AC-04-18

Algorithms for Mining Biological Sequences (COMP 680)

Instructor: Mathieu Blanchette
School of Computer Science and McGill Centre for Bioinformatics, 332 Duff Building
McGill University, Montreal, Canada
Phone: 398-5209
blanchem@mcb.mcgill.ca
http://www.mcb.mcgill.ca/~blanchem/ProposedCourse

Rationale:
The biotechnology revolution relies heavily on the ability to analyze the huge amount of data produced. Biological sequences (DNA and proteins) are of little use without a functional annotation, which can best be obtained through computational analysis (bioinformatics). This course will explore the cutting-edge algorithmic and machine learning techniques required for mining biological sequences, thus preparing students for innovative work in the industry and in research institutions while helping fuel research in applied molecular biology.

With the move of COMP 562 (Computational biology methods) to COMP 462 (Computational biology methods), the School of Computer Science does not offer a graduate course focused on algorithms for the analysis of biological sequences. The course proposed here gives graduate students the opportunity to get a solid background on sequence analysis, focusing on algorithmic, statistical, and machine learning approaches. Others (sequence alignment and hidden Markov models) will have already been introduced in COMP 462. These topics will only be very briefly reviewed before taking over where COMP 462 left off, introducing advanced algorithms and statistics not covered by COMP 462. The course is complementary to COMP 563 (Molecular Evolution Theory) and COMP 564 (Computation Gene Regulation) but shares little content with them. Because of its computational focus, the course proposed has no significant overlap with any other course offered at McGill. This course will involve an important final project, and students will also have to prepare a 60-minute in-class paper presentation. This amount of work is beyond the usual assignment workload.

Course abstract:
DNA sequencing techniques are now so advanced that new complete genomes are sequenced yearly. The goal of this course is to study the computational techniques needed to make sense of this huge amount of data. After a brief review of main biological processes of molecular biology, we investigate algorithms related to the annotation of biological sequences. Biological problems addressed include: sequence assembly, pair-wise and multiple sequence alignments and statistical significance of sequence similarity, comparative sequence analysis, gene discovery strategies, discovery of regulatory elements, detection of RNA genes, and identification of transposable elements. Computational approaches studied include dynamic programming algorithms (for sequence alignment, hidden Markov models, RNA secondary structure prediction) and heuristics of speed them up; hidden Markov models and their generalizations to pair, phylogenetic, and profile HMMs; expectation-maximization and Gibbs sampling algorithms for motif detection; statistical analysis of word and word-combination frequencies in DNA sequences.

Credits: 4. This course will involve a substantial final project, and students will also have to prepare a 60-minute in-class paper presentation. This amount of work is beyond the usual homework and assignment workload.

Expected attendance: 25 students

Prerequisites:
COMP 462: Computational Biology methods
or by permission of the instructor

Book Materials:
There will be no required textbook. Parts of the following books will be used:

[LH]	Chapter 1 of "Artificial Intelligence and Molecular Biology", by Lawrence Hunter. http://www.aaai.org/Library/Books/Hunter
[DEKM]	Chapters 2, 3, 4, 5, 6, 9, and 10 of "Biological Sequences Analysis", by Richard Durbin, Sean Eddy, Anders Krogh, Graeme Mitchison, Cambridge University Press 2001.
[BB]	Chapters 1 and 4, "Bioinformatics, the machine learning approach", by Pierre Baldi and Soren Brunak. MIT Press 2003.
[GE]	Chapter 9 from "Statistical methods in Bioinformatics", by Grant and Evans, Springer-Verlag 2001.

The following two landmark articles are also going to be used extensively:
"Initial sequencing and analysis of the human genome." Lander et al. Nature 2001, 409(6822):860-921.
"Initial sequencing and comparative analysis of the mouse genome." Waterston et al. Nature 2002, 420(6915): 520-62.

Articles from the following journals are going to be discussed: Nature, Science, Genome Research, Bioinformatics, Journal of Computational Biology.

Articles from the following conferences will be discussed: RECOMB, ISMB, PSB, WABI.

Computing resources:
We will use the computing resources of the School of Computer Science.

Course format:
During the first half of the semester the course will be in the format of lectures by the instructor. In the second half of the course, we will alternate between lectures by the instructor and students presenting an assigned paper related to the topics discussed.

Evaluation:
15%: In-class written midterm examination. This 90-minute examination will cover all material (including some basic molecular biology) discussed in class.
15%: 30-minute oral final examination. This examination will test the broader understanding of the concepts discussed and their relation to each other.
20%: Student paper presentation. Team of about five students (depending on enrolment) will be asked to give a 60-minute presentation of a research paper assigned at the beginning of the semester, and to lead a 20-minute class discussion of the paper.
50%: Final project. Students will work on a final project in teams of two or three. Topics will either be suggested by the teams themselves or by the instructor. The goal is that most projects undertaken may eventually lead to a Masters project and to publishable work.
The project is actually structured like a mini-Masters project and should help students in that process. A list of projects suggestions is available at:
http://www.mcb.mcgill.ca/~blanchem/ProposedCourse/projects.html

Project evaluation will be as follows:
5%: One-page project proposal describing the biological problem addressed and the computational and mathematical or computational tools being proposed.
5%: One-page literature review
40%: Final report, broken up as: 20% for methods developed; 10% for results obtained and their discussion; 10% for the overall quality of the report.

