CSCI 6595 -- Advanced Topics in Bioinformatics

Instructor:

Asst. Prof. Stephen Winters-Hilt
Phone: (504) 896-2761; (504) 280-2407
E-mail: winters@cs.uno.edu

Time/Location:

Lecture Location: CERM 217; Lecture Hours: F 11:00-1:45
Campus Office Location: CERM 217B; Office Hours: F 10:00-11:00
Institute Office Location: R&E 3215; Office Hours: Th 10:00-12:00

Prerequisites: CSCI 2125, or permission of instructor; either CSCI 4567 or CSCI 4569 (equivalent to any of the 4990/6990 coursework taught under Dr. Winters-Hilt), or permission of instructor.

Textbooks (required):

- --- TBA
- --- instructor notes

Reference Books (optional):

(1) Biological Sequence Analysis by R. Durbin et al.
(2) Programming Perl by Larry Wall et al.

Abstract & Course Objectives:

This *Advanced Topics in Bioinformatics* course will be an advanced graduate course that builds on the programming-intensive applications of research from CSCI 4589/6589 involving Hidden Markov Models and from CSCI 4590/6590 involving Support Vector Machines. The class projects will typically build on project efforts begun in prerequisite courses, and often result in peer-reviewed journal publications.

General Machine Learning & Bioinformatics Project Objectives:

Real-world deployment. Students should be familiar with training and testing in a real computational environment (including simple distributed computational arrangements on a networked cluster of computers to the extent that time permits).
Performance optimization. Students should understand how to obtain statistically valid (objective) scores of performance and how to use that information for performance optimization.
Peer-reviewed Publication. Some students are expected to have projects sufficiently mature that they will be asked, for their Final Project, to communicate their results as a paper submission.

Grading: (A) 90-100; (B) 75-89; (C) 65-74; (D) 55-64; (F) below 55.

Homework assignments	10%
Midterm	10%
Final/Project	80%

Policies:

- Omit documentation in your code at your own risk
- Academic honesty and fairness is expected at all times

Topics Covered:

I. Channel Current Cheminformatics
  - Time-domain Finite State Automata (FSAs) for signal acquisition (single-pass "scan" mode)
  - tFSAs for "spike" feature extraction
  - ad hoc preprocessing methods in other applications (ab initio gene finding)

II. Advanced Hidden Markov Model (HMM) Projects
  - Segmented-stationary signal analysis and kinetic feature extraction
  - Genome Structure Identification (generalized HMMs, with duration, possible SVM polarization)
  - Transcription Factor Binding Site (TFBS) Identification (using gIMM, SVM polarization)
  - Transcriptome Structure Identification (HMM feature extraction, SVM clustering)

III. Advanced Support Vector Machines (SVMs) Projects
  - SVM heuristic variants (alpha-selection, etc.)
  - Kernel Variants, Theory, and Mercer testing
  - SVM Variants for Multiclass Classification: Decomposition and Monolithic
  - SVM-based Clustering: Internal and External Formulations

IV. Boosting
  - Elimination of "hands-on" tuning of multiclass classifiers by boosting classifier ensembles
  - SVM boosting to eliminate tuning and/or explicit kernel selection