What is Bioinformatics? Bioinformatics is the application of computational techniques to the discovery of knowledge from biological databases.

"Bioinformatics is the marriage of molecular biology with computer science. It's an exciting, relatively young and emerging science, in which computer information technology greatly helps to integrate, manage, analyze, and visualize genetic and other biological information."

Sophisticated laboratory technology allow the collection of biological data faster than Biologists can interpret it. Vast volumes of DNA sequence data is growing at an exponential rate.

Why should we care?
The potential of bioinformatics for the good of society is enormous. Among other things, scientists apply this technology to

- Study biological processes in organisms
- Determine how these processes go wrong in diseases
- Discover and develop drugs to treat, cure and prevent human diseases
- Trace the evolutionary tree based on DNA sequences

CS Department Web-site: (http://www.cs.uni.edu/overview_bioinformatics.php)

Bioinformatics is an interdisciplinary curriculum combining computer science, mathematics, and biology that seeks to explore and elucidate life processes through modern genomic techniques and tools.

The [Bioinformatics BS] program seeks to prepare students to understand and manipulate vast genomic data bases through an understanding of the languages, tools, and techniques of computer science, mathematics, and biology. The program seeks to be truely interdisciplinary in nature and produce students who are conversant across all three disciplines.

Consider...

"... Career opportunities in bioinformatics are very, very good," said John M. Greene, senior staff scientist, bioinformatics research, at Gene Logic Inc., Gaithersburg, Maryland. "It seems that every time you turn around a company has decided to set up a bioinformatics group, or expand an existing group. Many scientists are turning their careers in this direction..."
-- Science Magazine: Focus on Careers: Bioinformatics Science

"...Experts have already dubbed bioinformatics - a hybrid profession pairing biology and computer science - the career choice of the decade. "
-- Industry Outlook: Biotechnology

"There is a crying need for experts in bioinformatics and this is not something that will just fade away," said Dr. Leena Peltonen, chairwoman of the Department of Human Genetics at UCLA. " -- Industry Outlook: Biotechnology

"...Frost & Sullivan, a San Jose consulting firm, has predicted a 10 percent annual growth rate in the bioinformatics market while the National Science Foundation has estimated 20,000 jobs will be created in the field by 2005."
-- Industry Outlook: Biotechnology

"Opportunities abound in bioinformatics, but qualified candidates are hard to find"

"The field of bioinformatics is surrounded with so much hype that thinking it's a brand-new field is forgivable. The word itself was coined in the early 1990s, but people have been using databases to manage biological sequence data -- which is a part of what bioinformatics encompasses -- since the 1960s. Although the field may not be new, it's moving in new directions and creating plenty of job opportunities." -- Chemical & Engineering News

Biological Data

DNA (deoxyribonucleic acid) sequence - contains 1-dimensional storage of genetic information

- self-replicating genetic material of inheritance
- encodes information used to create proteins
- a very long string consisting of a four-letter alphabet (ACGT)
  A (adenine), T (thymine), C (cytosine) and G (guanine)

RNA (ribonucleic acid) - an intermediary used in transferring a small part of DNA's information for the construction of a protein

- RNA consists of a long string consisting of a four-letter alphabet (ACGU)
  A (adenine), U (uracil), C (cytosine) and G (guanine)

Proteins - perform most functions in living organisms (their 3-D structure aids in the role of chemical reactions)

- Protien functions:
  - ✓ tissue building blocks, called structure proteins
  - ✓ biological catalysts for chemical reactions in cells, called enzymes
  - ✓ oxygen transport
  - ✓ antibody defense
- Proteins are chains of amino acid residues. There are 20 different amino acids.
- However, a protein's 3-D structure aids in the role of chemical reactions

Question: How many bases of an RNA sequence are needed to encode one amino acid of a protein? (Hint: RNA's alphabet is of size 4 and protein's alphabet is of size 20)

Bioinformatics Questions to be Answered:

- How do we figure out which parts of the DNA control the various chemical processes of life?
- We know the function and structure of some proteins, but how do we determine the function of other proteins?
- How do we predict what a protein will look like 3-dimensionally, based on knowledge of its sequence?
- How do we find new meaningful DNA subsequences for new proteins in a DNA sequence and add them to the DNA-protein dictionary?

Course Description: "Intermediate programming with emphasis on bioinformatics. Includes file handling, memory management, multi-threading, B-trees, introduction to dynamic programming including Wunsch-Neddleman and Smith-Waterman algorithms for optimal alignments, exploration of BLAST, FASTA and gapped alignment, substitution matrices."

Comments about the course:

- Being a new field it needs new computational tools to manage, mine, and analyze ever growing amount of genomic data
- Efficient algorithms are needed to process large amounts of data
- Our text does a great job of teaching algorithm development using bioinformatics examples, but we'll need to supplyment it with discussions of file handling and B-trees
- Textbook is written for senior/graduate level so problems at the end of the chapters are very hard
- We'll split the text approximately in halve between:

<u>Chapters: 1- 6 will be covered in Computing for Bioinformatics I</u>

Chapter 1: Introduction

Chapter 2: Algorithms and Complexity - some of this is written for non-CS majors

Chapter 3: Molecular Biology Primer - most of this is written of non-Biology majors

Chapter 4: Exhaustive Search - enumerate through all possible solutions looking for the best

Chapter 5: Greedy Algorithms - make decisions based on the what appears to be the best choice at the current time and live with the consequences (probably the most nature problem solving technique, but it can also get you in trouble)

Chapter 6: Dynamic Programming Algorithms - growing the answer to a problem by first solving smaller problems, remembering the answer to the smaller problems, and using the answer to the smaller problems to extend the answer to the larger problem

Chapters: 7-11 will be covered in Computing for Bioinformatics II

Chapter 7: Divide-and-Conquer Algorithms - split the problem into smaller problems, solve the smaller problems, and use the answer to the smaller problems to solve the original larger problem

Chapter 8: Graph Algorithms - uses many problem solving techniques

Chapter 9: Combinatorial Pattern Matching - hash tables, suffix trees, heuristics

Chapter 10: Clustering and Trees - uses many problem solving techniques

Chapter 11: Hidden Markov Models - popular machine learning approach in bioinformatics. Machine learning algorithms are presented with training data so the algorithm can derive important insights about often hidden/unknown parameters. After training, the algorithm uses the information learned in the training phase and applied it to target data of interest to answer the question.

Chapter 12: Randomized Algorithms - make random decisions throughout their operation. Useful to provide approximate answers for hard problems were no known optimal algorithm runs in polynomial-time.