

Northeastern University, Fall 2006
STATISTICS FOR BIOINFORMATICS (NEU: MTHG 340, BU:XRG NU500 A1),
DRAFT OF THE SYLLABUS

The aim of this four-credit one-semester course is to introduce students to statistical methods used in practice and applications of Bioinformatics. Some of the material is to become a working instrument in further work, while the rest is introduced for orientation in the multidisciplinary field of Bioinformatics. Basic relevant concepts from Probability, Stochastic processes, Information Theory, Statistics and Experimental Design will be introduced and illustrated by examples from Molecular Biology, Genomics and Population Genetics with an outline of algorithms and software.

Textbooks : Although “Statistical Methods in Bioinformatics. An Introduction” by Warren J. Ewens and Gregory R. Grant, Springer, N.Y., second edition, 2004, will NOT be used in the main part of the course, students might find its advanced part (including the intro to BLAST) of interest in their subsequent work. This book must be available from the NEU bookstore. This seems to be the first specialized textbook for a two-semester course taught for 6 last years in the University of Pennsylvania. To proceed more fast for covering diverse topics without Probability as a prerequisite in ONE semester, covering certain topics such as Multivariate joint distributions, need to be sacrificed.

For the introduction, you can download “Statistical Problems in Genetics and Molecular Biology” from www.math.neu.edu/~Malioutov/DrinkwaterLectures.pdf

using a generous permission of its authors N. Drinkwater and C. Denniston from the University of Wisconsin, which has an advantage of introducing the basic concepts of Probability and Statistics on examples from Genetics and Molecular Biology.

When appropriate, I shall distribute extracts from other sources, namely:

Deonier RC, Tavaré S & Waterman MS. [*Computational Genome Analysis: An Introduction*](#). Springer Verlag, New York. 2005,

“Biological Sequence Analysis” by E. Durbin et al, when introducing the Hidden Markov techniques.

For an introduction to the Population Genetics I shall use recent lectures by W. Ewens which you can download from:

www.math.neu.edu/~Malioutov/EwensCornellLect.pdf

Another free modern intro to Population Genetics by S. Tavaré can be downloaded from:

<http://www.cmb.usc.edu/people/stavare/stpapers-pdf/frbook.pdf>

A good additional reading in Probability might be: “Intro. to Probability”, second edition, by C. M. Grinstead and J.L. Snell, AMS, 1998, which is available (together with many java applets illustrating the material and games related to chance) from

http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html

Two labs based on the package SPSS (or EXCEL, or any other package with random number generation facility) and two labs on sequence alignment based on MatLab are part of the homework. Two additional labs in MATLAB are on the Hidden Markov modeling of sequences alignment, and one more lab is on Estimating distances between proteins based on Markov model. Projects based on examples from Bioinformatics covered in the textbook will be 10% grade each.

Time and Place : TBA.

Instructor : Prof. Mike Malyutov, Mathematics Dept., NEU, Phone: 373 5650, office: 545 Lake Hall, e-mail: MLTV@neu.edu,

Office hours TBA after discussion with students or by appointment via phone or e-mail.

Grade: 20% weekly homework, 20% four labs, 20% two best tests or projects, 40% take home final test.

