Math/Stat 547: Probabilistic Modeling in Bioinformatics.

Winter Term '04, MWF 9-10. (Room: 1060 East Hall, but this may be changed.)

Biological Sequence Analysis

The course will be about probabilistic models of proteins and nucleic acids, and their uses in molecular biology. The core of the course will be the analysis of sequences and its biological applications such as the searching of large databases for optimal comparisons or homologies of sequences (DNA nucleotide sequences or amino acid sequences for proteins), location of gene loci on a string of DNA, estimation of phylogenetic trees, structural motif recognition and structure prediction for proteins and RNA. Guest lecturers will address the class on applications in the pharmaceutical industry, as well as some earlier examples of these techniques applied to problems in linguistics and speech recognition.

More specifically, the topics will include a review of basic concepts of probability and very rudimentary molecular biology; probability and the design of similarity scoring functions; optimal local and global alignments of sequences: dynamic programming, Smith-Waterman algorithm, other algorithms available on the Web (BLAST and FastA, etc.), probabilistic (heuristic) versus rigorous algorithms; significance of scores and simulation; dependence of scoring functions and optimal alignments on parameters, comparison of standard tables; hidden markov models and neural network models; entropy and information content of a sequence; multiple sequence alignment methods and algorithms. The applications will be to gene finding; families of proteins; phylogenetic tree determinations; structure of proteins and recognizable patterns in amino acid sequences (motif recognition); DNA mechanics and duplex destabilization; *de novo* peptide sequencing from mass spectrometry data.

There will be no exam for the course. Students will be expected to complete five to six problem sets, most of which will hopefully be group projects. If the class demographics work out favorably, we will be mixing students with biological background and mathematical/statistical background in each group. The prerequisites are left very flexible for the moment, but students will be expected to try to familiarize themselves with possible gap areas in their background as the term progresses. Every effort will be made to accommodate students from biological or other relevant backgrounds. Students enrolling in this course are encouraged to enroll in Math/Stats 548: Computations in Probabilistic Modeling in Bioinformatics [1 credit].

Reference:

R. Durbin, S. Eddy, A. Krogh and G. Mitchison, "Biological Sequence Analysis", Cambridge, C.U.P., 1998.

Further papers, books and websites will be used during the term. Check the course webpage for more details:

`http : //www.math.lsa.umich.edu/˜dburns/547/547syll.html`

or contact the instructor: Dan Burns, 763-0152, `dburns@umich.edu`.