Bioinformatics I

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

Time/Location:
Lecture Location: Math 229; Lecture Hours: MW 1:00-2:15
Office Location: CERM 217; Office Hours: MW 10:30-12:00

Prerequisites: CSCI 2125, MATH 2314, or permission of instructor.

Textbooks (required):

Biological Sequence Analysis by R. Durbin *et al*. Cambridge University Press (1999). ISBN 0-521-62971-3.

The Elements of Statistical Learning by Trevor Hastie et al. Springer-Verlag (2001). ISBN 0-387-95284-5.

Reference Books (optional):

Programming Perl (3rd ed.) by Larry Wall *et al*. O'Reilly Media (2000). ISBN 0-596-00027-8.

Background:

Bioinformatics I is an introduction to informatics, real-world applications in the health and medical sciences, pattern recognition, feature extraction, statistical analysis, and kinetics analysis, and provides students with an attractive skill set to companies seeking to hire informatics specialists. These informatics courses also directly tie methods to applications, and, thus, are also meant to significantly enhance the skill-sets of non-computer-science students, including: biology students (*via* bioinformatics); pre-med students; medical students; pharmacology, toxicology, and forensics students (*via* biomedical informatics); and chemistry, biochemistry, and chemical engineering students (*via* cheminformatics).

This course involves introductory material similar to that given in CSCI 4568, but turns its focus to computational genomics applications, while CSCI 4568 focuses on channel current cheminformatics and related interdisciplinary Biomedical engineering and Biophysics issues. Either this course, or CSCI 4568, serve as prerequisites to the introductory and advanced CSCI projects courses: 4587, 4588, 4595; 6587, 6588, 6595. This course will include program writing for data mining, as well as analysis of stochastic sequential data, such as genomic or proteomic data. There will be a large project component to the course with a wide selection of problems, from programming intensive informatics solutions to explorations of theoretical & algorithmic methods.

Course Abstract:

An introduction to the algorithms and theory used in bioinformatics and cheminformatics, with current applications in computational genomics and biomedical informatics Covers statistical methods for identifying motifs in DNA and protein sequences and identifying structural features in stochastic sequential data (for gene finding and for feature extraction from protein-channel ionic current measurements), Bayesian statistical methods (e.g. Markov Chain Monte Carlo) for sequence motif discovery and discriminative methods for use in informatics, particularly Support Vector Machine approaches.

Course Objectives:

Task decomposition.

Students should understand how to decompose a complex informatics task into a collection of standard informatics tasks: feature identification and knowledge discovery, signal acquisition and filtering, feature extraction, classification, and data-rejection.

Method selection.

Students should understand how to analyze the general properties of their data and factor in their computational limitations in order to select the most efficient informatics method at each stage of the task decomposition.

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.

Grading:

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

Homework assignments	50%
Midterm	20%
Final Project	30%

Policies:

- Most of the assignments can done with others
- Final Projects must be done individually
- Homework is due in class on due date specified
- Omit documentation in your code at your own risk

Topics Covered:

The basic genomic code (week 1)

- Introduce the genome-wide sequence data format, background and literature review

Bioinformatics: Gene Finding (week 1,2)

- Cover the tough gene prediction problem, challenges and elaborate representative gene prediction methods such as: "gene finder".

Perl (week 3-4)

- Four computer lab sessions cover basic aspects of Perl language: syntax, motif and loops, regular expression and bioperl modules

Pairwise Alignment (week 5)

- Cover the fundamental dynamic programming algorithm for pairwise alignment

Markov Chains (week 6)

- Cover plain Markov chain models for sequence analysis both first-order and higher order, and introduce sample applications, such as CpG island prediction.

Hidden Markov Models (HMMs) (week 7,8)

- Cover theoretic aspects of hidden Markov Models. Applications include: revisit the CpG island prediction problem and predicting topology of transmembrane protein from primary amino acid sequences, such as TMHMM.

Midterm

Generalized HMM implementations: IHMM, GIHMM (week 9, 10)

- Introduce IHMM and GIHMM and include one sample application. For example protein family classification
- Kernel based discrimination methods

Support Vector Machines (week 11)

- Theory of SVM, includes frequently used kernel models, feature selection and unbiased access of classification error. Sample application
- Tree based discrimination methods: CART, Random Forest (week 12)
- Theory of CART and Random Forest. Sample application.

Final projects presentations (week 13-15)