Syllabus for EN 600.439/639: Computational Genomics

Fall 2013, Prof. Ben Langmead

Full description & administrative info is on Piazza site: https://piazza.com/jhu/fall2013/en600439639

There will be about 23 lecture-based class sessions, 1 session for the midterm exam, and 2 sessions for final project presentations. Auditors are required to attend the lecture-based class sessions, barring illness or religious holidays. Otherwise, JHU attendance policies apply.

The schedule is *not* set in stone. Things will shift if and when (a) I arrange for guest lecturers, (b) classes get canceled, (c) we go more slowly than I expected, etc. But every topic described below was covered in the Spring 2013 version of this course.

Don't forget to start thinking about your final project team and idea early on. Although the proposal isn't due until late October, you should have started by then. Final project materials are due on Friday, December 6.

"Jones & Pevzner" = An Introduction to Bioinformatics Algorithms (ISBN: 0262101068)
"Gusfield" = Algorithms on Strings, Trees and Sequences (ISBN: 0521585198)

Meeting 1: Sept 4 (Wed), 3 – 4:15pm

Lecture: Computational genomics: Administrative stuff, genomics, its importance, computational genomics, the role of computer scientists, notable successes, the role of DNA sequencing

Preparation:

- **Watch** videos 1 – 2: Links to videos are on Piazza page in "Videos" section of "Resources" tab.
- *Optional:* **Watch** videos 3 – 4
- **Read** class prerequisites document and review required material as appropriate.
- **Read** "Life and its Molecules." Link on Piazza page in "Readings" section of "Resources" tab.
- *Optional:* **Read** Jones & Pevzner Ch. 3

Assigned this class: **Homework 1** (due 1.5 weeks later on Mon, Sept 16)

Meeting 2: Sept 9 (Mon), 3 – 4:15pm

Lecture: Biological background: Genotype and phenotype, evolution, DNA and the genome, the central dogma, RNA and protein, DNA sequencing, sequencing-by-synthesis, FASTQ format

Preparation:

- *Optional*: **Read** "A decade's perspective on DNA sequencing technology" by Elaine Mardis. Link on Piazza page in "Readings" section of "Resources" tab.
- *Optional*: **Read** "Sequencing technologies — the next generation" by Michael Metzker. Link on Piazza page in "Readings" section of "Resources" tab.

Meeting 3: Sept 11 (Wed), 3 – 4:15pm

Lecture: Exact matching & indexing: DNA sequencing and the point-of-origin problem**,** naïve exact matching**,** Boyer-Moore**,** online versus offline**,** inverted indexing with substrings**,** index-assisted exact matching

Preparation:

- **Read** Gusfield 1.1 up to 1.1.1, 2.1, 2.2 up to 2.2.3 (ignore references to "fundamental preprocessing" or "Z algorithm" unless you're interested)
- *Optional:* **Read** Jones & Pevzner Ch. 9 up through 9.2, and section 9.6
- **Read** Jones & Pevzner section 9.7

Assigned this class: **Homework 2** (due 2.5 weeks later on Mon, Sept 30)

Meeting 4: Sept 16 (Mon), 3 – 4:15pm

Lecture: Inverted indexing: More index-assisted exact matching, Substring and interval lengths, hashing, approximate matching, hamming and edit distance, pigeonhole principle

Due this class: **Homework 1**

Meeting 5: Sept 18 (Wed), 3 – 4:15pm

Lecture: Suffix indexing: Tries and keyword trees**,** suffix tree**,** applications of suffix trees

Preparation:

- **Read** Jones & Pevzner: 9.4 (Keyword trees)
- **Read** Gusfield: Ch. 5 (all) (Suffix trees)

Meeting 6: Sept 23 (Mon), 3 – 4:15pm

Lecture: Suffix indexing: More suffix tree, applications of suffix tree, suffix array

Preparation:

- **Read** Gusfield: 7.14 through 7.14.2 (Suffix arrays)
- *Optional*: **Read** "Suffix arrays: a new method for on-line string searches" by Manber and Myers. Link on Piazza page in "Readings" section of "Resources" tab.

Meeting 7: Sept 25 (Wed)

I'll be away on the 25th, so we won't meet. Lecture will be posted online; watch it before next meeting.

Lecture: Suffix indexing: suffix array continued, longest common prefix (LCP), LCP-assisted suffix-array queries

Preparation:

- **Read** Gusfield: 7.14.3 through 7.14.5 (LCP-assisted suffix array)
- *Optional*: **Read** "Replacing suffix trees with enhanced suffix arrays" by Abouelhoda et al. Link on Piazza page in "Readings" section of "Resources" tab.

Meeting 8: Sept 30 (Mon), 3 – 4:15pm

Lecture: Suffix indexing: Burrows-Wheeler Transform, FM Index, methods used in practice

Preparation:

- **Read** "Introduction to the Burrows-Wheeler Transform and FM Index." Link on Piazza page in "Readings" section of "Resources" tab.
- *Optional*: **Read** "A Block-sorting Lossless Data Compression Algorithm" by Burrows and Wheeler. Link on Piazza page in "Readings" section of "Resources" tab.
- *Optional*: **Read** "Opportunistic Data Structures with Applications" by Ferragina and Manzini. Link on Piazza page in "Readings" section of "Resources" tab.

