Course proposal: CMSC701 - Computational Genomics

September 2010

Bioinformatics Field Committee

Motivation and Background

This course builds upon the success of two Computational Biology Courses offered over the past 5 years in the department: Algorithms for Biological Sequence Analysis (offered as CMSC858E - 2005,2006 , CMSC858P - 2008, CMSC858W - 2010), and Computational Gene Finding and Genome Assembly (offered as CMSC828N - 2006,2007,2008, CMSC828H - 2010). These courses were well attended and well received by the students, including by students with research interests outside of computational biology (e.g. theory, programming languages, language processing/linguistics, and software engineering).

The new course will combine the topics covered in the original courses and will primarily focus on string algorithms with relevance to biological research: exact/inexact matching, sequence reconstruction, pattern recognition (gene finding). The course is intended to complement CMSC702 - a course that covers the computational analysis of biological systems rather than sequences.

CMSC701 will be one of the few graduate courses in the department that cover string algorithms in detail, thus, we expect the course to attract a broad range of students in addition to those with a primary interest in computational biology.

Course description

An introduction to the algorithms and heuristics used in the analysis of biological sequences. Includes an introduction to string matching and alignment algorithms, phylogenetic analysis, string reconstruction (genome assembly), and sequence pattern recognition (gene and motif finding). A particular emphasis will be placed on the design of efficient algorithms and on techniques for analyzing the time and space complexity of these algorithms. Computational concepts will be presented in the context of current biological applications. No knowledge of biology necessary.

Prerequisites

CMSC 423, or Computer Science or Applied Mathematics graduate student, or permission of instructor.

Textbooks

There is currently no textbook that covers all the topics presented in this course. The following textbooks contain a number of topics that are discussed in this course and they will be listed as recommended reading. The relevant material from these books will be made available to the students as handouts.

Dan Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology.* Cambridge University Press. 1st Edition (1997) ISBN 0-521-58519-8.

Richard Durbin, Sean Eddy, Anders Krogh, Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press. 1st Edition (1999) ISBN 0-521-62971-3

William Majoros. *Computational Gene Prediction.* Cambridge University Press. 1st Edition (2007). ISBN 0-521-87751-0

Tentative Syllabus

- Exact string matching basics (1 week)
  - KMP, Boyer Moore, Aho Corasick
- Suffix trees and related topics (2 weeks)
  - Suffix trees construction, use in alignment
  - Suffix arrays
  - Burrows Wheeler transform for compression and searching
- Inexact alignment (3 weeks)
  - Basic dynamic programming approach: local, global alignment, affine gaps
  - Linear-space alignment
  - Landau-Vishkin algorithm
  - Seeding - spaced seeds, inexact seeds
  - Special applications (whole-genome alignment, short read alignment, spliced alignment)
- Multiple sequence alignment (2 weeks)
  - Computational complexity and approximation algorithm
  - Heuristic methods (tree alignment)
  - Special applications (whole-genome alignment, 16S rRNA alignment, structural alignment)
- Phylogenetic analysis (1 week)

    - Distance-based approaches (UPGMA, Neighbor-Joining)
    - Parsimony and maximum likelihood
    - Detection of lateral gene transfer/recombination/reassortment
- Genome sequence assembly (2 weeks)
    - Shortest common superstring and greedy algorithm
    - Graph-theoretic paradigms (overlap graph, deBruijn graph)
    - Computational complexity
    - Incorporation of additional information (mate-pairs, optical maps)
    - Advanced topics (error correction, comparative assembly, genome finishing, assembly validation)
- Gene and motif finding (3 weeks)
    - Bacterial gene finding, ORF detection, Markov chains
    - Eukaryotic gene finding, HMM algorithm (Viterbi, forward-backward, etc.)
    - Incorporation of additional signals (promoters, splice signals, etc.)
    - Motif finding