

**Course Number:** CSC 857

**Course Title:** Bioinformatics Computing

**Number of Credits:** 3

**Schedule:** Three hours of lecture/discussion per week.

## Introduction

Emerging at the intersection of Biology, Chemistry, Physics, Mathematics, and Computer Science, Bioinformatics is a discipline that will play a critical role in the coming years. This influence can already be seen in terms of increasing industrial job opportunities, upsurge in academic research and funding, and most importantly better understanding of the mechanism of life.

This course will give students broad and holistic background in Bioinformatics that will encompass basic technologies and algorithms, biological data storage and querying, software development issues, and advanced applications. We will focus on fundamentals of bioinformatics without assuming any prior knowledge of the area. The course will help evolve student understanding and culminate by considering many emerging R&D issues that are being addressed or need to be addressed both in industrial and academic settings. Opportunities to explore specific areas of bioinformatics analysis and processing in detail will be provided through course projects. It is expected that students who complete the course will have understanding of all basic methods and concepts in Bioinformatics at high to medium level.

## List of Topics

We expect to cover the following topics in this class.

**Fundamentals of Bioinformatics:** The molecular basis of life, molecular representations, structure of nucleotides, polynucleotide chains, structure of the DNA molecule, central dogma of molecular biology.

**Sequence Alignment and Comparison :** problem formulation, percent sequence identity, quality of alignments, introducing gaps, the optimization formulation, dynamic programming, Needleman-Wunsch algorithm, semi-global alignments, local alignments and the Smith-Wartermen algorithm, linear-space alignment and the Hirschberg algorithm.

**Proteins, Protein Sequences and Alignments:** protein biosynthesis, amino acids and their structures, the peptide bond and polypeptides, secondary structure, Ramachandran plot, amino acid coding, codon tables, degeneracy of the genetic code.

**BLAST:** high-throughput sequence comparison, Karlin-Altschul equation and its application, variations of the BLAST algorithm.

**Multiple Sequence Alignment:** problem formulation and complexity, relationship with pair-wise alignment, scoring schemes for MSA, heuristics for solving the MSA problem; progressive alignment, star alignment, the CLUSTALW algorithm.

**Substitution Patterns and Substitution Matrices:** mutations and substitutions, mutation rates, functional constraints on gene substitution, types of substitutions, estimating substitution numbers: Markov processes, the Jukes-Cantor model and its derivation, transitions and transversions, Kimura's model, more advanced models, bias-variance tradeoffs, substitution matrices: BLOSUM and PAM.

**Phylogenetics: Distance-Based Methods:** Introduction to molecular phylogenetics, phylogenetic trees, character and distance data, distance matrix methods, UPGMA, estimating branch lengths, anisotropic evolution, transformed-distance methods, neighbor joining methods.

**Phylogenetics: Character-Based Methods:** Principle of parsimony, small and large parsimony, the Fitch algorithm, maximum likelihood estimation of phylogeny, bootstrapping.

**Gene Finding:** techniques for detection of coding regions, detecting promoter regions, alternate splicing, Markov sequence models.

**Motif Discovery:** problem formulation and formalization, profiles and consensus scores, the median string problem and relation to motif discovery, partial consensus, branch-and-bound search techniques for motif finding.

**Microarray Expression Analysis:** principles of microarray design, microarray data extraction: gridding and spot detection, vector-space representation, measures of similarity, basic analysis techniques: agglomerative clustering, neighbor joining, average linkage clustering, k-means clustering, self-organising maps, correspondence analysis.

**Structural Bioinformatics:** secondary structure prediction: Chou-Fasman algorithm, lattice models, homology modeling,

**Introduction to Modern Pharmaceutical Drug Discovery:** key stages in the drug discovery pipeline and their inter-relationships, clinical trials.

**Advanced Topics:** aligning and matching molecules, whole genome alignments, modeling ADME-PK, finding and modeling active sites, other topics (variable)

**Textbooks:** To be discussed in the class

**Grading:** to be discussed in the class

Updated: September 2011 Prof. R. Singh