Schedule of lectures

1. Introduction:
    - Syllabus, topics covered, evaluation.
    - Bioinformatics as a reverse-engineering problem.
2. Basics of molecular biology
    - Prokaryotic and eukaryotic cell organization, proteins, DNA, RNA, chromosomes, genome, Central Dogma: transcription, translation, codon table.
    - Reading material: Chapter 1 of "Artificial Intelligence and Molecular Biology", by Lawrence Hunter. http://www.aaai.org/Library/Books/Hunter
3. Genome organization
    - Gene structure: introns, exons, splice sites
    - Other functional regions: RNA genes, regulatory regions
    - Non-functional regions: intergenic and intronic regions, repetitive regions, pseudo-genes
    - Genome annotation problem
    - Reading material: Same as above
4. Introduction to evolution and comparative genomics
    - DNA sequence evolution; substitutions, insertions, deletions, transposon activity, rearrangements, duplications.
    - Evolution as a random walk in sequence space
    - Phylogenetic inference problem
    - Genome comparisons: size, similarity.
    - Dogma of comparative genomics: Conservation implies function
    - Reading material: Same as above
5. Sequencing and assembly
    - Overview of sequencing technology.
    - Sequence assembly problem
    - Overlap-layout-consensus approach
    - Eulerian graph approach
    - Reading material: Pop, Salzberg, Shumway. Introduction to sequencing and sequence assembly. IEEE Computer 35(7): 47-54 (2002)
6. Sequencing and assembly
    - Eulerian graph approach
    - Reading material: Pevzner, Tang, Waterman. Euler sequence assembly algorithm. PNAS 98(17): 9748-9753 (2001)
7. Review of sequence alignment
    - Algorithms for pair-wise alignment: Needleman-Wunch, Smith-Waterman and variations.
    - Reading material: [DEKM] Sections 2.3 and 2.4
8. Fast local alignment heuristics
    - Blast algorithm
    - Notion of E-value, P-value, extreme-value distribution
    - Reading material: Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.
9. Statistics of local alignments
    - Introduction to Karlin-Altschul statistics
    - Reading material: [GE] Chapter 9.
10. Multiple sequence alignment algorithms
    - Scoring methods: sum-of-pairs, entropy, gap penalties, phylogenetic considerations
    - Exact dynamic programming algorithm
    - Progressive alignment (ClustalW) algorithm; profile alignment.
    - Word-based multiple alignment (DIALIGN)
    - Reading material: [DEKM] Chapter 6
      Morgenstern B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics. 1999 Mar;15(3):211-8.
11. Review of hidden Markov models
    - Definition, example

- Viterbi algorithm
- Forward-backward algorithm
- Reading material: [DEKM] Sections 3.1, 3.2

12. Hidden Markov models: Parameter estimation
    - Baum-Welch algorithm
    - Choice of topology
    - Reading material: [DEKM] Sections 3.3
13. Gene-finding with HMMs
    - Reminder of eukaryotic gene structure
    - Design of HMMs for gene finding
    - Reading material: Krogh. "Introduction to hidden Markov models for biological sequences", in Computational methods in Molecular Biology, 1998
14. Pair-HMMs and profile HMMs
    - Pair HMMs for sequence alignment
    - Viterbi variant, equivalence to Needleman-Wunch
    - Profile HMMs for remote homology detection.
    - Reading material: [DEKM] Chapter 4
15. **Midterm examination**
16. Finding transcriptional regulatory elements
    - Overview of regulatory mechanisms.
    - Motif finding problems and motif models.
17. Motif-finding algorithms: Expectation-Maximization
    - Expectation-maximization algorithms
    - Reading material: Bailey and Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proceedings ISMB94, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
18. *Student presentation: Statistical over-representation methods.*
    - Reading material: Sinha and Tompa. "A statistical method for finding transcription factor binding sites." Proc Int Conf Intell Syst Mol Biol. 2000;8:344-54.
19. Phylogenetic footprinting approaches
    - Reading material: Blanchette, Schwikowski, Tompa, "Algorithms for phylogenetic footprinting". Journal of Computational Biology, vol. 9, no. 2, 2002, 211-223
20. *Student presentation:* Comparative genomics in yeast.
    - Reading material: Kellis, Patterson, Birren, Berger, Lander. "Methods in comparative genomics: genome correspondence, gene identification, regulatory motif discovery." Journal of Computational Biology.
21. *Student presentation:* Discovery of regulatory modules.
    - Reading material: Sinha, van Nimwegen, Siggia. "A Probabilistic Method to Detect Regulatory Modules". ISMB 2003
    - Sharan, Ovcharenko, Ben-Hur, Karp. "CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.", Bioinformatics. 2003;19 Suppl 1:i283-91.
22. *Student presentation:* Detection of repetitive sequences
    - Reading material: Bao, Eddy. "Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes". Vol. 12, Issue 8, 1269-1276, August 2002
23. RNA secondary structure prediction.
    - Definition of RNA secondary structure
    - Nussinov algorithm
    - Zuker algorithm
    - Reading material: [DEKM] Sections 10.1, 10.2
24. *Student presentation:* RNA secondary structure models
    - Stochastic context-free grammars
    - CYK algorithm
    - Reading material: [DEKM] Chapter 9 and 10.3
25. Protein threading problem
    - Constraints graph

- Definition of optimization problem
- Branch-and-bound solution
- Reading material: Lathrop and Smith, "Global optimum protein threading with gapped alignment and empirical pair score functions." JMB, 255:641--665, 1996.

26. Hot topic of the day