BIOL 7023. Bioinformatics. Fall 2007

Dr. Mark Borodovsky, Whittaker Bldg., Room 4201.
Office hours: MF 12noon-1pm.
Office: 404-894-8432
Email: borodovsky@gatech.edu (preferable method)
Teaching Assistants:
Andrey Kislyuk, Wenhan Zhu
Email: kislyuk@gatech.edu, wemhan@amber.gatech.edu

Prerequisites: Modeling and Dynamics (MATH 6705); Computing Concepts for Bioinformatics (CS 4710); Introduction to Probability and Statistics (MATH 3215)

Recommended texts:
Durbin R., Eddy S., Krogh A., Mitcheson G. Biological sequence analysis. Cambridge University Press, 1998 (2006 reprinting).
Borodovsky M., Ekisheva S. Problems and solutions in biological sequence analysis, Cambridge University Press, 2006.

Recommended site:
Wikipedia - http://en.wikipedia.org/wiki/Bioinformatics

Bioinformatics is the field of science growing from the application of mathematics, statistics and information technology to the study and analysis of the very large biological and particularly genetic data sets.
This course is devoted to mathematical models and computer algorithms of DNA and protein sequence analysis. Students will practice on software programming of the algorithms studied in the course (in simplified settings) as well as get experience in using sequence analysis tools available either locally or via Internet.
Some lab time might be used for tests, student's presentations and additional lectures.

Grading rules:	Home works	15%	
	Tests	40%	(20% each)
	Lab projects	20%	
	Final exam	25%	

Homework: Small group efforts are encouraged.
Policy on examinations: Open books, open notes.

Important dates:

Test 1	September 26
Last day W	October 12
Test 2	November 7
Last day of classes	December 7
Final exam	December 11, 2:50pm

No lectures on campus: September 3, October 8, November 23

Course outline:
Introduction. The natural science paradigm. Molecular biology as a study of the information processing in the cell. Genes and Genetic Code. Sequencing genomes. The

interdisciplinary field of Bioinformatics. Major public data resources: US: Entrez (NCBI); European Union: EMBL-EBI (EBI) and SwissProt (SIB); Japan: DDBJ (NIG).

Probabilistic models of DNA sequences. Multinomial models. Models for protein-coding and non-coding regions. Bayesian inference. Estimation of model parameters. Supervised and unsupervised classification & clusterization of DNA sequences. Machine learning approach.

Developing sequence analysis tools. Strings, graphs and algorithms. Deterministic (string based) models and algorithms for string matching.

Pairwise alignment of biomolecular sequences. Search for similarities. Global alignment of two sequences. Needleman-Wunsch algorithm. Local alignment. Smith-Waterman algorithm.

Markov models. Homogeneous and inhomogeneous Markov models. Interpolated Markov models. Hidden Markov models. Pattern recognition with (Hidden) Markov models. Example: CpG islands in genomic DNA.

Probabilistic models of sequences conserved in evolution. Multiple alignment of conservative sequence domains. Gibbs sampling algorithm for multiple sequence alignment. Algorithms for prediction of functional sites in DNA sequences (RBS sites, promoters, splice sites).

Algorithms for gene identification in genomic DNA. Three-periodic Markov models. Hidden Markov models for prokaryotic and eukaryotic genes. Viterbi algorithm, Forward and Backward algorithms, posterior decoding algorithm.

Search for similarities in biomolecular sequences. Necessity of scoring system. Dot-matrix method. Statistical distributions of similar words. Common words in two random sequences. Distribution of the maximum length of a common word in two random sequences.

Markov models of DNA sequence evolution. Jukes-Cantor’s and Kimura’s models. Rate of mutation matrices and matrices of transition probabilities. Amino acid classification. Markov model of protein sequence evolution. Dayhoff's approach. Estimation of parameters of mutation matrices using alignments of closely related sequences.

Derivation of scoring functions for amino acid substitutions observed in pairwise alignments. The notion of relative entropy. Dayhoff's series of scoring matrices (PAM matrices). Derivation of scoring functions for amino acid substitutions observed in BLOCKS database. Series of BLOSUM matrices.

Evolutionary conserved regions in protein sequences. Multiple sequence alignment. Multidimensional dynamic programming. Progressive alignment methods.

Statistical models of protein domains. Analogy between models for functional sites in DNA and models of functional & structural motifs in protein sequences. The concept of profile. Profile HMM: Hidden Markov model for evolutionary conserved sequences.

Estimation of parameters for profile HMM. Predicting protein function by profile HMM. PSI-BLAST, PFAM and SMART local similarity search methods. Finding remote homologs. Assessment of the power of a homology search method by using SCOP database.

Protein secondary structure prediction. Profile-based neural network approach (PHD method). Information theory & Bayesian approach (GOR method). Prediction of the 3D structure of proteins.
Building phylogenetic trees. Construction of the tree by using pairwise distances. UPGMA clustering and neighbors joining clustering algorithm. Notion of parsimony.

CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.

Probabilistic approaches to phylogeny. Random genetic drift. Molecular clock. Synonymous and non-synonymous substitutions. Using maximum likelihood approach for phylogenetic inference. Orthologs and paralogs. Evolutionary and comparative genomics.

Study of gene expression with DNA microarrays. Detecting patterns in expression of multiple genes. Clustering methods of gene expression data:
k-means clustering, self-organizing maps.

Lab schedule:
Lab 1-2. UNIX environment. Biological data resources on Internet. Major data formats. Downloading sequence data from GenBank and SwissProt. Compiling test sets of sequences for using in following lab sessions.

Lab 3-4. Perl language review. Implementation of Needleman-Wunsch algorithm for global pairwise sequence alignment.

Lab 5. DNA sequence statistical model building and sequence generation.

Lab 6. Using the GeneMark and GeneMark.hmm gene prediction program for prokaryotic and eukaryotic genomic DNA.

Labs 7-8. Viterbi algorithm for DNA sequence segmentation. Implementation in Perl.

Lab 9. Fast sequence similarity searches. Perl scripts for parsing BLAST outputs.

Labs 10-11. Multiple sequence alignment by Gibbs sampling and simulated annealing. Implementation in Perl.

Labs 12-13. Confirmation of gene prediction and identifying protein function by PSI-BLAST or by profile HMM. Identification genes and proteins using EST database.

Labs 14-15 Building phylogenetic trees for sets of protein sequences retrieved from Entrez. Using CLUSTALW. Identification of laterally transferred genes.