Due this class: **Homework 2**

Assigned this class: **Homework 3** (due 2 weeks later on Tue, Oct 15)

Meeting 9: Oct 2 (Wed), 3 – 4:15pm

Lecture: Dynamic programming alignment: Hamming and edit distance, dynamic programming, finding edit distance, global alignment, sequence similarity

Preparation:

- **Read** Gusfield: Ch. 11 through 11.6.4 (edit distance, dynamic programming, sequence similarity)

Meeting 10: Oct 7 (Mon), 3 – 4:15pm

Lecture: Dynamic programming alignment: Smith-Waterman (local alignment), approximate occurrences of P in T, time and space improvements, real-world performance

Preparation:

- **Read** Gusfield: 11.6.5 through 11.7.3 (approximate P-in-T, local alignment)
- **Read** Gusfield: Ch. 12 up through 12.2.3 (Hirschberg's algorithm and banded dynamic programming),

Assigned this class: **Homework 4** (due 3.5 weeks later on Wed, Oct 30), **Final project proposal** (due 3 weeks later on Mon, Oct 28)

Meeting 11: Oct 9 (Wed), 3 – 4:15pm

Lecture: Index-assisted alignment: Why dynamic programming is too slow, combining inverted indexing with dynamic programming, string neighborhoods, suffix-index co-traversal

Preparation:

- *Optional*: **Read** "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome" by Langmead et al. Link on Piazza page in "Readings" section of "Resources" tab.
- *Optional*: **Read** "Fast and accurate short read alignment with Burrows–Wheeler transform" by Li and Durbin. Link on Piazza page in "Readings" section of "Resources" tab.

Meeting 12: Oct 15 (Tue), 3 – 4:15pm (this is the makeup for Fall Break Day: Mon, Oct 14)

Lecture: Index-assisted alignment / Assembly: a short history of alignment software, intro to de novo assembly, whole genome shotgun sequencing, coverage, overlaps, shortest common superstring (SCS) problem

Due this class: **Homework 3**

Meeting 13: Oct 16 (Wed), 3 – 4:15pm

Lecture: Assembly: repeats foil assembly, why SCS falls short, alternatives to SCS, intro to overlap-layout-consensus (OLC) assembly, finding overlaps, layout step

Preparation:

- **Read** "Assembly of large genomes using second-generation sequencing" by Schatz, Delcher and Salzberg. Link on Piazza page in "Readings" section of "Resources" tab.

Meeting 14: Oct 21 (Mon), 3 – 4:15pm

Midterm Exam, covering everything up to but not including assembly

Meeting 15: Oct 23 (Wed), 3 – 4:15pm

Lecture: Assembly: de Bruijn graphs, k-mers, why some de Bruijn graphs are Eulerian, error correction

Preparation:

- **Read** "How to apply de Bruijn graphs to genome assembly" by Compeau, Pevzner and Tesler. Link on Piazza page in "Readings" section of "Resources" tab.

- **Read** Jones & Pevzner: 8.7 - 8.9 (sequencing by hybridization, de Bruijn graph assembly). Note: Sequencing by Hybridization is not a very important technology these days; a better motivation would have been next-generation sequencing (NGS), but the book is a bit too old for that.

Meeting 16: Oct 28 (Mon), 3 – 4:15pm

Lecture: Assembly: error correction, refining the de Bruijn graph, paired-end sequencing and scaffolding, real-world assembly software

Due by this class: **Final project proposal** (i.e. you must have met with me to discuss)

Meeting 17: Oct 30 (Wed), 3 – 4:15pm

Lecture: Sequence classification: interpreting sequences, types of sequence signal, simple example: CpG islands, probability review, Markov property, Markov chains and log-ratio scores

Preparation:

- **Read** Jones & Pevzner: Ch. 11 up to but not including 11.2. (CpG islands and Markov models)

Due this class: **Homework 4**

Assigned this class: **Homework 5** (due 2 weeks later on Wed, Nov 13)

Meeting 18: Nov 4 (Mon), 3 – 4:15pm

Lecture: Sequence classification: Hidden Markov Models (HMMs), occasionally dishonest casino, Viterbi algorithm, simple examples

Preparation:

- **Read** Jones & Pevzner: 11.2 and 11.3 (HMMs)

Meeting 19: Nov 6 (Wed), 3 – 4:15pm

Lecture: Sequence classification & gene finding: finding CpG islands with HMMs, genes, eukaryotic gene expression, HMMs for gene finding

Preparation:

- **Read** "Computational Prediction of Eukaryotic Protein-Coding Genes" by Michael Zhang. Link on Piazza page in "Readings" section of "Resources" tab.

Meeting 20: Nov 11 (Mon), 3 – 4:15pm

Lecture: RNA sequencing: alternative splicing, RNA sequencing, isoform discovery and quantitation, spliced alignment, minimum path cover

Preparation:

- *Optional*: **Read** Bill Majoros' slides on HMMs for Eukaryotic gene prediction: http://www.geneprediction.org/book/hmm-part1.pdf

Meeting 21: Nov 13 (Wed), 3 – 4:15pm

Guest Lecture. Prof. Steven Salzberg: http://ccb.jhu.edu/people/salzberg/Salzberg/Salzberg_Lab_Home.html

Due this class: **Homework 5**

Meeting 22: Nov 18 (Mon), 3 – 4:15pm

Guest Lecture: Prof. Jeff Leek: http://www.biostat.jhsph.edu/~jleek/

Meeting 23: Nov 20 (Wed), 3 – 4:15pm

Lecture: RNA sequencing: quantitation, likelihood functions

Meeting 24: Nov 25 (Mon), 3 – 4:15pm

Lecture: **Wrap-up**: other types of sequencing, other computational challenges, computer science in the basic sciences.

(No meeting on Nov 27 (Wed), Thanksgiving Break)

Meeting 25: Dec 2 (Mon), 3 – 4:15pm

Final Project Presentations (about 15 minutes each, about 4 teams this meeting)

Meeting 26: Dec 4 (Wed), 3 – 4:15pm

Final Project Presentations (about 15 minutes each, about 4 teams this meeting)

Dec 6 (Friday)

Final project materials due 11:59pm