

Course Syllabus

Instructor: Mehmet Koyutürk

Course Objectives

This course provides an introduction to the analysis of biological data using computational methods. Topics include sequence analysis, gene finding, pairwise and multiple alignments, gene mapping and haplotyping algorithms, motif identification, polymorphisms, phylogenetic analysis, microarray data analysis, and analysis of proteomic data. In the context of these applications, the course focuses on algorithmic techniques such as dynamic programming, string algorithms, graph theory, and Hidden Markov Models. It is expected that, upon completion of this course, the students will achieve the following objectives:

- Become familiar with existing tools and resources for computational analysis of biological data, including sequences, phylogenies, microarrays, ontologies, and biomolecular interactions.
- Develop an awareness of the computational problems that arise in the modeling and analysis of living systems.
- Understand fundamental abstractions and computational approaches used to formulate and address these problems.
- Be able to use, manipulate, and extend existing computational infrastructure for analyzing biological data.

Class Meeting

MW 9:00 AM – 10:15 AM, Bingham 305.

Instructor

Mehmet Koyutürk

Office: Olin 512

Phone: 368-2963

e-mail: koyuturk@eeecs.case.edu

Office hours: MW 10:30 AM – 11:30 AM. The students are also welcome to visit the instructor outside office hours.

Textbook

None. Supplementary reading materials will be provided by the instructor.

Prerequisites

None. Prior knowledge of molecular/cell biology is a plus, but it is not required.

Course Work & Grading

Participation: (10%) Attendance is obligatory. This has three reasons: (i) The class meets in the morning. (ii) The class size is relatively small, so we have to make sure that we have enough people to make discussions interesting. (iii) The main objective in this course is for the students to acquire a vision in biology and develop research skills that involve critical thinking. For this reason, in-class discussions are an essential part of the course work. Students are not allowed to use laptops in class (the reason for this should be clear).

Assignments: (30%) There are three written assignments. These assignments are comprised of problems that aim to allow the students to practice with the methods covered in class and develop ideas to manipulate these methods.

Presentation: (20%) At the end of each “chapter”, a student will present a (generally recent) research article on the corresponding topic. Each paper is chosen by the instructor to be illustrative of the state-of-the-art, as well as emerging challenges related to the topic. However, students are welcome to suggest alternate papers to present on their selected topic. Each student will have an entire class meeting to present their paper; therefore the presentation is expected to provide in-depth understanding of the research paper and facilitate critical discussion. The selection of research papers will be on a first-come first-serve basis.

Project: (40%) Students will develop and conduct research projects on a topic in computational biology and bioinformatics throughout the semester. The projects can be done individually or in teams of two; however the teams have to be interdisciplinary. The projects are required to be (i) innovative, (ii) involve implementation of a computational method, and (ii) provide solid computational results. The development of projects comprises the following phases:

1. *Topic selection.* Students will for teams and select their topic by the end of Week 2. Suggested broader areas for the topic include the following:
 - Next generation sequencing
 - Metagenomics
 - Structural variation in human genome
 - Phylogenetic networks
 - Statistical significance of complex patterns in biological data
 - Genomics/Functional genomics/Systems biology of cancer

2. *Project proposal.* (30%) The teams will review the literature and develop ideas for their project. Based on these, they will put together a 5-page research proposal by the end of Week 7. The proposal will outline the motivation and the proposed idea and clearly argue for the significance and the intellectual merit of the proposed research. It should also clearly explain the research plan.
3. *Project proposal.* (20%) Each proposal will be reviewed in a double-blind fashion by three students in the class. The reviewers will evaluate the proposals based on the significance of the proposed research, soundness of the proposed idea, and the concreteness of the research plan. The instructor will gather reviews and provide full feedback on the proposals by the end of Week 8.
4. *Project report.* (50%) At the end of the semester, the students will return 10-page project reports. The project reports are expected to reflect the quality of a publishable research paper; they should clearly explain the approach and provide solid results and discuss the conclusions.

Calendar

1. Biological Basics & Overview of Bioinformatics.

- (a) Aug 24: Evolution, domains of life, chemistry of life, structure of the cell.
- (b) Aug 26: Central dogma, DNA, RNA, proteins, wet lab techniques, "omics".

2. Sequencing.

- (a) Aug 31: Sequencing technology, fragment assembly. [*Selection of Research Papers Due*]
- (b) Sep 2: Shortest common superstring problem, greedy algorithm, sequencing by hybridization. [*Selection of Project Topics Due*]
- (c) Sep 9: *Research Paper* – Next generation sequencing [1].

3. Sequence Alignment.

- (a) Sep 14: Dynamic programming, Needleman-Wunsch, Smith-Waterman.
- (b) Sep 16: Affine gap penalties, heuristic approaches, BLAST. [*Assignment 1 Out*]
- (c) Sep 21: Scoring matrices, multiple sequence alignment.
- (d) Sep 23: *Research Paper* – Statistics of sequence alignment [2].

