given. Students are strongly encouraged to bring a laptop to class so that they can practice executing commands as they are introduced.

Topics

1. Working with data in R
    a. R language fundamentals
        i. Command format
        ii. Data types and object classes
        iii. Input and output of data
        iv. Getting help
    b. Data manipulation
        i. Array and matrix operations
        ii. Data frames (sorting, extracting subsets, summary data, …)
        iii. Frequency tables
    c. Visualizing data
    d. Writing functions, control structures and scripts
    e. Using packages, like Bioconductor
2. Univariate statistics

    a. Probability distributions
    b. Density estimation and tests for normality
3. Linear statistical models
    a. What is a statistical model?
    b. Model formulas
    c. Extracting information from the “fit” object
    d. Finding an optimal set of explanatory variables
4. Microarray data analysis
    a. Preprocessing and quality assessment
    b. Differential expression methods
        i. The multiple testing problem and “solutions”
        ii. Bioconductor packages for differential expression analysis
    c. Connecting array data with biological information
        i. Using Affymetrix array annotation packages with NCBI data.
        ii. Using Gene Ontology and KEGG pathway information to find functionally relevant genes.