

## CSc 212 Course Description

**Academic Unit:** Department of Computer Science

**Department Chair:** Dr. Du Zhang

**Title:** Bioinformatics: Data Integration and Algorithms

**Units:** 3

**Justification:**

Biology has rapidly become a data-rich, information-hungry science because of recent massive data generation technologies. Our biological colleagues are designing more clever and informative experiments because of recent advances in molecular science. These experiments and data hold the key to the deepest secrets of biology and medicine, but we cannot fully analyze this data due to the wealth and complexity of the information available. Bioinformatics is the application of computational tools and techniques to the management and analysis of biological data. The primary task of Bioinformatics is to discover knowledge from biological data.

There are three levels of computational tools involved in Bioinformatics. They are (1) public tools and database web services provided by well established centers such as NCBI; (2) commercial or open source software packages for solving one type of problem; (3) creating an integrated information environment to address a particular set of problems in a research project or application. If a tool from (1) and (2) alone can not meet your needs, you have to combine several tools, make your own tools, or extend the utility of existing tools, finally integrating data generated from all the tools involved to provide a multi-dimensional view with cross inferences. In addition to becoming competent users of biological software tools at levels (1) and (2) (BIO 224 provides such education), an introductory training in level (3) is essential and in growing demand. This course is designed to address the education needs of students who will be tool builders for the field of Bioinformatics and is based on over two years of development with the Biological Sciences department at CSUS.

**Course Description (updated 2/17/05):**

The application of information technology and computer science to biological problems, in particular to biomedical science issues involving genetic sequences. Algorithms and their applications to DNA sequencing and protein database search; tools and techniques for data integration to transform genetic sequencing data into comprehensible information to study biological processes.

**Prerequisite (updated 2/17/05):** CSC 130, STAT 50, and graduate standing. Bio 010 recommended.

**Description of Expected Learning Outcomes:**

Demonstrate understanding of basic biological, statistical and algorithmic concepts, techniques and models underlying bioinformatics tools. Upon completion of this course the student will:

- Gain insight into the field of Bioinformatics from theoretical models to finished software.
- Understand how software design and methods can be integrated with existing tools to create productive information environment for bioinformatics practice.
- Understand how open source can be powerful in creating web-based applications in Bioinformatics.
- Understand important roles of programming languages and databases in Bioinformatics software development and service.
- Understand common string algorithms and their application in biological database search.
- Understand statistical importance of sequence alignment and database search.
- Understand Hidden Markov Models theory and applications in protein, DNA and RNA.
- Understand common data mining methods used in Bioinformatics and their implementation.
- Make decisions about which data mining methods to use in different knowledge discovery in biological data.
- Understand the mapping from sequence to protein structure, in particular, homology modeling and protein structure prediction methods.
- Understand basic research methods and current challenges in the field of Bioinformatics.

**Assessment Strategies:** Students will be evaluated on their performance in completing lab exercises, participation in discussions, presentations of term project, reading assignments. Both midterm and final exams will be given that will cover the material presented in lecture and assigned reading.

**Textbooks and readings:**

1. Cynthia Gibas, Per Jambeck. “Developing Bioinformatics Computer Skills”. O’Reilly & Associates; 2001
2. James Tisdall, “Beginning Perl for Bioinformatics”, O’Reilly & Associates; 2001
3. Dan Gusfield, “Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology”, Cambridge University Press, 1997
4. Pierre Baldi and Soren Brunak, “Bioinformatics: the Machine Learning Approach”, The MIT Press, 1999

5. David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Laboratory Press, 2001.

**Major topics covered in the course:**

1. Data integration: in addition to sequence data we input data from other dimensions that also support data viewing and relationship inference.
2. Open source in Bioinformatics: Linux, Apache, MySQL, PHP/Perl/ Bioperl, EMBOSS open source sequence analysis tools, the GeneX gene expression microarray database, etc.
3. XML: data representation and transformation for Bioinformatics
4. Integration of database and knowledge base for knowledge discovery in biological databases
5. String algorithms and their applications to DNA sequencing and protein database search
6. Dynamic programming, BLAST, and FASTA: algorithm comparison and implementation, parallel computing
7. Data mining methods for Bioinformatics
8. Hidden Markov Models: theory and application in Bioinformatics
9. Case studies: IBM Blue Gene Project, a Web-based Microarray Experiment Management System, Protein Secondary Structure Prediction

**Activities:**

The course will be taught using both lectures and lab sessions, and will include three components in addition to 2 hours lecture and 2 hours lab each week:

1. Web site construction and oral presentation:  
Each student will construct a web site progressively through the semester. Most lab exercises and the term project assignment will be posted on his/her web site for information sharing and functionality demonstration. Lab exercises will allow student to implement new tools or integrate existing tools to solve a data analysis or management problem. Each student will give an oral presentation of his term project at the end of the course.
2. Reading assignments:  
With selected case studies or fundamental concepts presented during lecture, a reading assignment will be given to allow students to develop skills in engineering and scientific problem solving by learning from examples.
3. Exercises and discussion:  
For each unit either an exercise or a discussion topic will be assigned to give hands-on experience with current topic covered in the lecture.