Bioinformatics

BIOL 4815-001 BIOL 5815-001 CSCE 4815-001 CSCE 5815-001 MATH 4815-001 MATH 5700-002

Course Information (Spring 2026)

Instructor: Rajeev Azad (Rajeev.Azad@unt.edu)

Lectures: Thursday 6:00 – 8:50 PM

Semester Credit Hours: 3

Location: Cury 211

Office hours: Friday, 8:15-9:30 AM via Zoom: https://unt.zoom.us/j/2298672063, 1:30 - 2:15 PM at LSC B314, or by appointment

Recommended Textbooks: The course content will be based on selected Bioinformatics textbooks considering diverse background of the prospective students and with no expectation of the prior exposure to this interdisciplinary area. The recommended textbooks are Biological Sequence Analysis by Durbin *et al.*, Bioinformatics and Functional Genomics by Pevsner, and Statistical Methods in Bioinformatics by Ewens & Grant. New research developments will also be covered in this course, based mainly on research articles and review papers.

Course objective: The aim of this course is to familiarize students with state-of-the-art methodologies in Bioinformatics and Computational Biology, and help them understand how to apply these techniques to solving biological and biomedical problems. This course will include the following topics:

- **An introduction to Bioinformatics**
  History of Bioinformatics; key milestones; genome projects status and statistics; perspectives on Bioinformatics- the cell, organism, and tree of life

- **Biological Databases and Genome Browsers**
  Major public DNA databases; NCBI key features: GenBank, PubMed, BLAST, OMIM, Gene, Protein, HomoloGene, Taxonomy, Structure; accession numbers; RefSeq; Genome Browsers- Ensembl and UCSC browsers; Galaxy

- **An introduction to probability and matrix operations**

- **Probabilistic model of protein evolution**
  Odds ratio; log odds ratio; substitution matrix; PAM matrices- derivation of score parameters; BLOSUM matrices- derivation of score parameters

- **Pairwise sequence alignment**

Global alignment algorithm; local alignment algorithm; heuristic algorithms; dynamic programming algorithms to find optimal alignment; dot plots; alignment with affine gap scores

- **Basic local alignment search tool (BLAST)**
  BLAST components; BLAST types; BLAST algorithm; understanding search parameters; interpreting search output- score, e-value, identity, positives, gaps, coverage; BLAST search strategies; PSI-BLAST- position-specific scoring matrix (PSSM) and algorithm; RPS-BLAST

- **Probabilistic models for interpreting biological sequence data**
  Markov chain- technicalities/properties, Markov chain models- transition probability matrix, homogeneous and inhomogeneous Markov models of DNA sequences; model training; stationary distributions; applications, interpolated Markov model; hidden Markov model- elements of HMM, standard and generalized HMM, Forward-Backward algorithm, Viterbi algorithm, Expectation-Maximization algorithm; pair hidden Markov model; profile hidden Markov model

- **Genome architecture and annotation**
  Gene prediction- GeneMark, GLIMMER, EasyGene, protein topology prediction- TMHMM and Phobius; protein family / domain prediction- HMMER / pfam

- **Multiple sequence alignment**
  Multiple sequence alignment (MSA): properties, features, and uses; scoring MSA- Minimum entropy, Sum of pairs (SP) scores, multidimensional dynamic programming; MSA methods: Feng-Doolittle progressive MSA, profile alignment; iterative refinement methods-Barton-Sternberg multiple alignment algorithm; MSA programs: ClustalW, ProbCons

- **Molecular phylogeny and evolution**
  Historical background; molecular clock hypothesis; nomenclature of trees- taxon, OTU, clade; rooting a tree; species trees versus gene/protein tree, phylogenetic tree construction-alignment, models of substitution- hamming distance, Jukes and Cantor one-parameter model, Kimura two-parameter model; tree-building methods- distance, character, and model based, UPGMA, neighbor-joining, maximum parsimony, maximum likelihood, and Bayesian; evaluating trees: bootstrapping

- **Genome evolution: vertical and horizontal modes of gene transfer**
  Horizontal gene transfer (HGT) as a potent force in prokaryotic genome evolution; HGT in eukaryotes; genomic islands; methods for HGT quantification

Course outcomes:

1. Appreciation of the interdisciplinary approaches to solving problems in biology

2. Understanding of the essence of computational and mathematical methods in biology and medicine

3. Familiarization with principles and models underlying standard bioinformatics methods/algorithms

4. Practical experience of using bioinformatics tools for biological data analysis.

5. Awareness of biological databases and their usage using available tools

6. Development of analytical skills to probe molecular sequences to understand biological processes at cellular and organismal levels.

Grading: Based on class participation (15%: 10% for attendance, 5% for discussions), in-class presentation (10%), homework assignments (35%), project (25%), and exam (15%). Above 90% ≡ A, 80-89.99% ≡ B, 70-79.99% ≡ C, 60-69.99% ≡ D, and Below 60% ≡ F

Attendance Policy: Attendance is essential and thus is expected.

Catalog description: Introduction to the interdisciplinary field of Bioinformatics. Databases and genome browser tools. Methods and algorithms for biological sequence analysis. Applications to problems in biology or medicine.