4. Hidden Markov Models.

- (a) Sep 28: Model inference, maximum likelihood, Markov models.
- (b) Sep 30: Viterbi algorithm, Baum-Welsh.
- (c) Oct 5: Profile HMMs, gene finding. [*Assignment 1 Due*]

(d) Oct 7: *Research Paper* – Copy number inference using HMMs [3].

5. Phylogenetics.

- (a) Oct 12: Tree reconstruction, distance-based methods. [*Project Proposals Due*]
- (b) Oct 14: Parsimony, maximum likelihood, assessment of reliability. [*Assignment 2 Out, Proposals Assigned for Review*]
- (c) Oct 21: *Research Paper* – Using phylogenetic trees for functional inference [4]. [*Proposal Reviews Due*]

6. String Matching & Motif Finding.

- (a) Oct 26: Exact string matching, suffix trees. [*Proposal Reviews Out*]
- (b) Oct 28: Regulatory motifs, rigid patterns, flexible patterns.
- (c) Nov 2: *Research Paper* – Phylogenetic footprinting [5].

7. Human Genetic Variation.

- (a) Nov 4: Haplotype inference from pedigree data. [*Assignment 2 Due*]
- (b) Nov 9: Haplotype inference from population data, tag SNP selection. [*Assignment 3 Out*]
- (c) Nov 11: *Research Paper* – Geographic distribution of human ancestral populations [6].

8. Gene Expression.

- (a) Nov 16: Transcriptional profiling, preprocessing of gene expression data, normalization, transformation.
- (b) Nov 18: Mining gene expression data: Differential expression, clustering, classification.
- (c) Nov 23: *Research Paper* – Synergistic dysregulation in complex phenotypes [7].

9. Systems Biology.

- (a) Nov 30: Molecular networks: protein-protein interactions, metabolic pathways, transcriptional regulation.
- (b) Dec 2: Module identification, network alignment, functional inference. [*Assignment 3 Due*]
- (c) Dec 4: *Research Paper* – Network based identification of disease markers [8]. [*Project Reports Due*]

Reading

1. Biological Basics & Overview of Bioinformatics.

- (a) L. Hunter. Life and its molecules: A brief introduction, *AI Magazine*, 2004.
- (b) A. R. Joyce and B. O. Palsson. The model organism as a system: integrating 'omics' data sets, *Nature Reviews Molecular Cell Biology*, 2006.

2. Sequencing.

- (a) S. Gopal, A. Haatke, R. P. Jones, and P. Tymann. Bioinformatics: A Computing Perspective, McGraw-Hill Higher Education, New York: 2008. (Chapters 2 and 3)

3. Sequence Alignment.

- (a) R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1999. (Chapter 2 and Sections 6.1–6.4)
- (b) W.R. Pearson, Protein sequence comparison and protein evolution. Tutorial, *ISMB*, 2000.

4. Hidden Markov Models.

- (a) R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1999. (Chapters 3, 5 and Section 6.5)

5. Phylogenetics.

- (a) R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1999. (Sections 7.1–7.4)

6. String Matching & Motif Finding.

- (a) M. Das and H. K. Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 2007.
- (b) K. D. MacIsaac and E. Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Computational Biology*, 2(4):e36, 2006.

7. Human Genetic Variation.

- (a) B. Halldorsson, V. Bafna, N. Edwards, R. Lippert, S. Yooseph, and S. Istrail. A survey of computational methods for determining haplotypes. *DIMACS/RECOMB Satellite Workshop*, 2004.

8. Gene Expression.

- (a) A. Schulze and J. Downward. Navigating gene expression using microarrays: A technology review. *Nature Cell Biology*, 2001.
- (b) J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 2002.
- (c) D. K. Slonim. From patterns to pathways: Gene expression data analysis comes of age. *Nature Genetics*, 2002.

9. Systems Biology.

- (a) M. Koyutürk. Algorithmic and analytical methods in network biology. *WIREs Systems Biology & Medicine*, in press.
- (b) A. L. Barabasi and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 2004.
- (c) T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 2008.

Research Papers

1. A. Sundquist, M. Ronaghi, H. Tang, P. Pevzner, and S. Batzoglou. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE*, 2007.
2. S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *PNAS*, 1990.
3. K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, and M. Bucan. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 2007.
4. R. Jothi, M. G. Kann, and T. M. Przytycka. Predicting protein-protein interaction by searching evolutionary tree automorphism space. *Bioinformatics*, 2005.
5. M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, 2002.
6. L. Parida, M. Méle, F. Calafel, J. Bertranpetit and The Genographic Consortium. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *Journal of Computational Biology*, 2008.
7. J. Watkinson, X. Wang, T. Zheng, and D. Anastassiou, Identification of gene interactions associated with disease from gene expression data using synergy networks, *BMC Systems Biology*, 2008.
8. K. Lage *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 2007.

Plagiarism Policy

Zero-tolerance policy on plagiarism is enforced. Following the departmental plagiarism policy, cheating on homeworks or tests will result in an F grade for the whole course and appropriate disciplinary action, independently of the extent of plagiarism. In case of doubt, the students are responsible for checking with the instructor on what is allowed and what is not.