ALGORITHMS FOR BIOLOGICAL SEQUENCE ANALYSIS

Computer Science Methods in Computational Biology

Description: This course aims at providing a unifying overview of sequence analysis methods of use in modern computational molecular biology. The main topics covered include pairwise alignment, hidden Markov models, multiple alignment, profile searches, RNA secondary structure analysis, and phylogenic inference. The course is intended for computer scientists wishing to pursue a research career in computational molecular biology.

Objectives: The central objective of the course is to connect the two rapidly convergent fields of computer science and modern molecular biology. As such, the course is conceived as a starting point for a specialty track for graduate or advanced undergraduate computer science students in algorithms that currently have, or seem to have in the future, a great potential application in molecular biology.

Specific Objectives:

1 To motivate computer science students to consider a variety of biological significant problems which are defined primarily on computational terms such as strings, trees, and grammars.

2 To provide computer science students with a knowledge and expertise in the design and use of classical computer science methods in bioinformatics such as string comparisons, database searches, determination of patterns and secondary structures.

3 To empower and foster the use of computer science skills such as correctness of proof, algorithm analysis, bounded approximation results, randomized algorithms, among others, in the search for effective solutions for old and new molecular biology problems.

4 To introduce and discuss in complete detail deterministic string methods and probabilistic models and methods of use in biological sequence analysis. As a result, students are expected to master a wide spectrum computing algorithms together with their underlying principles, ideas and derivations.

5 To develop in students an appreciation for the interplay between computer science and evolutionary aspects of molecular biology methods, emphasizing information processing and communication, and complexity.

Units: Three credits

Prerequisites: An undergraduate course in Data Structures.

Grading: Computing laboratories and homework exercises will account for about 75% of the final grade. The final project will be about 25% of the final grade.

Instructional Strategies:

The course material will be mainly presented in lectures. Students will be assigned computing laboratories and homework exercises to complement the lectures and provide an active learning experience. Early in the course, each student will select a project. This project will enhance the scope of the course and provide an important additional active learning experience. Each student will present each project results at the end of the course.

Topics to be covered in the course:

1. Introduction: evolution as an information processing phenomenon. Selection and mutation as operations on a space of genotypes.
2. Probabilities and probabilistic models.
3. Pairwise alignment algorithms: scoring model, complexity of brute force approach, dynamic programming methods, heuristic methods.
4. Significance of scores. Derivation of score parameters from alignment data.
5. Markov chains. Hidden Markov models (HMM). Parameter estimation for HMMs. HMM model structure. Numerical stability of HMM algorithms.
6. Pairwise alignment algorithms using HMMs.
7. Pair HMM versus Finite State Automata methods.
8. Profile HMM methods for sequence of families
9. Multiple sequence alignment methods.
10. Methods for building phylogenetic trees. Probabilistic approaches to phylogeny.
11. Transformational grammars and parsing of biological sequences.
12. Methods for RNA structure analysis.