Computational Linguistics of Biomolecules

CSCI 8610 Special Topics in Theoretical Computer Science
(Spring 2014)

Dr. Liming Cai, Computer Science Department, UGA
Email: cai@cs.uga.edu

Introduction

A biological molecule (DNA, RNA, or protein) is a sequence of linearly chained bio-residues (nucleic acids or amino acids). Neighboring and distant residues on a molecule sequence may physically interact to form tertiary (3D) structures that are critical to its function. While intriguing biological rules governing the formation of such structures are yet to be fully revealed, computational modeling and analysis of biomolecular structures have been made possible. Computational linguistics has proved a viable mean for such tasks.

Biomolecules are by nature formal languages whose sentences are linearly organized bio-objects. Structural patterns incurred by residues on a biomolecule sequence can be considered the co-occurrences of tokens or terminal symbols as on a language sentence. Structural patterns potentially admitted by the biomolecule are interpreted as the possible syntactic structures dictated by grammar rules that define the formation of such sentences. Hence, stochastic grammars, where rules are associated with probability distributions, may faithfully model biomolecules. Indeed, research in developing such models along with efficient parsing algorithms has delivered feasible solutions to various problems in analysis and prediction of bimolecular structures.

Content

This course will expose the students to a number of topics in formal languages and grammars, probability computation, and optimization algorithms in the context of modeling and analysis of biomolecules, spanning the following areas:

1. Chomsky grammars (regular and context-free); non-Chomsky (mildly context-sensitive grammars);
2. Stochastic grammar systems (Hidden Markov model, stochastic context-free grammar, stochastic mildly context-sensitive grammar);

3. Dynamic programming algorithms for probability computation (sequence parsing and probability parameter estimations); and

4. Biomolecule modeling and analysis (profiling and detection of DNA genes, secondary and tertiary structure prediction of RNAs and proteins).

Format

The teaching will be a mix of lectures by the instructor and presentations by students on their literature-readings and research projects. No textbook will be used. Grading will be based on project reports, presentations, and participation in classroom discussions.

Prerequisites

CSCI 2670 (Theory of Computation), or CSCI 4470/6470 (Algorithms), or CSCI 4490/6490 (Algorithms for Computational Biology), or the approval of the instructor.

No prior knowledge, but an interest, in molecular biology is essential for this course.